Journal of Statistics: Advances in Theory and Applications
Volume 5, Number 2, 2011, Pages 131-147

ESTIMATION METHODS FOR MISSING DATA IN UN-REPLICATED 2^k FACTORIAL AND 2^{k-p} FRACTIONAL FACTORIAL DESIGNS

MAHER QUMSIYEH and KRAIG KIRCHNER

Department of Mathematics
University of Dayton
300 College Park
Dayton OH 45469-2316
USA
e-mail: qumsiyeh@notes.udayton.edu

Abstract

In 2^k factorial and 2^{k-p} fractional factorial designs, each effect is dependent on every observation. Therefore, missing observations in factorial designs can drastically alter these effects. To restore the orthogonal structure to the design, estimation methods are needed. After estimation, the half-normal plot for the effects needs to be examined. If insignificant effects approximately point toward the origin, then we have a proper estimation. Current, popular estimation methods sacrifice effects in order to calculate missing observations. In this paper, we have attempted new estimation methods without explicitly sacrificing effects that seem to work well.

1. Introduction

A variety of factors contributes to missing or bad data in experimental designs. Some of these can be machine breakdowns,

2010 Mathematics Subject Classification: 62K15.

Keywords and phrases: un-replicated factorial and fractional factorial experiments, missing values, effect estimation.

Received June 13, 2011

damage to experimental units, faulty readings from measurement systems. Omitting the data is not a good option with this type of experimental design, because it can cause havoc to your model. Therefore, we need a method to estimation missing or bad data.

Large factorial designs require an exponential amount of experimental runs. A company may realize that an experiment is exceeding their budget during the experimentation process. Therefore, sometimes it may be impossible to finish an experiment, but there is still a need to find something from the experiment. Estimating missing data values could be an alternative to potentially salvage an experiment and save a company a lot of money. Even in small experiments, may be we realize post-research that conditions for a particular treatment combination did not remain constant. An outlier or bad data resulted from this particular combination. This data could greatly skew all of the effects in the experiment. In order to salvage the data and find something useful from the experiment, we can use a variety of methods to estimate the outcome of this particular treatment combination.

Several papers exist on estimating missing data. Draper and Stoneman [8] give a method to estimate the missing values, but their method depends on sacrificing some of the effects to estimate the missing values. John proposes a method similar to Draper and Stoneman. Wilkinson [17, 18] gives a method that can require considerable computations. Shearer [14] give a new procedure to use with factorial designs using an iterated method and convergence of such iteration, Qumsiyeh and Shaughnessy [13] proposed the bootstrap for estimating missing responses.

In this paper, we will outline two common methods for estimating missing data, the first introduced by Draper and Stoneman [8] and the second by John [9]. These methods look to minimize contrasts/effects in order to estimate missing data points. We will introduce alternative methods that do not require sacrificing contrasts/effects.

2. Estimation Methods for Missing Data

2.1. Draper and Stoneman

In order to tell if a factor or interaction is significant, we need to calculate contrasts and effects. The contrast is the total difference between responses where a factor is at the high level and responses where the factor is at the low level. In order to calculate this contrast relatively easily, we only need to think of it as product of the vector $y^t = (y_1, y_2, \dots, y_n)$ of the responses with the column of that factor say 1_i, where 1_i is vector of -1 or $+1$ elements. The Draper and Stoneman method estimates a missing data point by setting the highest order interaction equal to zero. Note that $1_i^t 1_j = 0$ for $i \neq j.$ This is the same as minimizing the absolute value of highest order contrast. The effect of a factor is the contrast divided by 2^{k-1}. Because the contrast is the numerator of the effect, this is also the same as minimizing the absolute value of the effect of the highest order interaction. If we have to sacrifice an effect to estimate a missing data point, it is very common to sacrifice the highest order interaction. Empirically, it is common that this contrast is relatively close to zero. This is a very popular method. However, we find it hard to fully buy in to this method because it is possible for any experiment to produce a significant effect with the highest order contrast.

Before estimation, Draper and Stoneman employ a method to detect a bias in the design. Without a missing response, if the insignificant effects in a half-normal plot do not point toward the origin, then there is a bias in the design. Research then needs to be used to find which point is causing this bias. Once this response is determined, they throw it out and use an estimation method. After estimation, they examine the half-normal plot. If insignificant effects in the new half-normal plot point toward the origin, then they have a proper estimation.

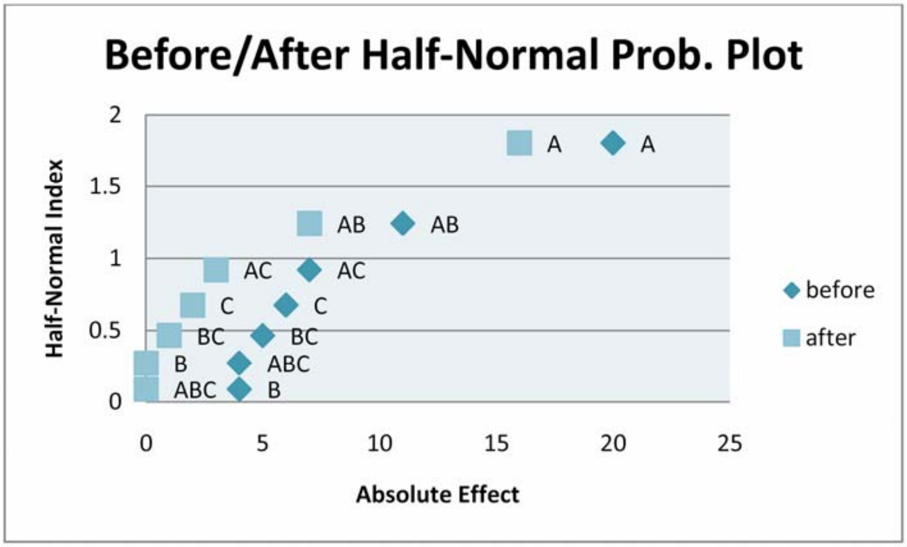
In their paper, Draper and Stoneman produce two specific examples. They look at a single replicate of a 2^3 factorial design. Responses in standard order are 10, 16, 2, 22, 8, 20, 2, and 44. The standard order of responses in a 2^3 factorial design, results in a nice grid of factors. The standard order here is (–), *a, b, ab, c, ac, bc,* and *abc*, where, for example, *ac* means both A and C are at the +1 setting. The response value 44 was produced under different circumstances than the rest of the responses, so it is deemed the bad data point causing the bias in the design. Setting the highest order interaction contrast equal to zero, results in an estimation of 28 instead of 44.

Experimental grid:

	A	B	C	Y
(–)	–1	–1	–1	10
a	1	–1	–1	16
b	–1	1	–1	2
ab	1	1	–1	22
c	–1	–1	1	8
ac	1	–1	1	20
bc	–1	1	1	2
abc	1	1	1	44

The half-normal plots before and after the estimation are given in the following table (Table 1):

Table 1



As you can see, the insignificant effects before estimation form a line that does not point toward the origin. However, after the bad data point was thrown out and an estimation method was employed, the significant effects form a line that approximates toward the origin slightly better. Therefore, this is a proper estimation.

2.2. Peter John

John [9] publishes a method that is a slight variation of the Draper and Stoneman method. This method involves minimizing the sum of squared contrasts for the highest order interaction and the $k - 1$ ordered interactions. John himself says that this method should only be used at times when the Draper and Stoneman method does not produce a proper estimation. Reasoning for this seems to be that we are sacrificing multiple effects.

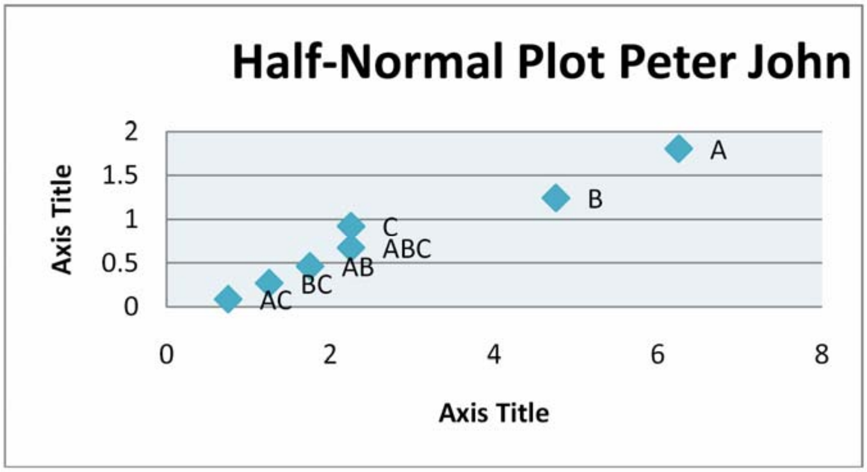
The Draper and Stoneman method only sacrifices one effect. Like Draper and Stoneman, John shows an example of a 2^3 factorial design. John starts out with an experimental design that has a missing data point. No bad data detection methods are used by John. Standard order of responses with missing point labelled as x is 23, 26, 25, 36, 25, 31, x, and 34. After minimizing the sum of squared contrasts for the single 3-way interaction and the three 2-way interactions, John finds an estimation of 29.

Experimental grid:

	A	B	C	Y
(–)	–1	–1	–1	23
a	1	–1	–1	26
b	–1	1	–1	25
ab	1	1	–1	36
c	–1	–1	1	25
ac	1	–1	1	31
bc	–1	1	1	x
abc	1	1	1	34

The half-normal plot after the estimation is given in the following table (Table 2):

Table 2



As you can see, the insignificant absolute effects of the half-normal plot form a line that approximately points toward the origin.

3. New Estimation Methods

Initially, to come up with a new estimation method, we look at particular circumstances where these methods are impossible to employ. For instance, in a half-fractional factorial design, the highest order interaction is the generator factor. Also, the $(k-1)$-way interactions are aliases with main effects. Therefore, we would not want to sacrifice those by minimizing their contrasts. Therefore, the previous methods of Draper and Stoneman and of John cannot be used. So, we need to come up with something different from the other methods. The following estimation methods provide an alternative.

For one missing response, estimation is done by simple average of all responses that share $k-1$ levels with the missing response, as we will see later.

For two missing responses, there are two different cases. If the missing responses do not share $k-1$ levels, then we can estimate each missing response by simple average of those responses that share $k-1$ levels with the missing response. This is the same as estimation for one missing response.

However, if the missing responses share $k-1$ levels, then we are missing a necessary response to estimate by a simple average. To simplify this method, assume X is the first missing response and Y is the second missing response. We should initially estimate X by simple average of all available points that share $k-1$ levels with X. Then, estimate Y by simple average of all data points that share $k-1$ levels with Y including the new X. Then re-estimate X by simple average of all responses that share $k-1$ levels including the Y. Because this method is determined by which missing response is calculated first, we should repeat this process starting with an estimation of Y and then following the same process. Now, we should have two estimates for each missing data point. Simple average of these estimates should provide a solid estimation for each missing response. We will illustrate this with an example later.

The following is an example of a 2^4 factorial design. This example was introduced by Yin and Jullie [19]. In this example, the response variable of interest is the etch rate for silicon nitride. Experimenters have come up with four factors that they believe affect the response variable.

Each of the four factors listed below with the unit of measure and high/low levels:

- Factor A: Anode cathode gap, low-0.8cm high-1.2cm.
- Factor B: Pressure in the reactor chamber, low-4.5mTorr high-550mTorr.
- Factor C: C_2F_6 gas flow, low-125 SCCM high-200 SCCM.
- Factor D: Power applied to the cathode, low-275W high-325W.

After running each of the treatment combinations, we create a grid of combinations.

Experimental grid:

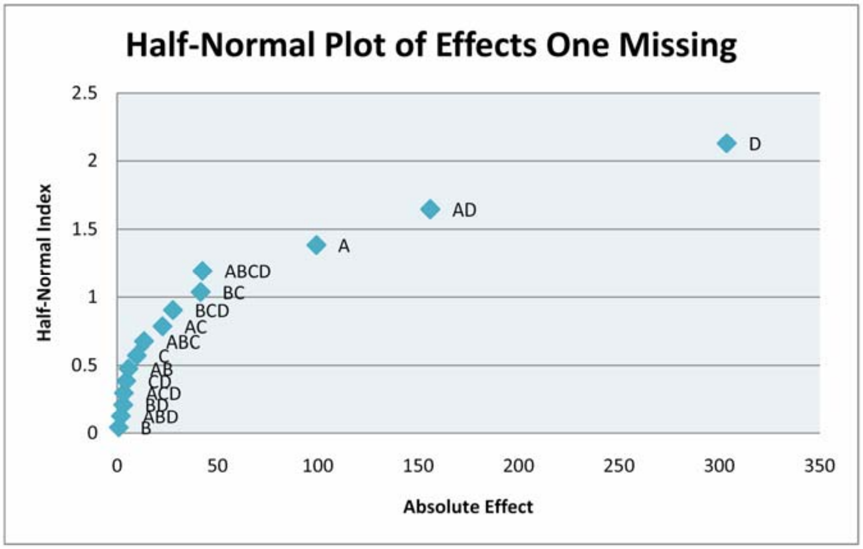
	A	B	C	D	Y
(–)	–1	–1	–1	–1	550
a	1	–1	–1	–1	669
b	–1	1	–1	–1	604
ab	1	1	–1	–1	650
c	–1	–1	1	–1	633
ac	1	–1	1	–1	642
bc	–1	1	1	–1	601
abc	1	1	1	–1	635(---)
d	–1	–1	–1	1	1037
ad	1	–1	–1	1	749
bd	–1	1	–1	1	1052
abd	1	1	–1	1	868
cd	–1	–1	1	1	1075
acd	1	–1	1	1	860
bcd	–1	1	1	1	1063
abcd	1	1	1	1	729

Using the half-normal plot, it was determined that factors A, D and their interaction AD are the active factors (factors that have an effect on the response).

3.1. One missing response in a 2^4 factorial design

Let us assume that the response 635 for the *abc* setting is missing. Since this design is complete, the missing response was chosen randomly. Again, the response variable of interest is etch rate for silicon nitride. The treatment combinations that share exactly $k - 1$ levels with this response (*abc*) are *ab, ac, bc,* and *abcd*. After we take a simple average of these responses, we have 653.75 as an estimate for the missing response. Now that we have our estimates, let us look at the half-normal plot of absolute effects (Table 3).

Table 3



Insignificant absolute effects form a line that approximately points toward the origin. Therefore, we have a proper estimation; also, factors A, D and their interaction AD are the active factors, which is the same as the case with no missing responses.

3.2. Two missing responses in a 2^k factorial design

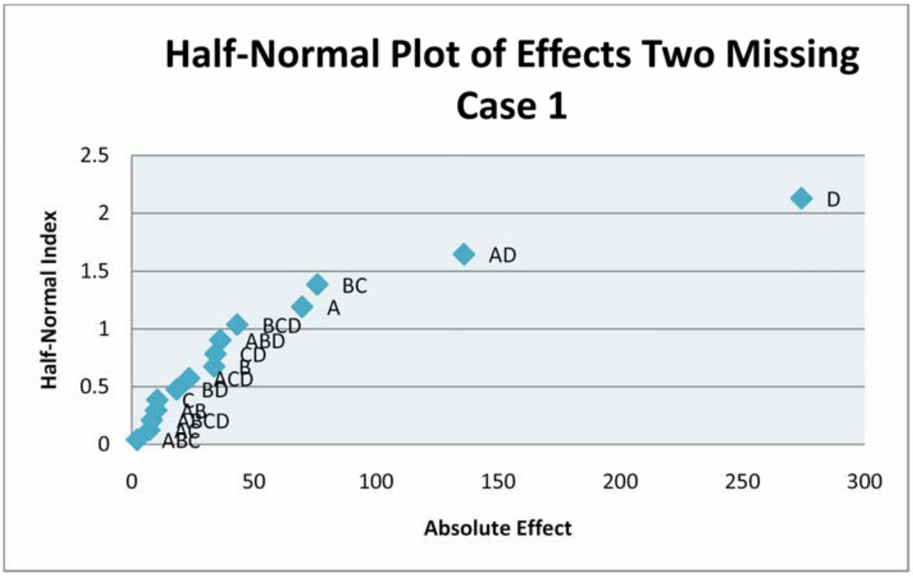
Case I. Two missing data points in a 2^4 factorial design that do not share $k - 1 = 3$ levels.

Experimental grid:

	A	B	C	D	Y
(–)	–1	–1	–1	–1	550
a	1	–1	–1	–1	669
b	–1	1	–1	–1	604
ab	1	1	–1	–1	650
c	–1	–1	1	–1	633
ac	1	–1	1	–1	---
bc	–1	1	1	–1	601
abc	1	1	1	–1	635
d	–1	–1	–1	1	1037
ad	1	–1	–1	1	749
bd	–1	1	–1	1	1052
abd	1	1	–1	1	868
cd	–1	–1	1	1	1075
acd	1	–1	1	1	860
bcd	–1	1	1	1	---
abcd	1	1	1	1	729

Assume we have two missing responses in a 2^k factorial design. Again, for this particular problem, we are going to use the design previously introduced as an example. Since this design is complete, we are going to assume two responses are missing. Responses *ac* and *bcd* are randomly selected as missing responses. The treatment combinations that share exactly 3 levels with response *ac* are *c, abc*, *a*, and *acd*. The treatment combinations that share exactly 3 levels with response *bcd* are *bc*, *cd*, *bd*, and *abcd*. These do not have any common combination. Taking a simple average of these responses for *ac*, we have 699.25 as an estimate for the missing response of *ac*. In addition, taking a simple average of these responses for *bcd*, we have 864.25 as an estimate for the missing response of *bcd*. Now that we have our estimates, let us look at the half-normal plot of absolute effects (Table 4).

Table 4



Insignificant absolute effects form a line that approximately points toward the origin. Therefore, we have a proper estimation. Factors A, D and their interaction AD are the active factors (A is slightly significant so it needs to be examined in an additional experiment), this is the same as the case with no missing responses.

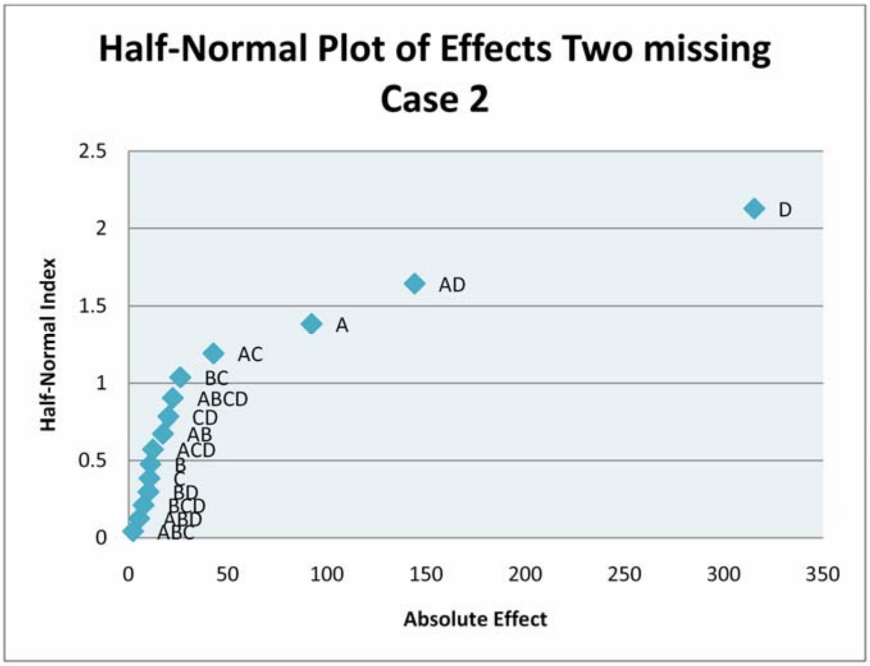
Case II. Two missing data points in a 2^4 factorial design that share $k - 1 = 3$ levels.

Experimental grid:

	A	B	C	D	Y
(–)	–1	–1	–1	–1	550
a	1	–1	–1	–1	669
b	–1	1	–1	–1	604
ab	1	1	–1	–1	650
c	–1	–1	1	–1	633
ac	1	–1	1	–1	642
bc	–1	1	1	–1	601
abc	1	1	1	–1	635
d	–1	–1	–1	1	1037
ad	1	–1	–1	1	---
bd	–1	1	–1	1	1052
abd	1	1	–1	1	868
cd	–1	–1	1	1	1075
acd	1	–1	1	1	---
bcd	–1	1	1	1	1063
abcd	1	1	1	1	729

Responses *ad* and *acd* that used to be 749 and 860 are randomly selected as missing responses. The treatment combinations that share exactly 3 levels with response *ad* are *d, abd, a,* and *acd* (note that *acd* is among those). The treatment combinations that share exactly 3 levels with response *acd* are *ac, ad, cd,* and *abcd* (note that *ad* is among those too). As mentioned previously, we have a step procedure for estimation here. Estimate the first missing response by simple average of all three available responses that share 3 levels with the missing response. Then, estimate the second missing observation by simple average of all responses that share 3 levels with the missing response including the previous estimation. Finally, re-estimate the first missing observation by simple average of all responses that share 3 levels with the missing response including the second estimation. Repeat the procedure again by estimating the second missing response first. Now, we have two estimations for each response. After we take a simple average of these estimates, we have 858 as an estimate for response *ad* and we have 826 as an estimate for response *acd*. Now that we have our estimates, let us look at the half-normal plot of absolute effects (Table 5).

Table 5



Insignificant absolute effects form a line that approximately points toward the origin. Therefore, we have a proper estimation; also, factors A, D and their interaction AD are the active factors, which is the same as the case with no missing responses.

3.3. One missing response in a 2^{5-1} fractional factorial design

Experimental grid:

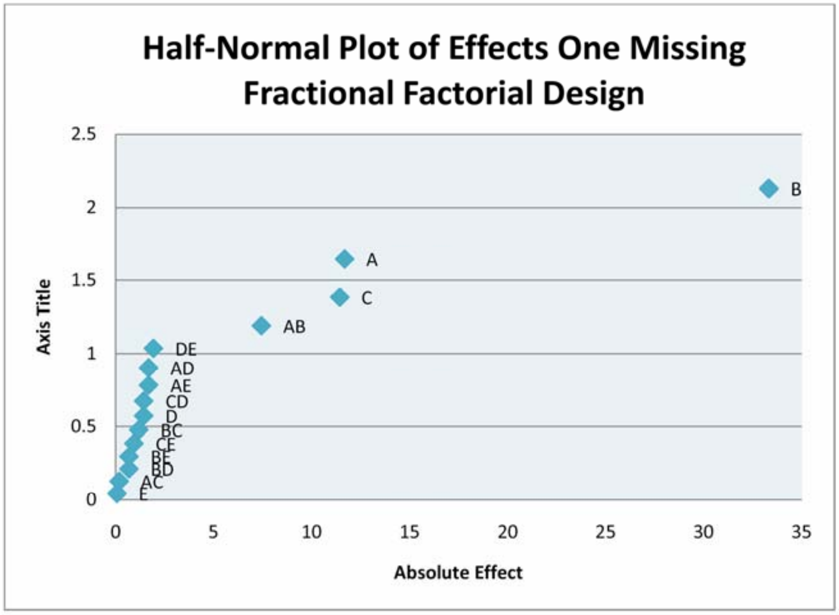
	A = BCDE	B = ACDE	C = ABDE	D = ABCE	E = ABCD	R
e	–1	–1	–1	–1	1	8
a	1	–1	–1	–1	–1	9
b	–1	1	–1	–1	–1	34
abe	1	1	–1	–1	1	52
c	–1	–1	1	–1	–1	16
ace	1	–1	1	–1	1	22
bce	–1	1	1	–1	1	45
abc	1	1	1	–1	–1	60
d	–1	–1	–1	1	–1	6
ade	1	–1	–1	1	1	10
bde	–1	1	–1	1	1	---
abd	1	1	–1	1	–1	50
cde	–1	–1	1	1	1	15
acd	1	–1	1	1	–1	21
bcd	–1	1	1	1	–1	44
abcde	1	1	1	1	1	63

This is a specific 2^{5-1} fractional factorial design. This is a half-fractional factorial design, therefore, we only have half of the observations and each effect is aliased with another effect. Unfortunately, for any missing response, we cannot estimate by simple average of all responses that share $k - 1$ levels with the missing response, because not all of these responses, for any possible missing observation, are in the fraction of responses we have. Therefore, we can estimate a missing response by simple average of all responses that share $k - 2$ levels with the missing response; this is explained in the next paragraph.

Assume we have one missing response in the 2^{5-1} fractional factorial design. We cannot use our original 2^4 factorial design example for this problem. This fractional factorial design was produced to see, if specific factors influenced production yield. The five factors were aperture setting

A, exposure time B, development time C, mask dimension D, and etch rate E. Half of the treatment combinations were run corresponding to the generator factor I = ABCDE. All combinations, where factor ABCDE was set at the higher level were used in the design. Response *bde* was randomly selected as missing observations. All responses that share $k - 1$ levels with *bde* are *bd, be, de, abde,* and *bcde.* These combinations are either have two factors at the higher level or four factors at the higher level. However, treatment combinations in our design have one, three, or five factors at the higher level. Any combination that is missing will not share $k - 1$ levels with the missing response. We introduce a method to estimate the missing response by simple average of all responses that share $k - 2$ levels with the missing response. *bde* shares exactly $k - 2$ levels with *b, d, e, abe, bce, ade, abd, cde,* and *bcd*. The simple average of all these responses results in a estimation of 25.56 for *bde*. Now that we have our estimates, let us look at the half-normal plot of absolute effects (Table 6).

Table 6



Insignificant absolute effects form a line that approximately points toward the origin. Therefore, we have a proper estimation, with A, B, C and the interaction of A and B are the active factors.

4. Conclusion

In all cases, examination of half-normal plots show that insignificant effects form a line that approximately points through the origin. Therefore, we have produced proper estimations. At this point, we can attempt to learn something useful from the experiment. This may not work perfectly for estimating every missing response. When half-normal plots of absolute effects are inconclusive, another estimation methods needs to be employed.

References

[1] A. Almimi, M. Kulahci and D. C. Montgomery, Estimation of missing observations in two-level split-plot designs, Quality and Reliability in Engineering International 24 (2008), 127-152.

[2] G. Box and J. Hunter, The 2^{k-p} fractional factorial designs, Technometrics 3, 311-351 (1961), 449-458.

[3] G. Box, W. Hunter and J. Hunter, Statistics for Experiments, John Wiley, New York, 1978.

[4] G. Box and D. Meyer, An analysis for un-replicated fractional factorials, Technometrics 28(1) (1986), 11-18.

[5] C. Daniel, Use of half-normal plots in interpreting factorial two-level experiments, Technometrics 1 (1959), 311-341.

[6] C. Daniel, Applications of Statistics to Industrial Experimentation, John Wiley, New York, 1976.

[7] O. Davis, The Design and Analysis of Industrial Experiments, Oliver and Boyd, London, 1954.

[8] N. Draper and D. Stoneman, Estimating missing values in un-replicated two-level factorial and fractional factorial designs, Biometrics (1964), 443-458.

[9] P. John, Missing points in 2^k and 2^{k-p} factorial designs, Technometrics 21 (1979), 225-228.

[10] J. La Pena and D. La Pena, A simple method to identify significant effects in un-replicated two-level factorials, Comm. Stat. Theory and Methods 21 (1992), 1383-1403.

[11] R. Lenth, Quick and easy analysis of un-replicated factorials, Technometrics 31 (1989), 469-473.

[12] D. Montgomery, Design and Analysis of Experiments, 7th Edition, John Wiley and Sons, Inc., 2008.

[13] M. Qumsiyeh and G. Shaughnessy, Bootstrapping un-replicated two-level designs with missing responses, J. Stat.: Adv. Theory and Appl. 4 (2010), 91-106.

[14] P. Shearer, Missing data in quantitative designs, Applied Statistics 22 (1973), 135-140.

[15] G. Tagushi and Y. Wu, Introduction to Off-line Quality Control, Central Japan Quality Control Association, Nagoya, Japan, 1980.

[16] D. Voss, Generalized modulus-ratio test for analysis of fractional designs with zero degrees of freedom for error, Comm. Stat. Theory and Methods 17 (1988), 3345-3359.

[17] G. Wilkinson, Estimation of missing values for the analysis of incomplete data, Biometrica 14 (1958), 257-286.

[18] G. Wilkinson, A general recursive procedure for analysis of variance, Biometrica 57 (1970), 19-46.

[19] G. Z. Yin and D. W. Jullie, Orthogonal design for process optimization and its application in plasma etching, Solid State Technology 30(5) (1987), 127-132.

■