Available online at www.ijo.ir

Iranian Journal of Optimization 2(2010) 536-554

Iranian Journal of Optimization

A statistical test for outlier identification in data envelopment analysis

Morteza Khodabin[1], Reza Kazemi Matin[2]

Islamic Azad University, Karaj Branch, Department of Mathematics

P.O.Box 31485-313, Karaj, Iran.

Abstract

In the use of peer group data to assess individual, typical or best practice performance, the effective detection of outliers is critical for achieving useful results. In these ‘‘deterministic’’ frontier models, statistical theory is now mostly available. This paper deals with the statistical pared sample method and its capability of detecting outliers in data envelopment analysis. In the presented method, each observation is deleted from the sample once and the resulting linear program is solved, leading to a distribution of efficiency estimates. Based on the achieved distribution, a pared test is designed to identify the potential outlier(s). We illustrate the method through a real data set. The method could be used in a first step, as an exploratory data analysis, before using any frontier estimation.

Keywords: Data Envelopment Analysis (DEA), Outlier, Efficiency, Paired Sample Test.

[1] Corresponding author, E-mail: m-kaodabin@kiau.ac.ir

[2] E-mail: rkmatin@kiau.ac.ir

1. Introduction

Nonparametric deterministic frontier models are very appealing because they rely on few assumptions; but, by construction, they are quite sensitive to extreme values and to outliers. Since extreme observations determine the production frontier in DEA models, the estimation of the frontier may be sensitive to measurement errors in the sample data. If an observation has been contaminated with noise that increases the observed output value or decreases the observed input values such that it gets rated as efficient, then it may also enter the reference set of other observations and distort their estimated efficiency scores. Detecting outliers is thus of primary importance: it is not an easy task in this multivariate setup.

Essentially there will always be problems with empirical data either because some decision making units (DMUs) are outliers or do not belong to the dataset. Before meaningful DEA results can be obtained, these outliers must be deal with.

The obvious first step is to double check the data of those DMUs that appear to performing too well, or too poorly. One way to find the former is to check the number of peers that use them as an efficient reference. It will be easy to see if this is much larger than for the other efficient DMUs. Hence, checking data integrity is the first step, followed by either correcting data errors or removing problematic DMUs. Such outliers may be influential in the estimation results obtained using a conventional DEA model. It is desirable, therefore, to consider a procedure that allows us to identify and remove such outliers.

Most of the standard geometrical methods for detecting outliers are very

computer intensive in multivariate set-ups and do not take the frontier aspects of the problem into account: we are mostly interested to detect super-efficient outliers which will be very influential to the efficiency measures and the obtained optimal weights of DEA models. The rest of this paper is organized as follows. A brief literature review is given in the next section. Section 3 is devoted to describe our approach in using paired test in identifying super-efficient outliers. An empirical application to the 42 educational departments of Islamic Azad University, Karaj Branch (IAUK) illustrates the method in section 4. Section 5 concludes.

2. A brief literature review

Many studies have been performed to measure sensitivity or robustness of DEA results and why this is closely related to many techniques for identifying outliers from different points of view in the recent DEA literature. Here we give a brief review to some of these works among the others.

Banker and Gifford (1988) suggested the use of the super-efficiency model to screen out observations with gross data errors, and obtain more reliable efficiency estimates after removing those identified outliers. Banker et al. (1989) applied this method for outlier identification to analyze cost variances for 117 hospitals. The Banker–Gifford method is designed for situations when some observations may be contaminated and, consequently, erroneously classified as efficient. Wilson (1993, 1995) proposed methods making use of influence functions to detect outliers in this framework but the methods become computationally

prohibitive as the number of observations increases. Ondrich and Ruggiero (2002) in their work analyze the resampling technique of jackknifing and its capability of detecting outliers in data envelopment analysis. Simar (2003) describes using a statistical order-m frontier method introduced by Cazals et al. (2002) to detect potential outliers. Banker and Chang (2006) present a super-efficiency based approach to identify and remove outlier units in DEA estimation of efficiency. And finally, in the more recent DEA literature, Johnson and McGinnis (2009) describe an outlier identification methodology by using the inefficient frontier.

In the next section of the present paper, we describe how the statistical pared test can also be used to identify outliers and distinguish the frontier units based on their influences on the distribution of the other units efficiency score.

3. A new super-efficiency method for outlier identification in DEA

A it is most common in the DEA literature, let's suppose that we have n DMUs which DMU_j utilizes inputs x_{ij} for $i = 1,...,m$ to produce outputs y_{rj} for $r = 1,...,s$ and $j = 1,...,n$. Also, let $(\mathbf{x}_j, \mathbf{y}_j) \in \Re^{m \times s}$ denotes the input/output vector of unit "j". In the classic CCR-DEA model by Charnes et al. (1978), it is derived that under the constant return to scale (CRS) assumption the underlying production possibility set (PPS) defined by $T = \{(\mathbf{x}, \mathbf{y}) \mid \mathbf{x} \in \Re_+^m$ can produce $\mathbf{y} \in \Re_+^s\}$ can be written as $T_c = \left\{ (\mathbf{x}, \mathbf{y}) \middle| \mathbf{x} \geq \sum_{j=1}^{n} \mathbf{x}_j \lambda_j ; \mathbf{y} \leq \sum_{j=1}^{n} \mathbf{y}_j \lambda_j ; \boldsymbol{\lambda} \geq 0 \right\}$. We

restrict our attention to the classic Farrell input efficiency measure, defined as $Eff\ (\mathbf{x}_0,\mathbf{y}_0)=\min\{\theta|(\theta\mathbf{x}_0,\mathbf{y}_0)\in T_c\}$, where vector $(\mathbf{x}_0,\mathbf{y}_0)$ refers to the DMU under evaluation. The Farrell efficiency measure for each production possibility $(\mathbf{x}_0,\mathbf{y}_0)$ can be calculated as the solution of following linear programming problem:

$$
\begin{aligned}
& e_o=\min_{\theta,\lambda} \quad \theta \\
& s.t. \quad \sum_{j=1}^{n} x_{ij}\lambda_j^o \le \theta x_{io},\ i=1,...,m \\
& \qquad \sum_{j=1}^{n} y_{rj}\lambda_j^o \ge y_{ro},\quad r=1,...,s \qquad (1) \\
& \qquad\qquad \lambda_j^o \ge 0, \qquad j=1,...,n
\end{aligned}
$$

This linear program is solved once for each observation, $j=1,...,n$ to compute efficiency estimates for that observation. For each producer, the solution of (1) provides not only a non-parametric measure of efficiency, i.e. e_j for $j=1,...,n$, but also a referent production 'possibility' composed of a convex combination of observed production possibilities define by the optimal intensity weights $\boldsymbol{\lambda}_o^*=(\lambda_1^{o*},...,\lambda_n^{o*})$. Therefore, the production possibility set that composes the Farrell efficient referent points can easily be identified.

We use an iterative technique to estimate the differences made by elimination of each observation on the efficiency score of the other units in an attempt to handle outliers. For this purposes the unit under analysis is deleted from the sample and the following linear programming is solved for all of the other observations to

obtain the effect of DMU_o on the frontier of the production set. For unit k, where $k \neq o$ we need the solution of the following LP:

$$
\begin{aligned}
e_k^o = \min_{\theta,\lambda} \quad & \theta \\
s.t. \quad & \sum_{j=1}^{n} x_{ij}\lambda_j^k \le \theta x_{ik}, \qquad i = 1,...,m \\
& \sum_{j=1}^{n} y_{rj}\lambda_j^k \ge y_{rk}, \qquad r = 1,...,s \qquad (2) \\
& \lambda_j^k \ge 0, \qquad j = 1,...,n \,\&\, j \neq o \\
& \lambda_o^k = 0
\end{aligned}
$$

The above model is feasible and bounded and also the added constraint $\lambda_o^k = 0$ deletes the unit "o" from the PPS. The results of this model lead to a distribution of $n-1$ efficiency scores for each observed units. To perform a pared test of sample size n, we also suggest using the *supper efficiency* score, e_o^o, computed by the above model.

Now, including the original Farrell measures as the results of the model (1), we require to solve $n^2 + n$ linear programs. However, based on the properties of the peer set of each observation, by means of the following simple propositions, we conclude that it is not necessary to solve the model (2) for observations which DMU_o is not a member of their peer set in the model (1).

Proposition 1. If $p, q \in \{1,...,n\}$ and $e_q < 1$, i.e. DMU_q be a technically inefficient unit then in solving model (2) we have $e_p^q = e_p$.

Proposition 2. If $p,q \in \{1,...,n\}$ and $\lambda_q^{p*} = 0$ in an optimal solution of the model (1) then in solving model (2) we have $e_p^q = e_p$.

So, we just need to solve the model (2) for a relatively small subset of original observations, i.e. efficient ones.

The required statistical background of paired tests is presented in the appendix in details. To investigate the potentiality of being an outlier for the efficient unit "j" based on this tests, we use two samples of size n; The first is the vector of efficiency scores obtained by solving model 1 for each observations and the second, $\mathbf{e^j}$, the vector consist of the results of model 2, i.e. e_k^j , for the efficient unit "j" where $k = 1,...,n$. The meaningful differences between these columns shows the influence of the super-efficient unit "j" on the efficiency score of the other units, our suggested criteria to detect potential outliers, which is described in the next section with an empirical study.

4. Outliers identification in an empirical study

Now we turn to illustrate proposed technique for outlier detection by applying it to the real-world data of 42 university departments of IAUK. These data are used for the internal performance assessment by the university. The input variables are the number of post graduate students (x_1), the number bachelor students (x_2), and the number of master students (x_3). The output variables are the number of graduations (y_1), the number of scholarships (y_2), the number of research

products (y_3), and the level of manager satisfaction (y_4). Note that all variables have integer structure and y_4 is an ordinal variable. Table 1 presents some descriptive statistics about the data.

Table1: Descriptive statistics of 42 units

Variables	Min	Max	Mean	Median	St. Dev
# post graduate students (x_1)	0	484	78.55	0	137.4
# bachelor students (x_2)	0	1202	394.6	304.5	360.3
# master students (x_3)	0	535	26.86	0	82.28
# graduations (y_1),	32	1158	385.4	327	297.3
# scholarships (y_2)	0	14	2.31	1	3.181
# research products (y_3)	0	12	1.905	1	2.783
manager satisfaction (y_4)	1	4	2.595	3	0.8851

The input oriented radial efficiency scores obtained by applying the model (1) are presented as the first column of Table 2 which shows that departments 14, 16, 17, 18, 19, 35, 36 and 37 are the efficient units in this evaluation. Now as a measure of influence of these units over the efficiency scores of the other units, the results of eliminating each efficient unit on the shape of production set are obtained through the optimal value of the model (2) and presented in the rest columns of

Table 2. For these computations we used EMS software[3] version 1.3.0 on an Intel (R), 512 Mbytes RAM, 1.73 GHz Laptop computer. The computational times were negligible for all LP programs.

Table 2: Efficiency scores before and after eliminating the efficient units

DMUj	e_j	$\mathbf{e^{14}}$	$\mathbf{e^{16}}$	$\mathbf{e^{17}}$	$\mathbf{e^{18}}$	$\mathbf{e^{19}}$	$\mathbf{e^{35}}$	$\mathbf{e^{36}}$	$\mathbf{e^{37}}$
1	0.8852	0.8852	0.8856	1.0000	0.8863	0.8852	0.8852	0.8852	0.8852
2	0.9564	0.9866	0.9564	1.0000	0.9564	0.9564	0.9564	0.9564	0.9564
3	0.9398	0.9634	0.9398	0.9678	0.9398	0.9398	0.9398	0.9398	0.9398
4	0.9405	0.9633	0.9405	0.9718	0.9405	0.9405	0.9405	0.9405	0.9405
5	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
6	0.9168	0.9168	0.9168	0.9168	0.9168	0.9168	0.9478	0.9168	0.9168
7	0.8709	1.0000	0.8709	0.8709	0.8709	0.8709	0.8709	0.8709	0.8709
8	0.5378	0.5378	0.5378	0.6331	0.5378	0.5378	0.5378	0.5378	0.5378
9	0.9285	0.9285	0.9285	0.9677	0.9285	0.9285	0.9285	0.9285	0.9285
10	0.9017	0.9017	0.9017	0.9263	0.9017	0.9017	0.9017	0.9017	0.9017
11	0.7727	0.9242	0.7727	0.7727	0.7727	0.7727	0.7727	0.7727	0.7727
12	0.2711	0.2711	0.2711	0.3046	0.2711	0.2711	0.2711	0.2711	0.2711
13	0.8823	0.8823	0.9028	0.9297	0.8883	0.8823	0.8823	0.8823	0.8823
14	1.0000	1.1961	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
15	0.7579	0.7579	0.7584	0.8571	0.7719	0.7579	0.7579	0.7579	0.7579
16	1.0000	1.0000	1.0459	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
17	1.0000	1.0000	1.0000	4.7518	1.0000	1.0000	1.0000	1.0000	1.0000
18	1.0000	1.0000	1.0000	1.0000	1.0250	1.0000	1.0000	1.0000	1.0000
19	1.0000	1.0000	1.0000	1.0000	1.0000	1.0018	1.0000	1.0000	1.0000
20	0.8923	0.8923	0.8923	0.8923	0.8923	0.8923	0.8923	0.8923	0.8923
21	0.8796	0.8796	0.8888	0.8799	0.8885	0.8796	0.8796	0.8796	0.8796
22	0.8742	0.8742	0.8817	0.8744	0.8858	0.8742	0.8742	0.8742	0.8742
23	0.8396	0.8396	0.8426	0.8563	0.8563	0.8396	0.8396	0.8396	0.8396
24	0.7569	0.7569	0.7570	0.7655	0.7711	0.7569	0.7569	0.7569	0.7569
25	0.7388	0.7388	0.7419	0.7391	0.7496	0.7388	0.7388	0.7388	0.7388
26	0.8901	0.8901	0.8901	0.9169	0.9071	0.8901	0.8901	0.8901	0.8901
27	0.9990	0.9990	1.0000	1.0000	1.0000	0.9990	0.9990	0.9990	0.9990
28	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

[3] Efficiency Measurement System, by Holger Scheel (H.Scheel@wiso.uni-dortmund.de)

29	0.6181	0.6181	0.6181	0.7957	0.6181	0.6181	0.6181	0.6487	0.6181
30	0.6217	0.6217	0.6217	0.9775	0.6217	0.6217	0.6217	0.6217	0.6217
31	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
32	0.7761	0.7761	0.7761	0.7947	0.7761	0.7761	0.7761	0.7761	0.7761
33	0.8163	0.8163	0.8163	0.8163	0.8163	0.8163	0.8446	0.8163	0.8163
34	0.9711	0.9711	0.9722	1.0000	0.9711	0.9711	0.9711	1.0000	0.9711
35	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0757	1.0000	1.0000
36	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.1744	1.0000
37	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	5.4828
38	0.9215	0.9267	0.9215	1.0000	0.9215	0.9215	0.9215	0.9215	0.9215
39	0.3640	0.3665	0.3640	0.4016	0.3686	0.3640	0.3640	0.3640	0.3640
40	0.9227	0.9710	0.9227	0.9227	0.9227	0.9227	0.9227	0.9227	0.9227
41	0.7772	0.7772	0.7772	0.7772	0.7772	0.7772	0.7772	0.7772	0.7772
42	0.7288	0.7288	0.7288	0.7288	0.7288	0.7288	0.7288	0.7288	0.7288

Since, using parametric statistical methods depends on normality condition of observations distribution, it is necessary to apply normality tests for the obtained data in the first step; here we use Kolmogorov-Smirnov empirical distribution function and visual tests. The results are summarized in the table 3

Table 3: Kolmogorov-Smirnov Test

		EFF	E14	E16	E17	E18	E19	E35	E36	E37
N		42	42	42	42	42	42	42	42	42
Normal Parameters(a,b)	Mean	.855943	.870450	.858140	.976410	.859060	.855986	.859157	.861512	.962676
	Std. Deviation	.1695478	.1784681	.1707749	.6163929	.1694246	.1695853	.1717203	.1747358	.7339701
Most Extreme Differences	Absolute	.202	.210	.198	.461	.206	.202	.194	.190	.456
	Positive	.198	.210	.179	.461	.179	.195	.182	.190	.456
	Negative	-.202	-.199	-.198	-.273	-.206	-.202	-.194	-.188	-.256
Kolmogorov-Smirnov Z		1.308	1.362	1.281	2.987	1.332	1.307	1.257	1.233	2.955
Asymp. Sig. (2-tailed)		.065	.049	.075	.000	.057	.066	.085	.096	.000

The significance probability is a real number that is determined by the sample, e.g. $Sig = 0.085$ in the last row of E35 column. In the statistical hypothesis testing we will reject H_0 if and only if $Sig < \alpha$, where α is the significance level that is fixed and known to the researcher and will takes the values like $\alpha = 0.05$.

The last row of above table shows that all units, except DMU_{17} and DMU_{37} follow the normality distribution. More diagnostic test for lack of normal fitting on these units are shown in the figures 1-4 bellow.

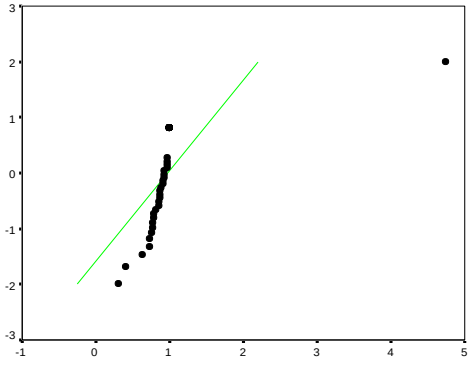

Figure 1: Q-Q plot for DMU_{17}

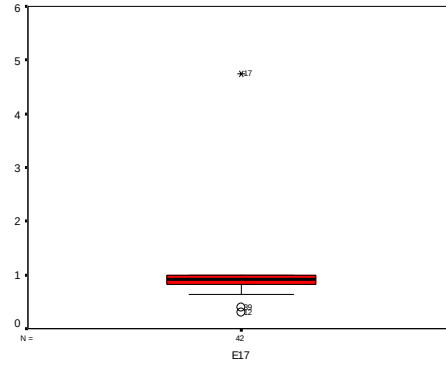

Figure 2: Box plot for DMU_{17}

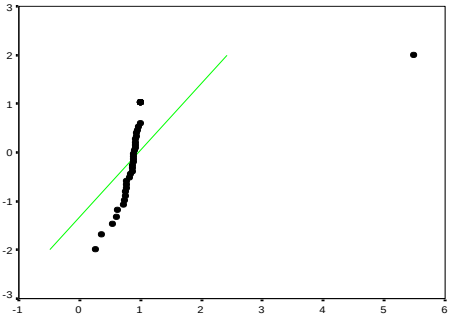

Figure 2: Q-Q plot for DMU_{37}

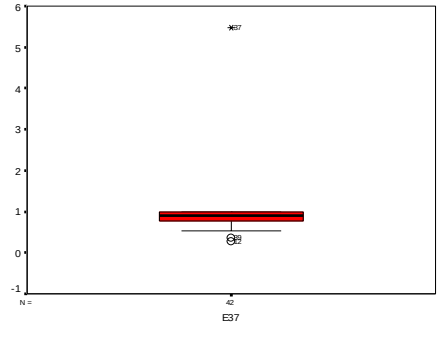

Figure 1: Box plot for DMU_{37}

For Q-Q plot if the selected unit matches the normal distribution, the points cluster around a straight line, whereas we don't see it here. The box plots show another useful visualization for viewing how the data are non-symmetric and hence non normal distribution.

So, except these two units, we can use parametric and nonparametric paired sample tests for detection of outliers. In the below table we show the results of parametric paired-samples t test

Table 4: Parametric paired sample t test results

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	EFF - E14	-.014507	.0423660	.0065372	-.027709	-.001305	-2.219	41	.032
Pair 2	EFF - E16	-.002198	.0077873	.0012016	-.004624	.000229	-1.829	41	.075
Pair 3	EFF - E18	-.003117	.0061927	.0009556	-.005046	-.001187	-3.262	41	.002
Pair 4	EFF - E19	-.000043	.0002777	.0000429	-.000129	.000044	-1.000	41	.323
Pair 5	EFF - E35	-.003214	.0131209	.0020246	-.007303	.000874	-1.588	41	.120
Pair 6	EFF - E36	-.005569	.0274459	.0042350	-.014122	.002984	-1.315	41	.196

Based on the significance probabilities values 0.032 and 0.002 for units 14 and 18, we can reject H_0 and conclude that in comparison to other units, these units have more capacity to be outlier.

Table 5: Non-parametric Wilcoxon signed ranks test

	E14 - EFF	E16 - EFF	E17 - EFF	E18 – EFF	E19 - EFF	E35 – EFF	E36 - EFF	E37 – EFF
Z	-2.666(a)	-2.934(a)	-4.286(a)	-3.059(a)	-1.000(a)	-1.604(a)	-1.604(a)	-1.000(a)
Asymp. Sig. (2-tailed)	.008	.003	.000	.002	.317	.109	.109	.317

Table 6: Non-parametric sign test

	E14 - EFF	E16 - EFF	E17 – EFF	E18 - EFF	E35 - EFF	E36 – EFF
Exact Sig. (2-tailed)	.004	.001	.000	.000	.250	.250

The non-parametric related sample tests for 17 and 37 units are shown in Tables 5 and 6.

In these two non-parametric tests, the results obtained for the significance probabilities of the units 14, 16, 17 and 18 are less than 0.05 and using parametric results we conclude that DMU_{14}, DMU_{17} and DMU_{18} can be considered as a potential outliers.

5. Conclusions

This paper presents a statistical pared sample t test to identifying outlier as the first step in using any DEA models. It is shown that the presented approach is a

powerful tool to remove any potential super-efficient outliers. An empirical study to an Iranian university department's further illustrated the importance of removing outliers in DEA estimation of production function.

Acknowledgement

This research was supported by the Islamic Azad University, Karaj Branch.

References

[1] Banker RD., Gifford JL., "A relative efficiency model for the evaluation of public health nurse productivity", Mellon University Mimeo, Carnegie, 1988.

[2] Banker RD., Chang H., "The super-efficiency procedure for outlier identification, not for ranking efficient units", European Journal of Operational Research, 175, 1311–1320, 2006.

[3] Corder G.W., Foreman D.I., "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach", New Jersey: Wiley, 2009.

[4] Gouri K., Bhattacharyya R., Johnson A., "Statistical concepts and methods". John Wiley & Sons, 1977.

[5] Johnson AL., McGinnis LF., "Outlier detection in two-stage semiparametric DEA models", European Journal of Operational Research, forthcoming, 2009.

[6] Ondrich J., Ruggiero J., "Outlier detection in data envelopment analysis: an analysis of jackknifing". Journal of the Operations Research Society, 53, 342-346, 2002.

[7] Simar L., "Detecting outliers in frontier models: A simple approach", Journal of Productivity Analysis. 20, 391–424, 2003.

[8] Wilcoxon F., "Individual comparisons by ranking methods". Biometrics, 1, 80-83, 1945.

[9] Wilson PW., "Detecting outliers in deterministic nonparametric frontier models with multiple outputs", Journal of Business and Economic Statistics 11, 319–323, 1993.

[10] Wilson PW., "Detecting influential observations in data envelopment analysis", Journal of Productivity Analysis 6, 27–45, 1995.

Appendix 1: Proofs

Proposition 1. If $p,q \in \{1,...,n\}$ and $e_q < 1$, i.e. DMU_q be a technically inefficient unit, then in solving model (2) we have $e_p^q = e_p$.

Proof. Since DMU_q is an inefficient unit then in any optimal solution of the model (1) for evaluating DMU_p we have $\lambda_q^p = 0$. Otherwise by the complementary slackness property of linear programming, the corresponding dual constraint will be bind. This means that in its multiplier dual form, we have a set of feasible weight by which the unit q reaches the maximum possible efficiency score 1. This contradicts with inefficiency of this unit.

Proposition 2. If $p,q \in \{1,...,n\}$ and $\lambda_q^{p^*} = 0$ in an optimal solution of the model (1) then in solving model (2) we have $e_p^q = e_p$.

Proof. It is easy to verify that both model have the same optimal solution, hence their optimal values are equal.

Appendix 2: The statistical preliminaries

Paired Test

There are two reasons for using a paired design: reduction of bias and increased precision. Both reasons may be true at once. Two measurements are paired when they come from the same observational unit: the efficiency scores obtained before and after elimination of the units.

Pairing seeks to reduce variability in order to make more precise comparisons with fewer subjects. When independent samples are used, the difference between treatment means is compared to the variability of individual responses within each treatment group. This variability has two components:

The larger component is usually the variability between subjects (between-subject variability). It's there because not every subject will respond the same way to a particular treatment. There will be variability between subjects.

The other component is within-subject variability. This variability is present because even the same subject doesn't give exactly the same response each time s/he is measured. There will be variability within subjects.

When both measurements are made on the same subject, the between-subjects variability is eliminated from the comparison. The difference between treatments is compared to the way the difference changes from subject to subject. If this difference is roughly the same for each subject, small treatment effects can be detected even if different subjects respond quite differently. If measurements are made on paired or matched samples, the between-subject variability will be reduced according to the effectiveness of the pairings. The pairing or matching need not be perfect. The hope is that it will reduce the between- subject variability enough to justify the effort involved in obtained paired data.

Sometimes pairing turns out to have been a good idea because variability is greatly reduced. Other times it turns out to be having been a bad idea, as is often the case with matched samples. Pairing has no effect on the way the difference between two treatments is estimated.

The estimate is the difference between the sample means, whether the data are paired or not. What changes is the uncertainty in the estimate.

- **Parametric Test**

The paired t test statistic is the difference between the paired observations, which is symbolized by d , which in our purposes denotes the difference between the efficiency scores of the units before and after the eliminations and $\bar{d}$ is average difference. Note that $\bar{d}$ has the same value as the difference between the means of the two samples (ave. after elimination minus ave. before elimination).The mean of the differences is the same as the difference between the means $\bar{d} = \mu_1 - \mu_2$. One can also calculate the standard deviation of d . The sample size (n) is simply the number of paired observations. Here we use null-hypothesis (H_0) for the case the unit under consideration does not appear as an outlier and alternative-hypothesis denotes by H_a:

$$\begin{cases} H_0 : \bar{d} = 0 \\ H_a : \bar{d} \neq 0 \end{cases}$$

The test statistics that we can use is $t_s = \dfrac{\bar{d}}{SE_{\bar{d}}}$ with $(n-1)$ degrees of freedom.

P-value (P_{value}: significance probability) is the probability of being wrong

(committing a type I error) if one rejects H_0 and the confidence interval is $\bar{d} \pm t_{a,df}\,(SE_{\bar{d}})$. We note that the d 's must be distributed normally when the sample size is small. This assumption is relaxed as the sample size gets large due to the effect of the central limit theorem.

- **Non-Parametric Test**

Occasionally, the assumptions of the t-tests are seriously violated. In particular, if the type of data you have is ordinal in nature and not at least interval. On such occasions an alternative approach is to use nonparametric tests. We are not going to place much emphasis on them in this unit as they are only occasionally used. But we should be aware of them and have some familiarity with them. In our purpose, when the units have both cardinal and ordinal data and we want detect the outliers.

Nonparametric tests are also referred to as distribution-free tests. These tests have the obvious advantage of not requiring the assumption of normality or the assumption of homogeneity of variance. They compare medians rather than means and, as a result, if the data have one or two outliers, their influence is negated.

- **Signed-Rrank Test**

This test is useful when sample size is small and no particular distribution is assumed and there are real doubts about whether or not a t-test can be used. It is based on the sign of the difference between paired observations. Let's restate the null and alternative hypotheses so that we are clear:

$$\begin{cases} H_0 : \bar{d} = 0 \\ H_a : \bar{d} \neq 0 \end{cases}$$

If H_0 is true, then the error between the observations is random. If this is so, then there should be an equal chance of getting a positive difference or a negative difference. That is, we expect half of the signs to be + and half to be -. The signs test is based on this assumption and the binomial distribution.

- **Wilcoxson Signed-Rank Test**

This uses a bit more information than does the signs test, so it is a bit more powerful.

To do this test, rank the d 's from smallest to largest (based on their absolute value).

Restore the + and - signs. Add the negative ranks and take their absolute value. Add the positive ranks. The test statistic (W_s) is whichever is the larger of the two sums above. There are few assumptions to use the Wilcoxson signed ranks test. If it is reasonable to assume that the d 's are as likely to be positive as negative and no particular distribution is assumed.