Jurnal Teknologi, 47(C) Dis. 2007: 21–36

THE MISCONCEPTIONS OF SOME STATISTICAL TECHNIQUES IN RESEARCH

HABSHAH MIDI[1], A. H. M. RAHMATULLAH IMON[2] & AZMI JAAFAR[3]

Abstract. In today's society, statistical techniques are being used with increasing rate in education, medicine, social sciences, and applied sciences such as engineering. They are crucial in interpreting data and making decisions. Based on our experience and observation through seminars, conferences and consultations, we noticed some statistics practitioners often misuse some of the statistical techniques in their researches. The easy availability of the statistical packages such as SAS, SPSS, has driven more researchers to use the packages in analysing their data. They need not to consult the statisticians and this led to greater abuse of statistics in data analysis. Consequently, meaningless and misleading conclusions are obtained from an incorrect analysis. Due to the lack of awareness, the policy makers often rely on these results to make decisions and that means disaster to a community or to a country. Therefore, it is imperative for the researchers to be aware of using the right statistical techniques so that a valid and objective conclusion can be made. In this paper, we will draw attention to some incorrect practices on selected topics caused by the lack of awareness of the researchers. Appropriate suggestions are offered to tackle this problem.

Keywords: Sampling frames; underlying assumptions; robust method; convenient sampling; central limit theorem; design and analysis of experiment

Abstrak. Dalam masyarakat hari ini, teknik berstatistik digunakan secara meluas dalam bidang pendidikan, perubatan, sains sosial dan sains gunaan seperti kejuruteraan. Ianya sangat penting sekali bagi mentafsirkan data dan untuk membuat keputusan. Berdasarkan pengalaman dan penilikan kami melalui seminar, konferensi dan khidmat perundingan, kami dapati beberapa pengamal statistik menyalahguna beberapa teknik statistik dalam penyelidikan mereka. Ketersediaan pakej statistik yang mudah diperoleh seperti SAS dan SPSS, telah menyenangkan penyelidik menganalisis data mereka. Mereka tidak perlu merujuk kepada ahli statistik dan ini membawa kepada penyalahgunaan statistik yang lebih parah dalam analisis data. Akibatnya, kesimpulan yang tidak bermakna dan mengelirukan diperoleh dari analisis yang salah. Oleh kerana kurang kesedaran, para pembuat polisi hanya bersandarkan kepada hasil tersebut untuk membuat keputusan dan ini mungkin membawa bencana kepada masyarakat atau negara. Dengan demikian, amat penting bagi penyelidik untuk menyedari penggunaan teknik statisitik yang betul, agar kesimpulan yang sah dan objektif dapat dikemukakan. Dalam kertas ini, kami akan menarik perhatian terhadap beberapa amalan silap bagi beberapa topik pilihan disebabkan kurangnya kesedaran para penyelidik. Beberapa cadangan yang sesuai diketengahkan bagi menangani masalah tersebut.

Kata kunci: Bingkai pensampelan; andaian termaktub; kaedah teguh; pensampelan selesa; teorem had memusat; reka bentuk dan analisis uji kaji

[1] Laboratory of Applied and Computational Statistics, Institute For Mathematical Research, University Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia
[2] Deparment of Mathematical Sciences, Ball State University, Muncie, Indiana, USA
[3] Faculty of Computer Sciences and Information Technology, University Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

1.0 INTRODUCTION

Every day we encounter data and they affect our decision making. In all aspects of our lives, an amazing diversity of data is available for inspection and analysis. Business managers, government officials, policy makers and professionals require solid facts based on data to justify a decision. Variations are inevitable in life! Every sample that we collect has variation. Variation can only be studied using statistical techniques. This is where statistics play an important role. They need statistical techniques to support their decisions since statistical analysis of data can provide investigators with powerful tools for making sense out of data. Some people are suspicious of conclusions based on statistical analyses. Extreme sceptics, usually speaking out of ignorance, characterise the discipline as a subcategory of lying, sometimes used for deception rather than for positive ends. However, we believe that statistical methods, used intelligently, offer a set of powerful tools to collect, analyse and interpret data relevant to their decision-making. Therefore statistical skills enable them to intelligently collect, analyse and interpret data relevant to their decision–making. Statistics are indicators, not definite guides to assist one in making decision under uncertainties based on data. It discounted personal opinion or beliefs. Just like weather, if we cannot control it, we should learn how to measure and analyse it using an appropriate statistical technique in order to predict it effectively.

Inferential statistics is concerned with making inferences about certain characteristics of a population based on information contained in a random sample of data taken from the entire population.

Once the data has been collected, the next step is to analyse it by using statistical analysis to support a hypothesis. Researchers no longer need to plug numbers into formulas to do the statistical analysis. Recent advances in computer sciences have invented statistical packages such as SAS, SPSS, MINITAB etc to do the work for them. Without adequate knowledge in statistical techniques, some researchers simply instruct the packages to analyse data using their convenient techniques. Unfortunately, they often are not aware of the fact that statistical packages just follow the instructions given to them and produce results accordingly. They do not know whether researchers have chosen the correct statistical techniques for their studies. Box [1] stated that "now it's really too easy, you can go to the computer and with practically no knowledge of what you are doing, you can produce sense or nonsense at a truly astonishing rate." Since statistical methods are used frequently to help policy makers in making decisions, it's very vital for the researchers to have basic understanding of statistics. In this paper, we will discuss inferential statistics, which allow us to draw reliable conclusions based on sample data if appropriate statistical techniques are used. The conclusions drawn from a study are to be trusted only when correct sampling methods are used to collect a sample. Furthermore, it is usually unwise to rely on the results of test procedures unless the validity of all underlying assumptions such as independence, normality and

purity (free from outliers) of observations, have been checked. Violations of these basic assumptions may produce sub-optimal or even invalid inferential statements and inaccurate predictions. In this paper, we will also draw attention to the common incorrect practices by the lack of awareness among researchers and give appropriate suggestions to remedy these problems.

2.0 METHOD

We have considered a variety of commonly used statistical methods in this paper.

2.1 Sampling Techniques

The important task of statistics is the scientific methodology for collecting, analysing, interpreting a random sample from a population in order to make inferences on the entire population of interest. There are many reasons for selecting a sample rather than obtaining information from a population. The main reason of not studying the entire population was that it was too expensive and too time consuming. Also the process could be destructive, as in measuring the breaking strength of cars or soda bottles, pathological tests, the sugar content of oranges, the life times of light bulbs etc where it would be simply impossible and/or foolish to study the entire population. Many studies had been done when a sample had been taken from a population. The aim is to generalise the conclusion from a sample to the corresponding population. As a result, it is important that the sample be representative of the population. Therefore, it is vital that the investigator defines carefully and completely the population and constructs the sampling frames before collecting the samples. A frame is a list of sampling units, which are non-overlapping collection of elements from the population such as objects or individuals.

It is very important to note that a commonly used method for selecting a random sample starts with creating a sampling frame first. Then we need to choose an appropriate sampling technique such as simple random sampling, stratified random sampling, cluster sampling, systematic sampling etc.

2.2 Non-Response Bias

One of the common types of bias encountered in sampling is non-response bias, which occurs when we do not get response from all the individuals included in the sample. Such bias can distort results if those who respond to a question differ from those who do not respond in a significant way. A serious effort must be made to minimise this non-response bias by following up with individuals who do not respond to an initial request for information. The non-response rate should be as low as possible and as a guide it should not exceed 5% of the target sample size. Otherwise an increase in the sample size does not help in reducing the sampling bias.

2.3 Hypothesis Testing

2.3.1 Underlying Assumption

It is a common practice to resort to parametric procedures while analysing data. We conventionally use z-test, t-test, Chi-Square test and F-test for testing hypotheses. One does not just learn formulae and plug in numbers in the formulae, but one has to learn about the conditions or assumptions under which the statistical testing procedures can be applied. The assumptions that are common to almost all statistical tests are that the observations are (i) random, (ii) independent and identically distributed, (iii) come from a normal distribution and (iv) equally reliable (there is no outlier in the data).

2.4 Regression

2.4.1 Linear Regression

A number of problems in analysing data involve the description of how variables are related. The simplest of all models describing the relationship between two variables is a simple linear regression. However, in many situations, the relationship between the dependent (response) variable and any single variable is not strong. The knowledge of values of several independent variables may considerably reduce uncertainty about the value of the response. For example, some variation in house prices in a large city can certainly be attributed to house size, but knowledge of size by itself would not usually enable a bank officer to accurately predict a home's value. Price is also determined to some extent by other variables, such as age, lot size, number of bedrooms, and distance from the schools. In similar situations, multiple linear regressions are more appropriate. Nevertheless, it is always best, to be parsimonious, to use as few variables as predictors as necessary to get a reasonably accurate predictions. The main objective of regression analysis is to predict the value of the dependent variable based on the values of one or more independent variables.

2.5 Unusual Observations and Underlying Assumptions

The assumptions underlying the linear regression model are that the independent variables are non-stochastic variables whose values are fixed and the errors are independently and identically distributed with zero mean and constant variance. Another assumption that has received much attention from statisticians in recent years is that the regression analysis must be free from the effect of any kind of unusual observations in the data set. Chatterjee and Hadi [2] enumerated that all observations are equally reliable and should have an equal role in determining the Ordinary Least Squares (OLS) results and influencing conclusions. In statistical data analysis, we have only one type of unusual observation that we call outlier, but in a regression problem

extra care should be taken because in this situation, we have three different types of unusual observations: outliers, high leverage points and influential observations. When all the model assumptions are met, according to the Gauss-Markov Theorem, the OLS estimates are unbiased and have minimum variances among all linear and unbiased estimators. We have noted earlier that the validity of model assumptions have to be checked so that valid inferences can be made.

2.6 The Effect of Unusual Observations

Unfortunately, many researchers are not aware of the immediate consequence of the presence of outliers. It may cause apparent non-normality and the entire classical inferential procedure might breakdown in the presence of outliers. Even one single outlier can have arbitrarily large effect on the estimates. The effects of unusual observations as described by Chatterjee and Hadi [2] are as follows:

(i) The estimation of regression parameters and different tests designed for regression problem are often badly affected in the presence of outliers, especially if they are influential.
(ii) Sometimes outliers may distort the homoscedasticity of errors or mask their inherent heteroscedastic pattern.
(iii) Outliers and high leverage points may affect the variable selection procedure of a regression model.
(iv) High leverage points often affect the identification procedure of outliers by causing the masking and/or swamping.
(v) High leverage points are mainly responsible for inducing or masking the multicollinearity problem.

2.7 Robust Methods

For the identification of unusual observations we often employ diagnostic techniques, which are basically designed to find problem with assumptions. Diagnostic methods are very simple and they are also very popular with the practitioners. They are also very effective in the presence of a single unusual observation. But diagnostic methods suffer a huge set back when a group of unusual observations are present in the data. To remedy this problem, a robust (resistant) method is put forward. A robust technique tries to make the effects of outliers as small as possible. Robustness signifies insensitivity to small deviations from the usual assumptions. When all the classical assumptions have been met, the robust technique is nearly as efficient as the classical procedure. However, when there is a small departure from the usual assumptions, the robust procedure is more efficient. A robust regression is extremely useful in identifying outliers and assessing the adequacy of a fit and suggesting suitable transformations.

All of these aspects can be detected in a single run by simply running a robust technique. Diagnostic and robust regression has the same goal, but they proceed in the opposite order. As mentioned by Rousseeuw and Leroy [3], in diagnostic setting, one first wants to identify the outliers and then fit the good data in the classical way, whereas the robust approach first fits a regression to the majority of the data and then discover the outliers as those points having large residuals from the robust fit. There are considerable papers related to robust regression such as Rousseeuw and Leroy [3], Ryan [4], Atkinson and Riani [5] and Montgomery *et al.* [6]. The commonly used robust techniques among others are the L, M, MM, Generalized M, Least Median of Squares (LMS), Least Trimmed Square (LTS) and Reweighted Least Squares (RLS).

2.8 Designs and Analysis of Experiment

Design and analysis of experiment is a process of planning an experiment with suitable design so that appropriate data will be collected and can be analysed by appropriate statistical methods, resulting in a valid and objective conclusion. In the experimental studies, the independent variables of interest (factor) are under the control of the experimenter. The experimenter controls each group by the randomisation process whereby the treatments (factor levels) are assigned at random to the experimental units. The objective of an experiment is to determine the effect of the manipulated factors on the response variables. Researchers must know how to choose the right design. The main difference between different types of design lies on the design objectives and different techniques of randomisation of the treatments (factor levels) to the experimental units. The consequence of using the wrong design will lead to invalid inferences.

2.9 One-way Analysis of Variance (Completely Randomised Design)

In this design, only a single factor is being investigated and no other factors will affect the experimental result because they are held fixed. The experimental units must be kept as uniform or as homogeneous as possible and the treatments are assigned at random to the experimental units.

2.10 Two Way Analysis of Variance (Completely Randomised Block Design)

In this design blocking systematically controls the extraneous factor other than the one being considered. The experimental units are first sorted into homogeneous groups called blocks. The treatments are then assigned at random to the experimental units within each block. The blocks are considered here chiefly as the means for reducing experimental error variability.

3.0 RESULTS AND DISCUSSION

In this section we present some results and examples from some diversified fields of statistics such as sampling, regression, hypothesis testing and experimental design to justify the misconceptions of some statistical techniques. These examples demonstrate the fact that the lack of awareness of not using appropriate statistical and sampling techniques in data analysis, would often produce meaningless and misleading conclusions.

3.1 Sampling

In practice, a researcher always claims that a random sample has been selected in his study. However, there is no way to verify his claim that it is a genuine representative of the population from which it was drawn just by looking at a sample. We can be sure only when we know that the appropriate method has been used to select the sample. We encounter many 'so called' random samples that have no sampling frames at all. Suppose a researcher wants to compare public opinion on certain issues between three ethnic groups in a country, over 21 years old. For this study, a common practice is to conduct a survey at a certain convenient place, say at a shopping complex. Then the researchers usually claim that a random sample has been collected and then generalise their conclusions for the whole population even though the sampling frames did not exist. They do not realise that the sampling method applied here is not random sampling but a convenient sampling. A voluntary response sampling is one of the most common forms of convenient sampling. Such samples rely entirely on individuals who volunteer to be a part of the sample, often by responding to an advertisement, calling a publicised telephone number to register an opinion, or logging on to an internet site to complete a survey. It is extremely unlikely that individuals participating in such voluntary response survey are representative of any larger population of interest. Consequently, results obtained from such sampling methods are rarely informative, and it is totally wrong to generalise the findings to any larger population. Only descriptive statistics can be applied to such convenient samples but not any kind of inferential statistics. Therefore, no statistical test is appropriate in this situation because to use test statistics such as *t*-tests required that the observations are a random sample from a population. Unfortunately, we have encountered many studies with convenient sampling, but statistical tests were conducted and generalised their results to a population.

3.2 Small Sample Size

It is a common misconception among researchers that a sample cannot accurately reflect the population if the sample size is relatively small compared to the population size. This misconception arises because the researches fail to realise the merit of random

sampling. Researchers tend to take a larger sample when dealing with very large population. The size of the population generally determines the size of the sample. However, there is no hard and first rule for the sample size determination. About 50% of the population should be chosen if the population size is 400 – 600. For larger group, such as thousands, 20% of the total population should be adequate. For a nation wide survey, a very small percentage of population, say 0.001% could produce a representative sample. Newport *et al.* [7] reported that for a very large population (nation wide survey), a sample size between 1200 – 1300 (e.g. Gallup polls with 1000 samples for a country like the USA) could be enough in simple random sampling to infer within 3% margin of error (for 1% margin of error, the required sample size for USA is 1,000). However, Cochran [8] discussed some sophisticated methods for estimating sample sizes, based on three criteria, i.e., the sampling techniques that are being used, the desired precision and the margin of error one would allow in the inferential procedure. The quality of random samples is high if the sampling is done very carefully and efficiently.

3.3 Hypothesis Testing

The assumptions regarding the hypothesis testing that we presented previously are crucial, not only for the method of computation, but also for the testing using resultant statistics. Therefore we have to check whether all of these assumptions have been met for a valid inferential statement. We may use diagnostic checking to confirm the validity of these assumptions. The run test, the pairs (serial) test and the gap test can be used to test for randomness. For testing independence, one may use the Chi-Square or Cochran's test, while the Shapiro-Wilk or Anderson-Darling test can be used for testing normality. The equality of variances can be tested using the F-test or Brown-Forsythe test. The exploratory data analysis methods such as stem and leaf plot and box plot are particularly useful for identifying extraordinary observations and detecting violations of traditional assumptions.

3.4 Non-parametric Procedure

When the assumptions are not satisfied, we may use other statistical procedures such as non-parametric or robust statistical procedures. The non-parametric tests are used when some specific conditions for the parametric tests are violated. In a non-parametric procedure, the probability distribution of the statistic upon which the analysis is based on does not dependent on specific assumptions about the populations from which the samples are drawn, but only on more general assumptions, such as continuity and/or symmetric of population distribution. For example, the Chi-square test concerning the variance of a given population is parametric since the test requires that the population distribution be normal, but the Chi-square test of independence does not assume normality, therefore it is a non-parametric test.

3.5 Central Limit Theorem

By using the Central Limit Theorem, when sample sizes are large, the observations tend to follow a normal distribution and we can use z-statistics to perform any kind of test. However, many researchers are not aware of the fact that the large samples (fixed size n, say more than 30) must be random and independent. This is another misconception that any larger sample size will automatically get closer to normal distribution even though the samples are not random.

3.6 Availability of Population Data

Occasionally, a researcher can obtained all the data from the population of interest. In this situation, a common mistake that we noticed is that researchers either take a random sample and then carry out a hypothesis test or simply carry out a hypothesis test based on the population data even when the complete information for population is available. Sometimes people forget this fact that it should be obvious that no test is needed to answer questions about a population if complete information are available and researchers do not need to generalise any conclusion from a sample.

3.7 Robust Location and Scale Estimators

In the presence of outlier, the sample mean and sample standard deviation are not robust. Robust estimator of location parameter such as sample median, trimmed mean and weighted mean are the alternatives to sample mean. As an alternative to standard deviation, we can use the robust scale estimator such as the Median Absolute Deviation (MAD).

3.8 Regression

The estimation of regression parameters and different tests designed for regression problems such as z, t, Chi-Squares and F is badly affected when the assumption of normality is violated. Judge *et al.* [9] pointed out that the violation of normality assumption may lead to the use of suboptimal estimators, invalid inferential statements and inaccurate predictions. It can be detected by using normal probability plot, Shapiro-Wilk, Anderson-Darling, Bowman-Shenton (Jarque-Bera), and rescaled moments tests (see Imon [10]). When the data are collected over time, the assumption of uncorrelated errors is frequently violated and the consequences can be very serious. The problem of no constant (heteroscedastic) error variances occurs quite frequently in practice and can be detected by residual plot and Brown-Forsythe and some other tests. This problem can be remedied by using the Weighted Least Squares (WLS) method if known weights are available. For unknown weights, we may use the Iteratively Reweighted Least Squares (IRLS) or the Transformed Both Sides (TBS) methods.

Outlier is a single or a group of observations, which are markedly different from the bulk of the data or from the pattern, set by the majority of the observations. Outliers occur very frequently in real data, and they often go unnoticed because nowadays computers process much data set. Hampel *et al.* [11] claim that a routine data set typically contains about 1 – 10% outliers, and even the highest quality data set cannot be guaranteed free of outliers.

3.9 Robust Regression

Example 1: Belgian Fire Data

This data set shows the trend of the number of reported claims of Belgian Fire Insurance Companies from 1976 to 1980 (see Rousseeuw and Leroy [3]). From the scatter plot, it can be seen that there is a slight upward trend over the years. However, one will notice that the number of fires for the year 1976 is extraordinarily high. The reason lies in the fact that in that year the summer was extremely hot and dry compared to Belgian standards, causing trees and bushels to catch fire spontaneously. The scatter plot is shown in Figure 1.

Figure 2 illustrates the OLS fits to the Belgian Fire Data with and without outlier. One will notice that even with one outlier the sign of the slope of the OLS fit can be changed. The OLS and the RLS fit of the Belgian Fire Data are presented in Figure 3. The OLS fit is very sensitive to outliers and it tends to go towards the direction of outliers. It is also observed that the RLS (robust) fit is not sensitive to the presence of outliers.

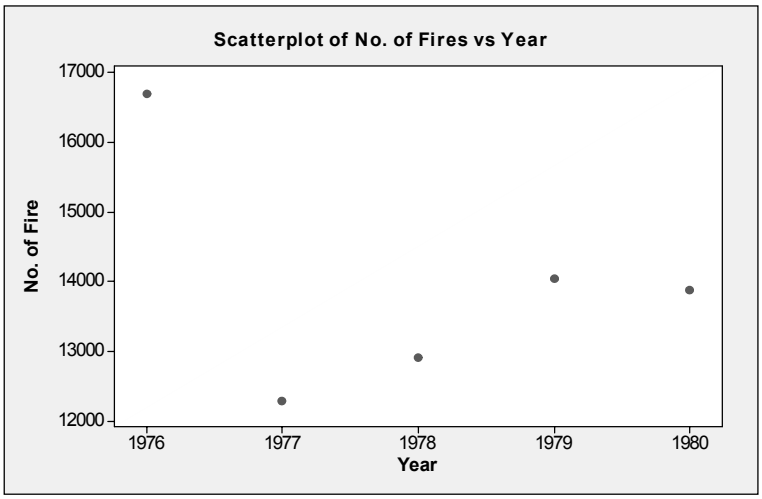


Figure 1 Scatter plot of Belgian fire data

Example 2: Belgian Telephone Data

The data set on the total number (in ten of millions) of international phone calls made between the years 1950 to 1973 was obtained from Rousseeuw and Leroy [3]. This time series data contains heavy contamination from 1964 to 1969. Upon inquiring, it turned out that during that period another recording system was used giving the total number of minutes of these calls.

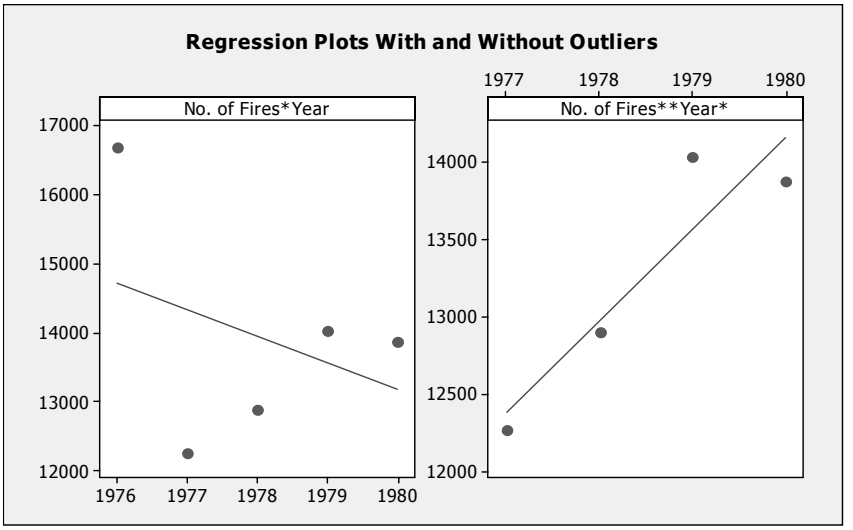


Figure 2 Regression lines with and without outlier

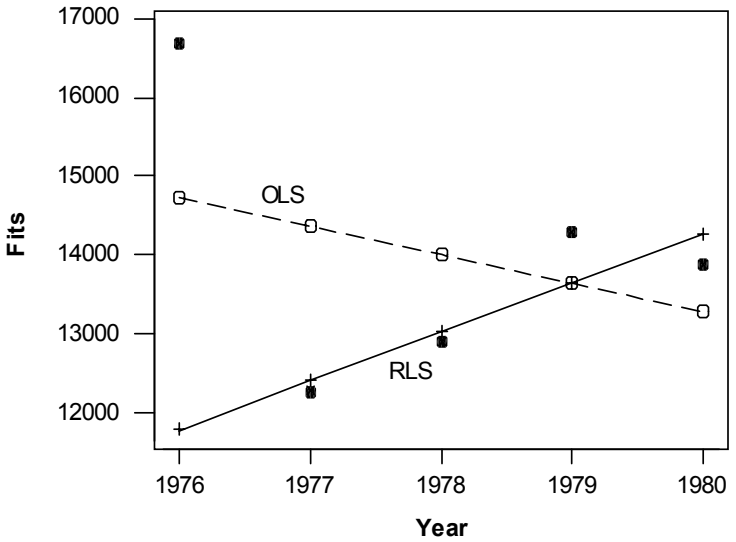


Figure 3 The OLS and the RLS fit to the Belgian fire data

Figure 4 illustrates the scatter plot of Belgian Telephone Data together with the OLS and RLS fits. Likewise the previous example, we observe that the OLS fit is very sensitive to outliers and it tends to go towards the direction of outliers. However, the RLS fit is not affected in the presence of outliers.

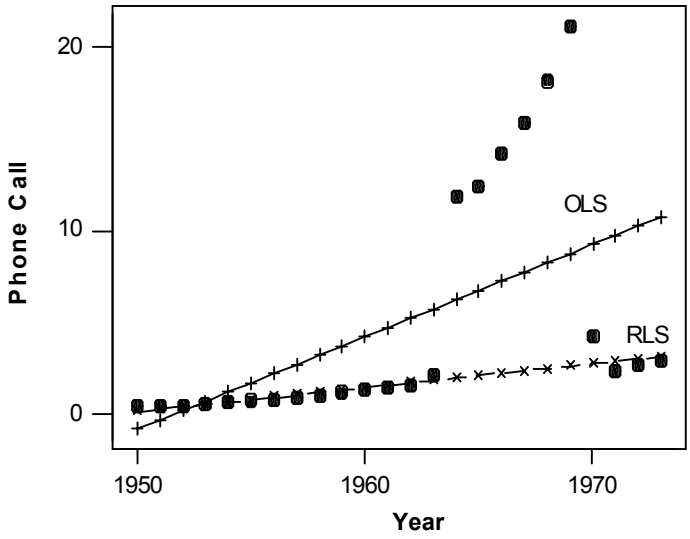


Figure 4 The OLS and the RLS fits to the Belgian telephone data

3.10 Goodness of Fit Test

Researchers often rely on R^2 to determine the goodness of fit of a model. This measure gives the ratio of the variation explained by the independent variables to the total variation of the response variable. The value of R^2 ranges from 0 to 1; the value closer to 1 indicates a good fit while the value closer to 0 indicates a poor fit. Even though R^2 is a well-accepted measure, it has some serious drawbacks.

- For a single X increasing the range of X can increase the value of R^2.
- The value of R^2 may also be artificially large if the sample size is small relative to the number of regressors.
- For each variable that is added to a known model, will increase the R^2.
- The presence of outlying observations will affect the value of R^2.

With these drawbacks, it will give a misleading indicator to the goodness of fit of a model. One can use robust R^2 as a remedy to this problem.

3.11 Multicollinearity

When dealing with multiple linear regression one should be aware of the problem of multicollinearity. Multicollinearity exist when the independent variables of a multiple linear regression is correlated with each other or correlated with other important variables that is not included in the model. The OLS estimation technique may break down and may give wrong signs, inflated variances and insignificant regression coefficients. The multicollinearity problem can be detected by examining the Variance Inflation Factor (VIF), condition indices or eigen/singular values decomposition. One may use the Ridge Regression or Latent Root Regression as remedy to this problem. The problem becomes more complicated when outliers come together with multicollinearity. A robust ridge or robust latent root regression is proposed to handle this situation.

3.12 Design and Analysis of Experiment

Outliers may have an adverse effect on the analysis of variance techniques. Figure 5 is a one-way design and illustrates how a single outlier can reverse the conclusion of the analysis of variance. Data Set 1 shows a clean data. A researcher has misreported the value of the fifth observation of treatment A. Instead of typing 144, he has typed 744. We called this, Data Set 2. We can see from the given table that by wrongly recording the observation as 744 instead of 144 has made the test not significant (p value equals 0.353). As an alternative, one can use robust designs.

The example in Figure 6 illustrates the fact on how a wrong choice of design could lead to a misleading conclusion and invalid inferences. Suppose a researcher would like to determine if the student's performance on computer programming scores for the final year students majoring in Computer Science are essentially the same in four universities. Three final year students majoring in Computer Science are selected at random from each university and they are given a programming test. The results (the students scores and ANOVA table) are presented in Figure 6. It can be seen from ANOVA Table 1 that the test is not significant. The explanation for this is, perhaps, the variability which is due to the ability (extraneous factor) of students, is not taken into consideration and not measured. This variability was included in the experimental errors that may inflate the MSE and lead to small value of F statistics (p value = 0.37).

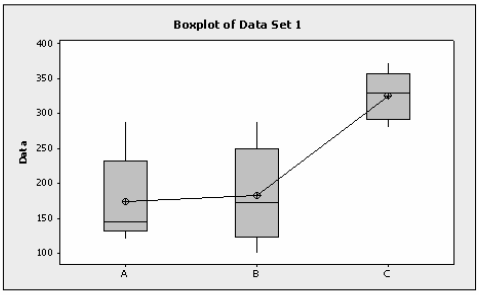
The experiment was repeated by randomly selecting from each university, one computer science final year student with high Cumulative Grade Point Average (CGPA), medium CGPA and low CGPA. The results (the students scores and ANOVA table) are presented in Figure 6. This table shows that when this known source of variability is measured by blocking the ability of students according to their CGPA, this variability is filtered out from the experimental error and hence reduces the MSE and lead to significant conclusions (p value = 0.007). This example illustrates how serious is the consequences of using the wrong design. However, we encountered

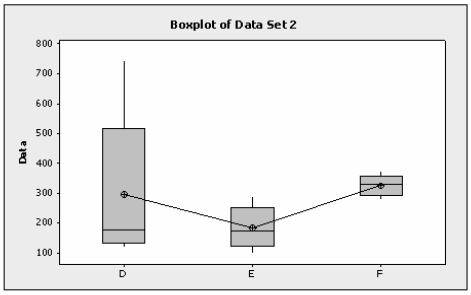
Data Set 1

A	144	120	176	288	144
B	101	144	211	288	172
C	330	280	341	374	302

Data Set 2

D	144	120	176	288	744
E	101	144	211	288	172
F	330	280	341	374	302





Source	SS	DF	MS	*F*	*p*
Factor	71832	2	35916	9.99	0.003
Error	43125	12	3594		
Total	114957	14			

Source	SS	DF	MS	*F*	*p*
Factor	55912	2	27956	1.14	0.353
Error	294645	12	24544		
Total	350557	14			

Figure 5 Data with one outlier

Data 1

University	Scores		
A	65	85	90
B	80	80	60
C	50	45	85
D	70	85	80

Data 2

	CGPA		
	Below 2.5	2.5 – 3.5	Above 3.5
A	70	85	92
B	45	50	85
C	50	64	70
D	65	80	86

ANOVA TABLE 1

Source	SS	DF	MS	F	*p*
University	739.58	3	246.53	1.17	0.37
Error	1683.33	8	210.42		
Total	2422.91	11			

ANOVA TABLE 2

Source	SS	DF	MS	F	*p*
University	1128.33	3	376.11	7.05	0.007
Ability	1327.17	2	663.58	12.44	0.00
Error	320.17	6	53.36		
Total	2,775.67	11			

Figure 6 Wrong design example

many studies that ignore the basic principles of design and analysis of experiments. They are contented and satisfied with their statistical results as they claimed that the statistical packages that they used couldn't make mistakes. But they are not aware of the fact that these packages are just 'dumb machines', they just follow instruction from the researcher as to what design to use. They do not 'understand' whether the researcher has chosen the right design.

Another important concept to remember is that even if a correct design has been chosen, the result of analysis of variance cannot be trusted unless all the assumptions underlying the model assumptions has been checked and met for valid inferences.

4.0 DISCUSSION

No statistical technique can be used to eliminate or explain all of the uncertainties in the world. However, statistics can be used to quantify that uncertainty. That is the reason why statistical techniques have been used widely to help policy makers make decisions. One cannot just use statistical techniques blindly without prior knowledge or sound knowledge in statistics. We have discussed some topics in statistical analysis where researchers often are not aware of the adverse consequences of using incorrect or incomplete analysis. In today's society, it is very unfortunate that many researchers with little knowledge of statistics rely on statistical packages to analyse their data. They do not even bother to consult statisticians since they think that statistical packages can provide them with all the necessary analysis they require. To get a valid inference, a right sampling and statistical techniques and the correct design should be chosen. In any statistical technique, a proper adequacy checking of the underlying assumptions are to be performed. When the basic assumptions are not satisfied, proper remedial measures should be taken into considerations such as transformation of either the response or the regressor to provide an appropriate fit to the data. One should prefer the parametric model, especially when subject area theory supports the transformation used. A robust procedure should be used if one suspects the existence of outlier in the data. One should use (robust) nonparametric procedures if no simple parametric model yields an adequate fit to the data, when there is little or no subject area theory to guide the analyst, and when no simple transformation appears appropriate. By ignoring the correct sampling, correct statistical methods, correct design and adequacy checking will lead to invalid inferences and inaccurate predictions. Consequently, policy makers become ignorant of the fact and they are bound to rely on meaningless and misleading results to make decisions and that may bring disaster to a community or to a country.

REFERENCES

[1] Box, G. E. P. 1969. The Challenge of Statistical Computation. Statistical Computation. 3-10.

[2] Chatterjee, S. and A. S. Hadi. 1988. *Sensitivity Analysis in Linear Regression.* New York: John Wiley & Sons.

[3] Rousseeuw, P. J. and A. M. Leroy. 1987. *Robust Regression and Outlier Detection.* New York: John Wiley and Sons.

[4] Ryan, P. T. 1997. *Modern Regression Methods*. New York: John Wiley and Sons.
[5] Atkinson, A. C. and M. Riani. 2000. *Robust Diagnostic Regression Analysis*. 1st ed. New York: Springer.
[6] Montgomery, D. C., E. A. Peck and C. G. Vining. 2001. *Probability and Statistics: Introduction to Linear Regression Analysis*. New York: John Wiley and Sons.
[7] Newport, F., L. Saad, and D. W. Moore. 1997. *How Polls are Conducted*. In Where America Stands. M. Golay, editor. New York: John Wiley & Sons.
[8] Cochran, W. G. 1977. *Sampling Techniques*. 3rd ed. New York: John Wiley and Sons.
[9] Judge, G., W. E. Griffiths, R. C. Hill, H. Lutkepohl and T. C. Lee. 1986. *The Theory and Practice of Econometrics*. 2nd ed. New York: John Wiley and Sons.
[10] Rahmatullah Imon, A. H. M. 2003. Regression Residuals, Moments, and Their Use in Tests for Normality. *Communications in Statistics: Theory and Methods*. 32: 1021-34.
[11] Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley and Sons, Inc.