

Bootstrap analysis of designed experiments

Ron S. Kenett
KPA Ltd. and University of Torino
ron@kpa.co.il

Effi Rahav
Tel Aviv University
ne-rahav@bezeqint.net

David M. Steinberg
Tel Aviv University and KPA Ltd.
dms@post.tau.ac.il

Summary

In the first ENBIS conference in Oslo, Kenett and Steinberg proposed to apply Bootstrapping to the analysis of data derived from moderate size designed experiments [1].

In this paper we present follow up work with emphasis on designed experiments with censored data, heteroscedasticity and constraints related to the structure of the experiment. Our results show clear advantages to bootstrap based data analysis, which is both robust and relatively easy to apply. The bootstrap analysis often points to problems in linear model analyses that might easily be overlooked. These findings suggest that bootstrapping can contribute significantly to the design of experiments methodology.

Key words: Design of experiments, bootstrapping, mathematical models, errors of the third kind

Introduction to Bootstrapping

Bootstrapping was first introduced by B. Efron in 1979 as a computer intensive, yet conceptually simple, technique that leverages the collected data to draw statistical inferences [2]. An important advantage of the bootstrap is that it provides a common paradigm for inference that can be applied both to standard and non-standard statistical analyses. Bootstrapping is an extremely useful methodology for analyzing data sets of intermediate size, a typical situation in analyzing data from designed experiments.

An excellent in depth presentation of bootstrapping is available in [3], a practical introduction to bootstrapping is presented in [4].

How does Bootstrapping work? Suppose we have a sample of data and use it to compute a statistic, say T , that estimates a population parameter of interest. The following steps provide a Bootstrap Confidence Interval for the parameter.

- 1) Take a Random Sample With Replacement (RSWR) from the data and compute the statistic T
- 2) Resample M times and compute the statistic T for each new sample
- 3) Derive the Empirical Bootstrap Distribution (EBD) of T from the M samples
- 4) Compute the Empirical Bootstrap Confidence Interval for the population parameter by identifying appropriate quantiles of the EBD

For example, a sample of 32 hybrid microcircuits was collected to estimate the inter-layer resistance in ohms. The mean resistance in the sample was computed as 2143.4 ohms.

Does the sample indicate a significant departure from the design target of 2100 ohms?

Resampling one thousand times with replacement yields the Empirical Bootstrap Distribution of the sample mean. The 95% Empirical Bootstrap Confidence Interval of the inner-layer resistance using the 2.5th and 97.5th percentiles of the 1000 bootstrap values, yields an interval between 2109.5 ohms and 2179.9 ohms (see Figure 1).

Since 2100 ohms is outside the 95% bootstrap confidence interval we can state, with 95% confidence, that the hybrids are manufactured at a resistance level higher than the design target.

The approach outline above is derived from the original suggestion of Efron and has been labeled the percentile method. It has been proved to work asymptotically and is approximately correct for small samples. In order to improve coverage probabilities some later suggestions have been proposed. These include the Bootstrap-t, in which standardized bootstrapped values are computed, and bias corrected and accelerated (BCA) bootstrap estimates (see [4] and [5]).

Analyzing Data from Designed Experiments with Bootstrapping

Designed experiments typically consist of balanced arrays of experimental runs that allow for efficient estimation of factor effects and their interactions. However, in running designed experiments one often meets anticipated and unanticipated problems.

Some anticipated issues can be dealt with in the design phase. For example, the potential impact of raw materials or operating conditions can be accounted for by running the experiment in separate blocks. Practical constraints may dictate that some factors will be "nested" within others or that there will be limitations on the run order. In other examples, there may be some experimental points that turn out to be impossible to execute because of logistics or technological requirements.

Unexpected problems may arise when the experiment is carried out. For example, experiments can produce non-quantifiable results or experimental points may generate "outliers," observations whose value appears quite incompatible with the overall pattern in the data.

Bootstrapping provides a working approach to statistical inference for factor effects when faced with such practical considerations. The minimal requirement for applying the bootstrap is the presence of replicate observations (or a suitable model for generating replicates) at all levels of the experiment.

To demonstrate how bootstrapping can be applied to analyze an experiment we use data from an experiment that examined conditions for optimal growth of bean sprouts. The experiment was a full 2³ factorial with 3 replications. Three factors were studied: 1) Type of Water (tap vs. bottled), 2) Growth Medium (1 vs. 2 layers of cotton) and 3) Location (inside vs. outside). The response variable is height (in cm) of the bean sprout.

=====
 Table 1: The bean sprout experiment
 =====

The main goal of the experiment is to assess the main effects of the three factors and their interactions. Regression analysis of the results shows a large main effect for the growth medium (p-value < 0.0001), a smaller, but clearly significant main effect for the location (p-value < 0.0001) and a possible interaction of these two factors (p-value = 0.042). However, there is also a moderate outlier (the observation with result of 24.8 cm), casting some doubt on the validity of the p-values from the regression.

The bootstrap provides a simple and direct way to carry out statistical inference despite the presence of the outlier. The bootstrapping implementation must reflect the structure of the experiment, with 3 replicate outcomes at each of the 8 experimental conditions. Separate RSWR's are taken at each of the conditions, and for each repeated sample estimates of main effects and interactions are computed. This produces Empirical Bootstrap Distributions for the main effects and interactions. It is worth emphasizing that even though there are only 10 possible distinct samples for each experimental run the overall number of combinations across all experimental points is very large.

=====
 Table 2: EBD mean and 95% Bootstrap Confidence Intervals of main effects and interactions
 =====

The bootstrap analysis shows that the medium by location interaction is clearly significantly different from zero at the 5% level. Each of the regression coefficients is determined up to an interval of about 1.5 units.

The power of bootstrapping is in its simplicity and its ability to provide answers with minimal assumptions. Theoretical studies of the bootstrap have shown both the accuracy and the robustness of bootstrap inferences. For more references to bootstrapping see ([6]-[10]). Surprisingly, the bootstrap has been applied rarely to the analysis of designed experiments. In a sense, randomization tests for analysis of data gathered in designed experiments can be seen as an application of bootstrapping with all permutations being considered. However they are limited to small experiments. Recently, an application of the Jackknife to designed experiments was proposed ([11]) in a way also applicable mostly to small size experiments. The application of bootstrapping to designed experiments is not affected by such limitations and can be considered an expansion of randomization tests and application of the Jackknife. For a general discussion of computer intensive methods in the context of designed experiments see [12]. The next section will expand on the application of bootstrapping in the context of designed experiments.

Model Diagnostics with Bootstrapping Data from Designed Experiments

In analyzing data an underlying model is fitted to the data. Two-level factorial experiments are used to estimate parameters of a linear model which, in turn, depends on the estimability properties of the experimental design. When a model is misspecified an error of the third kind is said to have occurred. Bootstrapping can be used to flag errors of the third kind or, alternatively, validate a specific model. Use of an inadequate model will often lead to an overestimate of the residual variance and to inflated standard errors for the model parameters. Comparison of bootstrap SE's to those from a regression analysis is thus a valuable diagnostic. If the bootstrap standard errors are clearly smaller than those from fitting a regression model to the experimental data, this is a likely sign that the model is inadequate. We demonstrate this property with an example from [13, Section 13.1, p. 563].

A 2⁷⁻³ fractional factorial experiment was conducted to investigate the impact of 7 factors on wave soldering defect levels. The factors were:

- A - prebake temperature
- B - flux density
- C - conveyor speed
- D - preheat condition
- E - cooling time
- F - solder agitator
- G - solder temperature

The experiment was conducted with 3 replicates and is therefore amenable to bootstrapping as described in the preceding section.

For each bootstrapped sample, estimates of the linear model coefficients are computed and respective EBD estimates and confidence intervals are computed (see Table 3)

=====

Table 3: EBD mean and 95% Bootstrap Confidence Intervals of linear model regression coefficients

=====

As evident from Table 3 factors D, E and F have non-significant main effects, suggesting that they can be ignored and providing an opportunity to fit a model with interactions.

In order to properly compare regression based estimates and bootstrap estimates normality was induced by applying the square root transformation to the solder defects data.

In Table 4 we present a comparison of regression estimates computed on the transformed data and the corresponding bootstrap estimates.

=====
Table 4: EBD means and 95% Bootstrap Confidence Intervals and regression analysis of linear model
=====

From Table 4, the regression analysis implies that only factors C and G are significant, whereas bootstrapping points to A, B, C and G.

By ignoring factors D, E and F a model with interaction can also be fitted. Table 5 compares the standard errors of the effects derived from the first-order regression model with all 7 factors and the collapsed model with interactions, with regression and bootstrap analyses. The relative difference in variability, Delta, is computed and expressed as a percentage.

=====
Table 5: Standard errors of bootstrapping and regression estimates of first-order regression model with all 7 factors and a model with interactions.
=====

The gap between the regression and bootstrap standard errors in the first-order model is about 18%. When using the collapsed model with interactions, this gap drops to about 2%.

The sizable difference between the bootstrap and regression estimates of variability in the first-order model is an indicator of lack of fit or error of the third kind. A rule of thumb could indicate such a gap when Delta is above 5%.

When comparing models, the standard error in regression estimates drops by 20% when using the linear model versus the model with interactions, but the bootstrap standard error drops only by about 1%.

The practical implication of this phenomenon is that bootstrap estimates are more robust than standard regression estimates and that comparing variability of regression estimates to bootstrap estimates is a good test for errors of the third kind.

Reanalyzing the experiment on wave soldering from [13] suggests that the appropriate model to use for this data involves factors A, B, C and G and their interactions. The improvement from using this model rather than a main effects model is also supported by a cross-validation analysis. The cross-validation mean squared error for the main effects model is 5.1 as compared to 4.1 for the model using factors A, B, C and G and their interactions, despite the fact that the cross-validation “penalizes” the latter model for adding parameters.

Handling Heteroscedasticity in Data from Designed Experiments

[13, Section 12.1.2, p. 531] provides yet another example that can be used to demonstrate the advantage of bootstrapping. The experiment involves a saturated 12 run Plackett-Burman design for analyzing the effect of 11 factors on the number of cycles until first failure of a thermostat in a heater. In each experimental run 10 heaters were tested with right censoring at 7342 cycles. Heaters that reached such a number of cycles were dropped from the experiment and the figure 7342 was recorded. The data and variances of the 12 experimental runs are displayed in Table 6.

 Table 6: Plackett-Burman experiment with censoring at 7342 cycles

The data show significant heteroscedasticity. In such a case, the correct standard errors of the coefficients need to reflect the unequal variances. Regression analysis of the original model (all main effects, no interactions) gives all standard errors as 127.5; bootstrap analysis of the same model gives all standard errors as close to 125. Wu and Hamada suggest that the model with only f5, f8 and their interaction gives a very good fit to the data. We re-analyzed the experiment, including the f5*f8 interaction instead of f2. Regression and bootstrap analysis for this model give the SE's in Table 7.

 Table 7: Standard errors of the model coefficients from a standard regression, from the bootstrap and from application of the unequal error variances to the regression.

The large difference in standard errors between the regression and the bootstrap serves as a diagnostic that, most likely, something is wrong in the assumptions for the regression model. In fact, as already noted, there is a severe problem here of heteroscedasticity. As a result, the standard errors from the regression model, which assume constant error variance, are not correct. The standard errors of the coefficients need to reflect the unequal variances. Corrected standard errors, computed from the sample variance at each design point, are presented in the final column of Table 7.

The bootstrap standard errors are perfectly compatible with the corrected regression standard errors. A data analysis that fails to take account of the heteroscedasticity would lead to greatly biased conclusions. The bootstrap, on the other hand, gives accurate standard errors automatically. We think that this robustness of the bootstrap, as opposed to regression based analysis, should make bootstrapping a method of choice for practitioners.

Discussion

The following observations are typically experienced by industrial statisticians involved in process and product improvement, in a variety of industries:

- 1) Successful designed experiments accelerate learning – a critical ability of competitive organizations, however designed experiments require planning and discipline, a difficult requirement in most work environments.
- 2) Designed experiments require close collaboration between statisticians and content experts. Good communication channels are a necessary condition for success.
- 3) Prospective and retrospective constraints impact the design and analysis of statistical experiments. Missing data and other constraints are common in DOE implementations.

These three observations create the need for robust and intuitive statistical methods and their implementation in a broad management context. Bootstrapping provides an efficient, robust and intuitive method supporting the implementation of designed experiments.

From a larger view experimentation is an iterative process between:

- Questions for investigation
- Experimental design and data acquisition
- Data analysis
- Answers and interpretations

An experimentation data analysis strategy with bootstrapping involves six steps:

1. Evaluation of experimental conditions including the identification of experimental constraints and de fact constraints, not originally planned in the experimental design. These constraints reflect missing or extra experimental runs, constraints on the setting of factor levels or randomization and run order issues.
2. Design of bootstrap strategy. This involves specifying the underlying mathematical model used in the data analysis and the bootstrapping algorithm.
3. Bootstrap analysis. This is an iterative step where an initial pilot run of resampled data is evaluated using mostly graphical displays to validate the bootstrapping algorithm accuracy.
4. In order to generate in depth diagnostics, a fit of the data using regression is followed by computation of standard errors from the regression model and the empirically bootstrapped distribution.
5. A diagnostic check is performed by comparing standard errors of regression coefficients and bootstrapped standard errors.
6. Gaps are interpreted through a second iterative cycle until the analysis is completed. The iteration involves sequential adaptation of regression models until a match is achieved with bootstrapping results.

Figure 2 presents a sketch of the above strategy.

=====

Figure 2: The six steps experimentation data analysis strategy with bootstrapping analysis

=====

Bootstrapping reinforces the iterative loop, within the analysis step, between the statistician and the content expert who can provide critical inputs to the data analysis both in terms of the statistics of interest, the evaluation of the experimental conditions and the bootstrapping strategy.

The bootstrap resampling should reflect all design constraints including unplanned constraints that entered during the course of the experiment. When possible, direct resampling of the data should be used. When there are too few replicates to facilitate resampling and reasonable knowledge of the data distributions, the parametric bootstrap is a good alternative.

The application of bootstrapping to analyzing data from designed experiments is a relatively new discipline. Many questions need to be answered in the context of non-standard situations where dynamic effects, structural constraints, retrospective insight and other effects need to be addressed explicitly in the data analysis. It is expected that further research will provide insights into these issues from both a theoretical and practical perspective.

References

1. Kenett, R. S., Steinberg, D. M. On the Application of Bootstrapping in Analyzing Designed Experiments. *Proceedings of the First Annual Conference on Business and Industrial Statistics*, Oslo, Norway, 2001: September 17-18.
2. Efron, B. Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics* 1979: 7, pp. 1-26.
3. Davison, A. C. and Hinkley, D. V. *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge, 1997.
4. Kenett R. S., Zacks, S. *Modern Industrial Statistics: Design and Control of Quality and Reliability*, Duxbury Press, San Francisco, 1998.
5. Efron, B. and Tibshiriani, R. J. *An Introduction to the Bootstrap*, Chapman and Hall, New York: 1993.
6. Fisher, N. I. and Hall, P. Bootstrap algorithms for small samples. *Journal of Statistical Planning and Inference* 1991: 27, 157–169.
7. Liu, R. Y. Bootstrap procedures under some non-i.i.d. models. *The Annals of Statistics* 1988: 16, 1697–1708.
8. Polansky, A. M. Selecting the best treatment in designed experiments. *Statistics in Medicine* 2003: 22, 3461–3471.
9. Shao, J. and Tu, D. *The Jackknife and Bootstrap*. Springer: New York, 1995.
10. Wu, C. F. J. Jackknife, bootstrap and other resampling methods in regression analysis (with discussions). *The Annals of Statistics* 1986: 14, 1261–1350.
11. Variyath, A., Abraham, B, Chen, J. Analysis of Performance Measures in Experimental Design Using Jackknife, *Journal of Quality Technology* 2005: 37, 2, pp. 91-100.
12. Kenett, R. S., Steinberg, D. M. New Frontiers in Design of Experiments: From Fisher to DACE. *Quality Progress*, to appear, 2006.
13. Wu, C. F. J. and Hamada, M. *Experiments: Planning, Analysis and Parameter Design Optimization*, J. Wiley & Sons, New York, 2000.

Factors			Responses		
Water	Medium	Location	Y1	Y2	Y3
Tap	1 Layer	Inside	10.7	7.9	7.8
Bottled	1 Layer	Inside	7.8	6.6	6.1
Tap	2 Layers	Inside	21.2	22.6	21.2
Bottled	2 Layers	Inside	23.7	24.2	24.7
Tap	1 Layer	Outside	12.3	13.2	9.8
Bottled	1 Layer	Outside	10.9	13.7	10.1
Tap	2 Layers	Outside	33.5	29.4	30.7
Bottled	2 Layers	Outside	34.7	24.8	32.6

Table 1: The bean sprout full factorial experiment

Effect	EBD mean	EBD LCI	EBD UCI
Water (W)	-0.03	-0.78	0.66
Medium (M)	8.59	7.84	9.31
Location (L)	2.96	2.21	3.68
WxM	0.52	-0.32	1.15
WxL	-0.18	-0.92	0.55
MxL	1.03	0.26	1.74
WxMxL	-0.61	-1.4	0.11

Table 2: Empirical Bootstrap Distribution (EBD) mean and 95% Bootstrap Confidence Intervals of main effects and interactions

Factor	EBD Mean	EBD Variance	EBD Std	95% LCI	95% UCI
Average	32.73	11.59	3.4	27.9	38.33
A: Prebrake temperature	10.2	11.66	3.41	5.29	16.04
B: Flux density	-8.79	12.1	3.48	-14.42	-4.0
C: Conveyor speed	18.13	11.43	3.38	13.1	23.5
D: Preheating condition	-0.78	11.82	3.44	-6.46	3.92
E: Cooling time	-3.6	12.04	3.47	-8.52	2.4
F: Solder agitator	-1.03	11.2	3.35	-6.65	3.67
G: Solder temperature	-11.04	11.95	3.46	-15.88	-5.35

Table 3: Empirical Bootstrap Distribution (EBD) mean and 95% Bootstrap Confidence Intervals of linear model regression coefficients

Regression analysis

	Value	Std.	t value	Pr(> t)
A	0.5404	0.2986	1.8095	0.0779
B	-0.5393	0.2986	-1.806	0.0784
C	1.4238	0.2986	4.7678	0
D	-0.153	0.2986	-0.5123	0.6113
E	-0.282	0.2986	-0.9442	0.3507
F	-0.1563	0.2986	-0.5235	0.6035
G	-0.8932	0.2986	-2.991	0.0047

Bootstrap estimates

	Mean	Variance	Std	95% BCI	
				LCI	UCI
A	0.5476	0.0596	0.2441	0.1929	0.9477
B	-0.5462	0.0597	0.2443	-0.935	-0.186
C	1.4333	0.0604	0.2458	1.0754	1.8512
D	-0.1468	0.0577	0.2402	-0.5253	0.2114
E	-0.2822	0.0577	0.2402	-0.6391	0.1014
F	-0.166	0.0611	0.2472	-0.5817	0.1939
G	-0.8878	0.062	0.249	-1.2623	-0.4692

Table 4: Regression analysis of linear model, Bootstrap estimates and 95% Bootstrap Confidence Intervals (BCI).

	No Interactions			With Interactions		
	Regr.	Bootstrap	Delta	Regr.	Bootstrap	Delta
A	0.299	0.244	18%	0.239	0.249	-4%
B	0.299	0.244	18%	0.239	0.237	1%
C	0.299	0.246	18%	0.239	0.243	-2%
D	0.299	0.240	20%			
E	0.299	0.240	20%			
F	0.299	0.247	17%			
H	0.299	0.249	17%	0.239	0.246	-3%
A*B				0.239	0.240	-1%
A*C				0.239	0.244	-2%
B*C				0.239	0.235	2%
A*G				0.239	0.240	-1%
B*G				0.239	0.239	0%
C*G				0.239	0.245	-3%
A*B*C				0.239	0.245	-3%
A*B*G				0.239	0.239	0%
A*C*G				0.239	0.245	-3%
B*C*G				0.239	0.239	-0.002
A*B*C*G				0.239	0.235	0.015

Table 5: Standard errors of regression (Regr.) and bootstrapping (Bootstrap) estimates of first-order regression model with all 7 factors and a model with interactions, and relative difference (Delta).

ID#	<u>1</u>	<u>2</u>	<u>3</u>	<u>4</u>	<u>5</u>	<u>6</u>	<u>7</u>	<u>8</u>	<u>9</u>	<u>10</u>	<u>Variance</u>
1	957	2846	7342	7342	7342	7342	7342	7342	7342	7342	5460292
2	206	284	296	305	313	343	364	420	422	543	8775
3	63	113	129	138	149	153	217	272	311	402	10928
4	75	104	113	234	270	364	398	481	517	611	35194
5	97	126	245	250	390	390	479	487	533	573	28505
6	490	971	1615	6768	7342	7342	7342	7342	7342	7342	9172099
7	232	326	326	351	372	446	459	590	597	732	23835
8	56	71	92	104	126	156	161	167	216	263	4214
9	142	142	238	247	310	318	420	482	663	672	37165
10	259	266	306	337	347	368	372	426	451	510	6424
11	381	420	7342	7342	7342	7342	7342	7342	7342	7342	8566204
12	56	62	92	104	113	121	164	232	258	731	40185

Table 6: Plackett-Burman experiment with censoring at 7342 cycles

<u>Factor</u>	<u>Regr.</u>	<u>Bootstrap</u>	<u>WRegr</u>
f1	180.3	195.0	197.9
f5*f8	382.4	385.1	382.4
f3	180.3	200.0	202.2
f4	180.3	20.4	20.2
f5	127.5	130.2	127.5
f6	180.3	19.0	19.1
f7	180.3	200.3	202.5
f8	127.5	130.4	127.5
f9	180.3	199.4	202.1
f10	180.3	195.3	198.2
f11	180.3	195.8	198.1

Table 7: Standard errors of the model coefficients from a standard regression (Regr.), from the bootstrap (Bootstrap) and from application of the unequal error variances to the regression (WRegr).

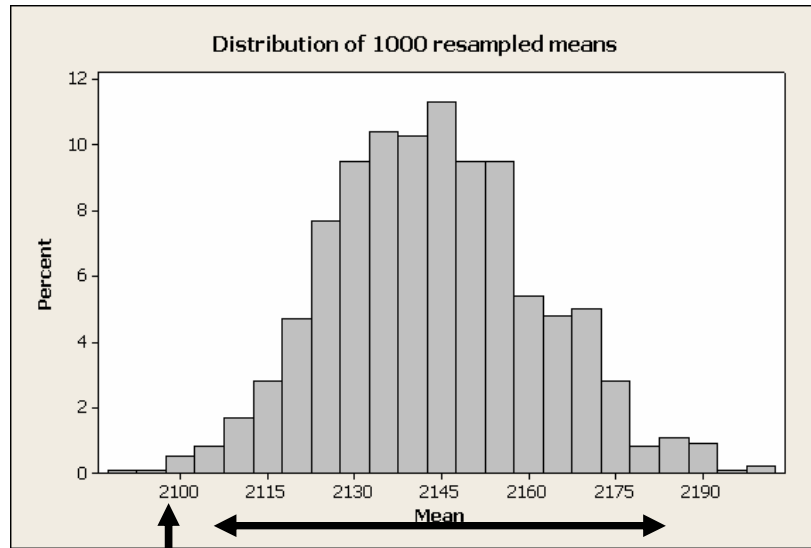


Figure 1: Histogram of 1000 resampled means from sample of 32 hybrid microcircuits inter-layer resistance in ohms with 95% Empirical Bootstrap Confidence Interval and Design Target.

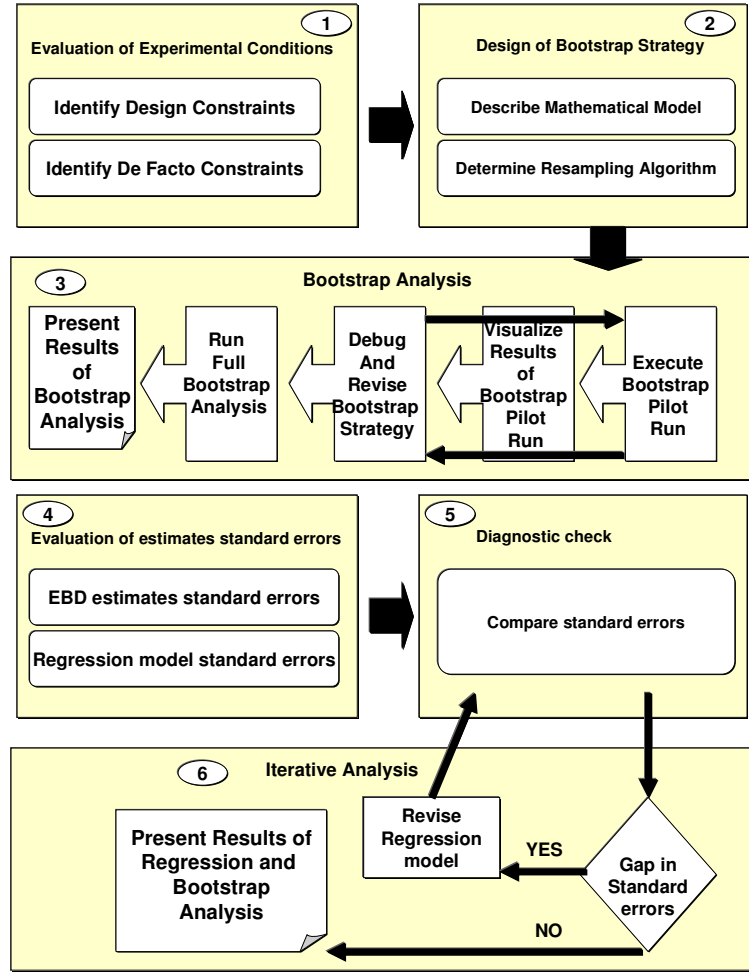


Figure 2: A six steps experimentation data analysis strategy with bootstrapping analysis (1-3) and bootstrapped diagnostic checks (4-6)