Statistica Sinica 2(1992), 381-391

ON ESTIMABILITY PROBLEMS IN INDUSTRIAL EXPERIMENTS WITH CENSORED DATA

M. Hamada and S. K. Tse

University of Waterloo and University of New Hampshire

Abstract: In industrial experiments for improving reliability, censored data are often observed because of cost and time constraints. Associated with censoring are estimability problems, however. We expose these problems in the industrial context by presenting some striking examples which show that it is difficult to tell, just by looking at the data, whether the estimates exist or not. Thus, in practice, there is a potential danger of using meaningless results when the estimates do not exist. Estimability is easily characterized for small two level factorial experiments such as 4 and 8 run designs. Because characterization becomes difficult for larger experiments, using a linear programming (LP) algorithm to check the estimability conditions is recommended. For industrial experiments whose run sizes are typically small, we propose a simple alternative LP problem that can be solved directly by a standard LP algorithm. These results apply to popular reliability models including the Weibull, lognormal and exponential regression models.

Key words and phrases: Maximum likelihood estimation, fractional factorial design, linear programming, lognormal, Weibull and exponential distributions.

1. Introduction

To improve a product's reliability, an engineer can perform an experiment to identify important factors that affect this quality characteristic. Taguchi (1987) pioneered the use of highly fractionated designs for these experiments, which was studied further by Hamada and Wu (1991). In such lifetesting experiments, censored data are often collected because of cost and time constraints. The censored data can then be analyzed by calculating maximum likelihood estimates (MLEs) for popular reliability models such as lognormal, Weibull, and exponential regression. While less costly to implement, censoring can cause estimability problems; i.e., the MLEs may be infinite and are said to not exist. The non-existence of MLEs for censored data has not been an important issue because it seldom occurs in the medical or social science context where data are abundant. It becomes an acute problem, however, for lifetesting data from highly fractionated designs, because of the small amount of data that is usually collected. Silvapulle and Burridge (1986) and Hamada and Tse (1988) showed that the question of

the MLEs' existence for these models reduces to solving a linear programming (LP) problem. For simple linear regression, Hamada and Tse (1988) described how the LP problem can be reduced to checking a few data configurations. In this paper, we consider such data configurations for more than one covariate, the situation that one faces in industrial experiments, because it is more efficient to study many factors simultaneously.

In practice, the experimenter uses a software package to calculate the MLEs. A serious problem can arise when the stopping rule of the optimization algorithm is based on the increase of the likelihood function in successive iterations. Namely, there may be no indication of anything going wrong even when some of the MLEs do not exist. While some of the estimates should diverge in theory, the stopping criteria may be met first since the likelihood becomes flat as the estimates diverge. Therefore, because of the potential danger of making decisions based on meaningless results, detecting such estimability problems is important. Recently, Clarkson and Jennrich (1991) presented algorithms for computing the MLEs when all of them do not exist.

In Section 2, we present a table that summarizes the extent of the estimability problem for the 8 run design. For a particular data set, it is generally hard to tell, just by looking at the data, whether the MLEs exist or not. Two striking examples are presented that should convince the reader. For small two level factorial designs, however, estimability problems are easily characterized. Results for the 4 and 8 run designs are given in Section 3. Characterization becomes difficult for larger designs, so that using an LP algorithm to check the estimability conditions is recommended. In Section 4 we propose a simple alternative LP problem for industrial experiments which can be solved directly by a standard LP algorithm. A check for estimability can then easily be incorporated into existing software which calculates the estimates. In Section 5, we observe that the results from the previous sections suggest an analysis strategy which has been incorporated into a method proposed by Hamada and Wu (1991). An interesting question is what additional experimentation is needed to guarantee the existence of the MLEs. Results from the previous sections also suggest a simple way to do this.

2. The Estimability Problem Exposed

We first introduce some necessary notation and then review how the question of the MLEs' existence for popular reliability regression models is answered by solving an LP problem.

By modeling the log lifetime, lognormal, Weibull, and exponential regression models fall into the following framework. Consider the model for n observations, $y_i = x_i\beta + \sigma\epsilon_i$ $(1 \leq i \leq n)$, where β, the regression parameters, and x_i, the

covariates, are p dimensional vectors. The ϵ_i are independent and identically distributed with known density. For the lognormal model, ϵ is Gaussian, whereas for the Weibull and exponential models, ϵ is the standard extreme value distribution. Note that σ equals one for the exponential model. Assume that for $0 \leq r \leq n$, the observations y_i are (i) $-\infty < a_i < y_i$ for $1 \leq i \leq r$ (right-censored) and (ii) known exactly for $r+1 \leq i \leq n$.

Provided there is at least one exactly known observation, Silvapulle and Burridge (1986) and Hamada and Tse (1988) showed that the necessary and sufficient conditions for the MLEs' existence for all these models are the same. The MLEs exist if and only if there does not exist a non-zero $e \in R^p$ for which: (i) $x_i e \geq 0$ for $1 \leq i \leq r$; and (ii) $x_i e = 0$ for $r+1 \leq i \leq n$. Thus, the question of the MLEs' existence reduces to solving an LP problem.

For designed experiments, in contrast with the general regression setup, there are a finite number of possible covariate combinations. The columns of the design matrix are orthogonal, and the entries in the design matrix for the two-level designs are either -1 or 1. This structure implies that only a finite number of data configurations need to be investigated. Although the number of data configurations increases when the experiment is replicated, we need only classify each run of the experiment. Suppose that the design has n runs. Then classify each run as R or E: R if all observations are right-censored and E (exactly known) otherwise. Then, regardless of the number of replications, the necessary and sufficient conditions above simplify to: there does not exist a non-zero $e \in R^p$ for which $xe \geq 0$ for an R run and $xe = 0$ for an E run, where x is the appropriate row from the regression design matrix for the model being fitted. Thus, we need only solve an LP problem with n constraints in p variables.

Table 1 displays the potential estimability problems of fitting eight different models for the 8 run two-level designs studied later in Section 3. Let 2_c^{a-b} denote the model for a two level factors and c two factor interaction (f.i.) based on a 2^{-b} fraction of a full factorial design. The first and second columns (#R and Total) give the number of R runs (out of 8) and the total number of such data configurations (8 choose #R), respectively. The body of the table gives the number of configurations for which the MLEs do not exist. The table shows that the estimability problem increases as the number of parameters in the model increases and as the number of R runs increases.

While Table 1 displays the extent of the estimability problem for the 8 run design, it does not answer the question of whether the MLEs exist for a particular data set. For the main effects model, the conditions of Silvapulle and Burridge (1986) and Hamada and Tse (1988) suggest a geometric approach for verifying them: if there exists a p-dimensional hyperplane such that all the R runs fall on one side, then the MLEs do not exist; otherwise, the MLEs do exist. Two data

Table 1. MLEs' non-existence for the 8 run design

		Design							
#R	Total	2^3	2^{4-1}	2^{5-2}	2^{6-3}	2^{7-4}	2_1^3	2_1^{4-1}	2_2^{4-1}
0	1	0	0	0	0	0	0	0	0
1	8	0	0	0	0	8^8	0	0	0
2	28	0	0	8	16	28	4	8	16
3	56	0	0	40	48	56	24	40	48
4	70	6	16	66	68	70	56	66	68
5	56	24	32	56	56	56	56	56	56
6	28	24	24	28	28	28	28	28	28
7	8	8	8	8	8	8	8	8	8
8	1	1	1	1	1	1	1	1	1

configurations are easily checked using this approach. First, if two E runs have design matrix rows with opposite signs, then the MLEs exist; it is impossible to have all the R runs on the same side of the hyperplane which passes through the pair of runs. We refer to this configuration as the *opposite sign pair.* This result has implications for subsequent experimentation which is discussed in Section 5. Second, if all the runs in the design with the same level of a factor are R runs, then the MLEs do not exist. Here, a hyperplane can be fit through the E runs with all the R runs on the same side of the hyperplane. We refer to this configuration as *complete separation.*

Generally, it is hard to tell just by inspecting the data whether the MLEs exist or not as the two examples that we present next suggest. Although there is much structure in designed experiments, the structure is sufficiently complex to prevent a simple method to determine estimability. As seen by the next example as well as Table 1, existence is not necessarily guaranteed if the number of E runs exceeds the number of parameters to be estimated.

Example 1. Consider the data from a 16 run design in 9 factors as shown in Table 2. Assume an exponential regression model ($f(y) = \theta \exp\{-\theta y\}$, where $\theta = \exp\{x\beta\}$) with an intercept and nine main-effects (10 parameters). Note that 12 out of 16 run are E runs! Using ISMOD (Lawless and Singhal (1987a,b)) to fit the model, the optimizer went through 7 iterations yielding the estimates and standard errors given in Table 3. Although the ISMOD output did not indicate a problem, the MLEs do not exist for this data configuration.

This example shows that although the MLEs do not exist, the optimization

Table 2. Design and data for Example 1

			Design Matrix								
Run	Data	Type	A	B	C	D	E	F	G	H	I
1	2.0	R	1	1	1	1	1	1	1	1	1
2	0.5	E	1	1	1	−1	1	−1	−1	−1	−1
3	0.6	E	1	1	−1	1	−1	−1	−1	1	−1
4	2.0	R	1	1	−1	−1	−1	1	1	−1	1
5	0.7	E	1	−1	1	1	−1	−1	1	−1	−1
6	2.0	R	1	−1	1	−1	−1	1	−1	1	1
7	2.0	R	1	−1	−1	1	1	1	−1	−1	1
8	0.8	E	1	−1	−1	−1	1	−1	1	1	−1
9	0.9	E	−1	1	1	1	−1	1	−1	−1	−1
10	1.0	E	−1	1	1	−1	−1	−1	1	1	1
11	1.2	E	−1	1	−1	1	1	−1	1	−1	1
12	1.3	E	−1	1	−1	−1	1	1	−1	1	−1
13	1.4	E	−1	−1	1	1	1	−1	−1	1	1
14	1.5	E	−1	−1	1	−1	1	1	1	−1	−1
15	1.6	E	−1	−1	−1	1	−1	1	1	1	−1
16	1.7	E	−1	−1	−1	−1	−1	−1	−1	−1	1

program can terminate since the likelihood becomes flat as the estimates diverge. While many iterations of the optimization routine can signal problems, the defaults in a software package may preclude this possibility. Unless the stopping criteria are suitably chosen, our concern is that practitioners will not be aware of an estimability problem and consequently make decisions based on meaningless results.

Example 2. Consider the data configuration for a different 16 run design in 8 factors as displayed in Table 4. Here only 2 out of 16 runs are E runs. Surprisingly, the MLEs exist for an exponential regression model with 8 main effects (9 parameters including the intercept). The opposite sign pair configuration explains why the MLEs exist.

3. Results for 4 and 8 Run Designs

In this section, we use the two easily checked configurations, complete separation and opposite sign pair, as well as some other rules to characterize the estimability problems for the 4 and 8 run designs. The results for the 4 run design are based on the two easily checked configurations and are presented first.

Table 3. Estimates and standard errors for Example 1

Parameter	Estimate	Standard error
INT	−1.95 + 00	3.92 + 00
A	−1.66 + 00	3.92 + 00
B	1.67 + 00	3.06 − 01
C	7.43 − 02	3.06 − 01
D	2.17 − 02	3.06 − 01
E	5.06 − 03	3.06 − 01
F	−1.99 + 00	3.92 + 00
G	−1.94 − 02	3.06 − 01
H	−4.04 − 03	3.06 − 01
I	−1.99 + 00	3.92 + 00

Table 4. Data configuration for Example 2

		Design Matrix							
Run	Type	A	B	C	D	E	F	G	H
1	E	1	1	1	1	1	1	1	1
2	R	1	1	1	−1	1	−1	−1	−1
3	R	1	1	−1	1	−1	−1	−1	1
4	R	1	1	−1	−1	−1	1	1	−1
5	R	1	−1	1	1	−1	−1	1	−1
6	R	1	−1	1	−1	−1	1	−1	1
7	R	1	−1	−1	1	1	1	−1	−1
8	R	1	−1	−1	−1	1	−1	1	1
9	R	−1	1	1	1	−1	1	−1	−1
10	R	−1	1	1	−1	−1	−1	1	1
11	R	−1	1	−1	1	1	−1	1	−1
12	R	−1	1	−1	−1	1	1	−1	1
13	R	−1	−1	1	1	1	−1	−1	1
14	R	−1	−1	1	−1	1	1	1	−1
15	R	−1	−1	−1	1	−1	1	1	1
16	E	−1	−1	−1	−1	−1	−1	−1	−1

Define the E (R) set for a particular data configuration to be the run numbers of the E (R) runs where #E (#R) denote the number of runs in the set. The E (R) lists are simply lists of E (R) sets with mE (mR) denoting lists of m size sets. In the following, we represent a set of runs by combining all the run numbers into a single number; e.g., 12 represents runs 1 and 2 from the design matrix.

3.1. 4 run design results

The 4 run design matrix is given in Table 5. The 2^2 and $2^{3-1}(=2_1^2)$ designs are obtained by using the first two and three columns, respectively.

Table 5. 4 run design matrix

	Design Matrix		
Run	1	2	12
1	1	1	1
2	1	−1	−1
3	−1	1	−1
4	−1	−1	1

The results for 2^2 are: (1) The MLEs exist for all 1R cases. (2) Of the 6 2R cases, 4 have complete separation (MLEs do not exist) and 2 are opposite sign pairs (MLEs exist). (3) All 3R cases have complete separation. For $2^{3-1}(=2_1^2)$, there must be all E runs for the MLEs to exist. These results can be summarized as follows: for 2^2, if 12, 13, 24, or 34 are contained in the R set, then the MLEs do not exist; for $2^{3-1}(=2_1^2)$, if the R set is contained in 1234, then the MLEs do not exist.

3.2. 8 run design results

Table 6. 8 run design matrix

	Design Matrix						
Run	1	2	3	123	12	13	23
1	1	1	1	1	1	1	1
2	1	1	−1	−1	1	−1	−1
3	1	−1	1	−1	−1	−1	1
4	1	−1	−1	1	−1	1	−1
5	−1	1	1	−1	−1	1	−1
6	−1	1	−1	1	−1	−1	1
7	−1	−1	1	1	1	−1	−1
8	−1	−1	−1	−1	1	1	1

The 8 run design matrix is given in Table 6 with columns used for different models displayed in Table 7. Note that the models not listed with fewer

Table 7. Columns used for 8 run design models

	Columns used						
Model	1	2	3	12	13	23	123
2^3	x	x	x				
2^{4-1}	x	x	x				x
2^{5-2}	x	x	x	x	x		
2^{6-3}	x	x	x	x	x	x	
2^{7-4}	x	x	x	x	x	x	x
2_1^3	x	x	x	x			
2_1^{4-1}	x	x	x	x			x
2_2^{4-1}	x	x	x	x	x		x

main effects and more two factor interactions are identical to the main effects models listed; namely, $2_2^3 = 2^{5-2}$, $2_3^3 = 2^{6-3}$, $2_3^{4-1} = 2^{7-4}$, $2_1^{5-2} = 2^{6-2}$, $2_2^{5-2} = 2_1^{6-2} = 2^{7-4}$. For each given model, there are 256 data configurations to consider. We developed some rules to eliminate the need to check every data configuration (see Hamada and Tse (1989a)) and used then together with the two easily checked configurations to obtain the results for the eight different models displayed previously in Table 1.

A simple summary for all 8 run design models is given below.

- For 2^3, if the R set contains 1234, 5678, 1256, 3478, 1357, or 2468, then the MLEs do not exist.
- For 2^{4-1}, if the R set is contained in 123678, 234567, 134568, or 124578, then the MLEs exist.
- For 2^{5-2}, if the R set is contained in 1467, 2358, 1368, or 2457, then the MLEs exist.
- For 2^{6-3}, if the R set is contained in 1467 or 2358, then the MLEs exist.
- For 2^{7-4}, if the R set contained in 12345678, then the MLEs do not exist.
- For 2_1^3, if the R set contains 12, 34, 56, or 78 or complete separation (1234, 5678, 1256, 3478, 1357, 2468), then the MLEs do not exist.
- For 2_1^{4-1}, if the R set is contained in 1458, 1368, 2457, or 2367, then the MLEs exist.
- For 2_2^{4-1}, if the R set is contained in 1368 or 2457, then the MLEs exist.

Next, we make some comments about Table 1. These results demonstrate that for a given design, estimability problems increase as the censoring becomes heavier (larger #R). For a given amount of censoring, estimability problems

increase as the number of parameters increase; in other words, larger designs or designs with fewer factors provide more protection. Also, note the dramatic increase in the MLEs' nonexistence in moving from 2^{4-1} to 2^{5-2} and the following surprising cases where the MLEs exist for large size R sets and do not exist for small size R sets: for the former where $p \geq \#\text{E}$, see (4R, $L_8(2^{5-2})$), (6R, $L_8(2^3)$, $L_8(2^{4-1})$); for the latter where $p \leq \#\text{E}$, see (2R, $L_8(2^{5-2})$).

The complexity of larger designs quickly increases so that an exhaustive study of all the data configurations becomes prohibitive. Nevertheless, the MLEs' existence for a particular data set can still be checked for any design no matter what its size by solving the LP problem given in Section 2. In the next section, we propose a simple alternative LP problem for industrial experiments.

4. A Simple Alternative LP Problem for Industrial Experiments

The LP problem to check estimability given in Section 2 cannot be solved directly by standard LP algorithms since they find the zero vector which is always a solution; recall that a non-zero solution is needed for non-existence. Instead, we propose a simple alternative LP problem particularly suited for industrial experiments. Silvapulle and Burridge (1986) also proposed an alternative problem based on reducing the size of the problem, which included several intermediate steps to obtain an alternative LP problem. Their approach reduces the problem size at the expense of simplicity. It is especially appealing for biomedical applications where the data sets can be large. In the industrial context, however, reducing the original problem size is unnecessary since the run size of designed experiments is typically small. Recall that even if several replicates are taken, the problem can still be solved in terms of the runs and not the individual observations.

The following simple restatement of the original LP problem given in Section 2 is as follows. Suppose that n is the design run size and #R is the number of R runs. We change each inequality associated with an R run into an equality by adding a new slack variable s. Since this new variable must be negative, we introduce the corresponding constraint, $s \leq 0$. Thus, the alternative LP problem is to minimize the sum of all the new slack variables given all the constraints consisting of n equalities (one for each run) plus #R new inequalities (one for each new negative slack variable). There are two solutions to this alternative problem: either the zero vector is the only solution or the problem is unbounded from below. The MLEs exist for the former case, but not for the latter. This alternative problem can be handled directly by a Phase 1–Phase 2 algorithm (Best and Ritter (1985)). See Hamada and Tse (1989b) for an implementation which can easily be added as a front end to the optimization program which calculates the estimates. Note that our approach, in fact, increases the size of the original

LP problem. Consider, however, that the worst case for a 32 run design would be 63 constraints in 63 variables which is still a small problem by LP standards.

5. Discussion

The results for the 8 run design in Table 1 contain some important information that suggest a general analysis strategy: estimability problems tend to occur with saturated or nearly saturated models, i.e., models whose number of parameters are equal or nearly equal to the number of experimental runs. Consequently, the MLEs for a comprehensive model will usually not exist. This observation suggests a strategy of building up a model rather than starting with a comprehensive model and then looking for good submodels. Hamada and Wu (1991) proposed an iterative scheme of model fitting, imputation of censored observations, and model selection which incorporates this strategy. Although their procedure uses MLEs, most estimability problems are avoided by building up the model. See their paper for details and applications of their method to two actual experiments.

While these results suggest that larger designs provide better protection against estimability problems, what can be done when the MLEs do not even exist for the main effects model? Two interesting questions are how many additional runs are needed to guarantee the MLEs' existence and what are they. The opposite sign pair configuration suggests a simple solution. If we perform one additional run so that there are a pair of E runs with opposite signs, then the MLEs exist. Thus, for one E run, perform an experiment at the opposite combination until exact data are observed. A referee asked whether a combination could be run that guarantees existence whatever its outcome. The following example suggests this possibility. Consider the 4 run design for two factors with two R runs at the same level of one of the factors, say runs 1 and 2; the MLEs do not exist for the main effects model. If an additional run $(-1.5, 0)$ is performed (assuming that both factors are quantitative), then the MLEs exist for the main effects model as well as the full model (including the 2 f.i.) no matter whether the run is exact or censored. More work along these lines for larger designs is needed.

Acknowledgements

We thank M. Best for helpful discussions and R. J. MacKay and C. F. J. Wu for comments on an earlier version. We also thank the referees for their valuable comments. M. Hamada was supported by research grants from General Motors of Canada Limited, the Manufacturing Research Corporation of Ontario, and the Natural Sciences and Engineering Research Council of Canada.

References

Best, M. J. and Ritter, K. (1985). *Linear Programming Active Set Analysis and Computer Programs*. Prentice-Hall Inc., Englewood Cliffs.

Clarkson, D. B. and Jennrich, R. I. (1991). Computing extended maximum likelihood estimates for linear parameter models. *J. Roy. Statist. Soc. Ser.B* **53**, 417-426.

Hamada, M. and Tse, S. K. (1988). A note on the existence of maximum likelihood estimates in linear regression models using interval-censored data. *J. Roy. Statist. Soc. Ser.B* **50**, 293-296.

Hamada, M. and Tse, S. K. (1989a). When do estimates exist from censored data in industrial experiments. Research Report 89-03, University of Waterloo, Institute for Improvement in Quality and Productivity.

Hamada, M. and Tse, S. K. (1989b). MLECHK: a FORTRAN program for checking the existence of maximum likelihood estimates from censored, grouped, ordinal and binary data in designed experiments. Research Report 89-09, University of Waterloo, Institute for Improvement in Quality and Productivity.

Hamada, M. and Wu, C. F. J. (1991). Analysis of censored data from highly fractionated experiments. *Technometrics* **33**, 25-38.

Lawless, J. F. and Singhal, K. (1987a). ISMOD: an all-subsets regression program for generalized linear model. I. Statistical and computation background. *Comput. Methods and Programs in Biomedicine* **24**, 117-124.

Lawless, J. F. and Singhal, K. (1987b). ISMOD: an all-subsets regression program for generalized linear model. II. Program guide and examples. *Comput. Methods and Programs in Biomedicine* **24**, 125-134.

Silvapulle, M. J. and Burridge, J. (1986). Existence of maximum likelihood estimates in regression models for grouped and ungrouped data. *J. Roy. Statist. Soc. Ser.B* **48**, 100-106.

Taguchi, G. (1987). *System of Experimental Design*. Unipub/Kraus International Publications, White Plains.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

Department of Mathematics, University of New Hampshire, Kingsbury Hall, Durham, NH 03824, U.S.A.

(Received June 1990; accepted October 1991)