23rd International Congress on Modelling and Simulation, Canberra, ACT, Australia, 1 to 6 December 2019
mssanz.org.au/modsim2019

# Two common pitfalls applying design of experiments (and hopefully how to avoid them!)

**A. Gill** [a]

[a]*Defence Science and Technology Group, PO Box 1500, Edinburgh, SA, 5111*
*Email: andrew.gill@dst.defence.gov.au*

**Abstract:** The Defence Science and Technology Group, as part of their Modelling Complex Warfighting Strategic Research Investment, has been developing a prototype simulation depicting joint warfighting. The Joint Future OpeRating Concept Explorer (JFORCE) is an agent-based, stochastic simulation, where the parameters define the physical attributes of the entities, control their behaviour, or define a particular scenario. The Design of Experiments (DOE) is a structured investigation through this high-dimensional parameter-space and the simulation's stochastic response in order to support a particular analytical objective. Characterising the form and strength of the sensitivity of the simulation's response to changes to factor settings can provide insight into sub-system/attribute contributions to joint warfighting operational effectiveness and the trade-space between them. This paper sets out to highlight two of the more common pitfalls analysts might face when conducting such a sensitivity analysis of stochastic simulations.

Regression fits a model $\hat{y}(\mathbf{x}, \hat{\boldsymbol{\beta}})$ where the coefficients $\hat{\boldsymbol{\beta}}$ (which reflect the sensitivity of the parameters) are chosen to make the model close to the simulation response at a number of user-specified design points and replications. A very common choice is to consider a *baseline* scenario and other scenarios where only one parameter is changed at a time. This *One Factor At a Time* (OFAT) design intuitively makes sense, but it's a trap for new players. The second issue is that some regression software generally assume that the simulation responses at the design points are *independent and identically distributed* (iid), which allows the analysis to be conducted using common (and simpler) *Analysis of Variance* (ANOVA) procedures. But for simulations that employ *common random numbers* the assumption of independence is not met (by design) and the assumption of identically distributed simulation responses at each of the design points can often be found wanting. The aim of this paper is to convince the reader to avoid the temptation to use OFAT designs and to be cautious when using DOE software that rely on iid assumptions.

Now, one should consider the fitted regression coefficients as a point estimate of a random variable $\hat{\mathbf{B}}$, which ideally should have the properties of minimum bias ($\min |E[\hat{\mathbf{B}}] - \boldsymbol{\beta}|$) and maximum precision ($\min var[\hat{\mathbf{B}}]$). A simple example using the JFORCE simulation will hopefully be sufficient to demonstrate the negative implications of relying on OFAT designs and/or iid assumptions. First, it will be shown that the OFAT design contains more bias than an equivalent sized superior design, as well as suffering false negatives (two of three sensitive parameters were not picked up as such). Secondly, even when using this superior design, the iid assumptions will be shown to either under-estimate or over-estimate the regression coefficient confidence intervals, potentially causing false positives (claiming a sensitive parameter when it is not).

The first pitfall (OFAT design) can be avoided if one reads just about any text on DOE. However, one of the classic texts, and some DOE software packages, still espouse the use of traditional ANOVA, thus making avoiding the second pitfall (iid assumptions) less easy for practitioners. This paper, by detailing the required mathematical formulation and illustrating through a small but typical example, potentially offers a useful path forward.

***Keywords:*** *Design of experiments, combat simulation, independent and identically distributed, bias, precision*

## 1 INTRODUCTION

The Defence Science and Technology (DST) Group, as part of their Modelling Complex Warfighting Strategic Research Investment, has been developing a prototype simulation depicting joint warfighting. The Joint Future OpeRating Concept Explorer (JFORCE) is an agent-based model written in the NetLogo language, and is thus a closed-loop, stochastic simulation, where the parameters of the model define the physical attributes of the entities, control their behaviour, or define a particular scenario in which the warfighting is taking place (Au et al. (2018)).

Experimental design is a structured investigation through this high-dimensional parameter-space and the simulation's stochastic response in order to support a particular analytical objective. Thus, design and analysis go hand-in-hand. There are several distinct analytical objectives that are of practical interest in the use of JFORCE:

- From all of the parameters of the simulation which *may* affect its response, identify only the subset that *significantly* affect it (significance here may be both statistical (is the effect greater than zero) and practical (is the effect greater than an indifference threshold)). This is motivated by the parsimony principle (or Occam's Razor) which has often been observed anecdotally. This *Factor Screening* objective can be tackled using specially crafted experimental designs, such as sequential bifurcation (Bettonvil and Kleijnen (1997)).

- Identify the combination of parameter settings that optimise the simulation's response. In joint warfighting this might seek the behavioural parameter values governing tactics that maximises the Blue Force probability of winning. This *Simulation Optimisation* objective is often approached using a response surface methodology (Myers et al. (2016)).

- Characterise the form and strength of the sensitivity of the simulation's response to changes to factor settings. This *Sensitivity Analysis* objective often occurs after Factor Screening and can provide insight into sub-system/attribute contributions to joint warfighting operational effectiveness and the trade-space between them, often by employing generalised linear regression (Dunn and Smyth (2018)).

It is this third analytical objective that this paper will explore. The intent is to expose the reader to two common pitfalls that analysts may encounter when performing Sensitivity Analysis - sometimes through no fault of their own - and to provide details of effective remedies. A simple example using JFORCE will be used for illustration, and it is hoped that this paper contributes to the conversation amongst the design of experiment (DOE) community both at DST Group and further abroad. In particular, an explicit mathematical formulation for the characterisation of the bias and precision of estimated regression coefficients, as a function of a general design and without the typical simplifying assumptions, is provided and its application demonstrated.

## 2 SENSITIVITY ANALYSIS AND ORDINARY LEAST SQUARES REGRESSION

Let $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{iq})$ denote the $i-$th design point (combination of level settings of the $q$ parameters) and let $y_{ir}$ denote the simulation's response at the $i-$th design point and for the $r-$th replication (remembering that the simulation contains stochastic processes). Regression fits a model $\hat{y}(\mathbf{x}, \hat{\boldsymbol{\beta}})$ whereby the regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_q)^T$ are chosen to make the model close to the simulation response $y_{ir}$, $r = 1, \ldots, M$ at each of the design points $\mathbf{x}_i$, $i = 1, \ldots, N$.

While non-linear functions of the $q$ simulation parameters (e.g., $x_j^2$ or $x_j x_k$) can easily be included (by simply defining them as new parameters and increasing the value of $q$), for ease of illustration attention will be restricted to a *main effects* model $\hat{y}(\mathbf{x}, \hat{\boldsymbol{\beta}}) = \mathbf{x}\hat{\boldsymbol{\beta}}$.

The matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T$ is called the *design matrix* and in Ordinary Least Squares (OLS) regression, the model's predicted values $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ are made close to the simulation average responses $\bar{\mathbf{y}} = (\bar{y_1}, \bar{y_2}, \ldots, \bar{y_N})^T$ where $\bar{y_i} = \sum_{r=1}^{M} y_{ir}/M$ by minimising $(\bar{\mathbf{y}} - \hat{\mathbf{y}})^T(\bar{\mathbf{y}} - \hat{\mathbf{y}})$ which results in the so-called *normal equations* $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}^{OLS} = \mathbf{X}^T\bar{\mathbf{y}}$ so that $\hat{\boldsymbol{\beta}}^{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\bar{\mathbf{y}}$.

Thus, given a design matrix $\mathbf{X}$, there exists an equation which estimates the *best* regression model (best here in the sense of minimising the sum of the squares of the residuals, other definitions of best exist as well). This leads to the natural question of whether there is a *best* design matrix? This is where the first common pitfall often arises.

### 2.1 JFORCE SCENARIO

A scenario developed in JFORCE examined the military value of information in a fictitious, geographically-symmetrical scenario. Jets were tasked with the mission of destroying land targets but could engage other (enemy) jets if within sensor range or cued in via a Cooperative Engagement Capability (CEC) system, while other assets could target the jets if within both sensor and weapon ranges. These agent behaviours were controlled by fairly simplistic rules. Further details can be found in Au et al. (2018).

To illustrate the first pitfall the sensitivity of the JFORCE simulation's response to just three parameters ($q = 3$) related to Blue Force's capability will be explored. These are the number of Blue jets ($x_1$), the speed of the Blue jets ($x_2$) and whether Blue has its CEC system turned on ($x_3$). While it is likely that the complexity of having multiple entities and the scenario environment may mean that other (perhaps many other) parameters equally affect the simulation's response, we focus on just these three for illustration purposes.

The minimum and maximum values of the considered ranges of these parameters are linearly scaled to $-1$ and $+1$ as follows: $x_1 = [10, 15], x_2 = [1500 \text{ km/h}, 2000 \text{ km/h}], x_3 = \{\text{FALSE}, \text{TRUE}\}$. The simulation's response of interest ($y$) is the fraction of Blue jets remaining, so JFORCE was replicated one hundred times ($M = 100$) and the average response used. Again, for ease of illustration, a *main effects-only* model will be considered, so that $\hat{y}(\mathbf{x}, \hat{\boldsymbol{\beta}}) = \mathbf{x}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$.

### 2.2 ONE FACTOR AT A TIME DESIGNS

Now OLS regression is in fact just a numerical curve fitting procedure, so having four parameters to estimate implies (at least) four design points ($N = 4$). A very common choice is to consider a *baseline* scenario and the scenarios where only one parameter is changed at a time. This *One Factor At a Time* (OFAT) design intuitively makes sense, as the parameter sensitivities $\hat{\beta}_j$, $j = 1, 2, 3$ can actually be estimated by comparing each simulation response with the baseline (and avoiding the need for calculating matrix inverses). Thus $\mathbf{x}_1 = (1, -1, -1, -1), \mathbf{x}_2 = (1, 1, -1, -1), \mathbf{x}_3 = (1, -1, 1, -1), \mathbf{x}_4 = (1, -1, -1, 1)$ and for these design points 100 replications of JFORCE resulted in $\bar{\mathbf{y}} = (0.305, 0.259, 0.157, 0.325)^T$. Applying the OLS equations results in the estimated regression model $\hat{y}(\mathbf{x}, \hat{\boldsymbol{\beta}}) = 0.218 - 0.023x_1 - 0.074x_2 + 0.010x_3$.

Relative to the baseline scenario, which has an estimated fraction of Blue jets remaining of $30.5\%$, the estimated effects of the three Blue attributes can be clearly seen. Increasing the number of Blue jets (from 10 to 15) decreases the fraction of Blue jets remaining by an estimated $4.6\%$, while increasing the speed of the Blue jets (from 1500 km/h to 2000 km/h) decreases the fraction of Blue jets remaining by an estimated $14.8\%$. Finally, the estimated effect of Blue turning its CEC system on is to increase the fraction of Blue jets remaining by an estimated $2\%$. Analysts might use this information to advise decision-makers on the relative merits of differing capability options.

So where is the pitfall alluded to? Well, the analyst should first be asking him/herself the questions *is* $-0.023$ *(or* $-0.074$ *or* $0.010$*) statistically different to* $0$*?* That is, are the effects *real* or are they an artefact of the stochastic nature of the simulation. Secondly, the analyst should be wary of the choice of a main effects-only regression model, and should be asking *are the estimated effects due solely to the parameters explicitly modelled?*

These questions call for the consideration of the regression coefficients $\hat{\boldsymbol{\beta}}$ as *random variables* and the associated properties of *bias* and *precision*, and it is here that the OFAT design proves wanting. While the DOE literature does discuss limitations with OFAT designs (e.g., Law (2007), Montgomery (2012), Kleijnen (2015)) they do so without considering equal sized designs (making efficiency comparisons harder), nor the impact on bias or hypothesis testing (making accuracy comparisons harder). The JFORCE example in this paper will provide a clearer illustration of these aspects.

## 3 REGRESSION COEFFICIENTS AS RANDOM VARIABLES

In the example above, the estimated regression coefficients were based on the mean simulation response from a sample of $M = 100$ replications (for each of the design points). If another sample of $M$ replications were produced, then one might expect to obtain different estimates, as the sample means would likely differ. This suggests that one should consider $\hat{\boldsymbol{\beta}}^{OLS}$ as a point estimate of the random variable $\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\bar{\mathbf{Y}}$. Ideally, this $\hat{\mathbf{B}}$ should have the properties of minimum bias ($\min |E[\hat{\mathbf{B}}] - \boldsymbol{\beta}|$) and maximum precision ($\min var[\hat{\mathbf{B}}]$). The question now is how does the design matrix $\mathbf{X}$ affect bias and precision?

The fact that OLS regression has coefficients which are *linear* functions of random variables (the mean response $\bar{\mathbf{Y}}$) makes this possible. If the true model is $\bar{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + \boldsymbol{\zeta}$ (with $E[\boldsymbol{\zeta}] = \mathbf{0}$) where $\tilde{\boldsymbol{\beta}}$ is the vector of simulation parameters not included in the regression model (here it would consist of all the interaction terms $x_i x_j$ and $x_i x_j x_k, \forall i < j < k \in \{1, 2, 3\}$) and where $\tilde{\mathbf{X}}$ is composed by multiplying the corresponding columns in $\mathbf{X}$ according to the entries in $\tilde{\boldsymbol{\beta}}$ then:

$$\begin{aligned} E[\hat{\mathbf{B}}] &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E[\bar{\mathbf{Y}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\left(\mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} + E[\boldsymbol{\zeta}]\right) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{X}}\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + A\tilde{\boldsymbol{\beta}} \end{aligned}$$

where $A = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\tilde{\mathbf{X}}$. Matrix $A$ is known as the *alias matrix*, as it not only quantifies the amount of bias in the regression estimate, but more importantly it indicates which other parameter(s) might be contributing some effect to the sensitivity of the simulation's response (but not explicitly modelled). For the OFAT design with $\tilde{\boldsymbol{\beta}} = (\beta_{12}, \beta_{13}, \beta_{23}, \beta_{123})^T$, the alias matrix is:

$$A = \left(\begin{bmatrix} 4 & -2 & -2 & -2 \\ -2 & 4 & 0 & 0 \\ -2 & 0 & 4 & 0 \\ -2 & 0 & 0 & 4 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & -1 \\ -1 & -1 & 1 & -1 \\ -1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} -1 & -1 & -1 & 2 \\ -1 & -1 & 0 & 1 \\ -1 & 0 & -1 & 1 \\ 0 & -1 & -1 & 1 \end{bmatrix}$$

so (for example) $E[\hat{\beta}_1] = \beta_1 - \beta_{12} - \beta_{13} + \beta_{123}$. What this means is that (on average) the estimate for the effect of the number of Blue jets is equal to the true effect minus the effects of the two-way interactions between the number of Blue jets and the Blue jet speed and between the number of Blue jets and the Blue CEC system, plus the effect of the three-way interaction between all three parameters.

Now it is possible that these interactions are not strong (as assumed) and it is generally the case that the higher the order of the interaction the weaker its effect size is. But a design that generates a sparse alias matrix $A$ with small-magnitude non-zero entries would be preferred.

The linearity of the OLS regression coefficients also allows the covariance function for $\hat{\mathbf{B}}$ to be explicitly derived. Using the bi-linearity property this can be written as $\Sigma[\hat{\mathbf{B}}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\Sigma[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}/M$. Ideally, we want the entries in the covariance matrix to be small (specifically the diagonal entries) as they control the width of the confidence intervals associated with each of the regression parameters. The common assumption is that $\mathbf{Y}$ are independent and identically distributed (iid) random variables in which case $\Sigma[\mathbf{Y}] = \sigma^2\mathbf{I}$ (where $\sigma^2$ is the population constant variance) and the covariance matrix for $\hat{\mathbf{B}}$ simplifies to $\Sigma[\hat{\mathbf{B}}] = \frac{\sigma^2}{M}(\mathbf{X}^T\mathbf{X})^{-1}$. Some simple matrix calculations shows that for the OFAT design:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 1 & 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1/4 & 1/4 \\ 1/2 & 1/4 & 1/2 & 1/4 \\ 1/2 & 1/4 & 1/4 & 1/2 \end{bmatrix}$$

so that $var[\hat{\mathbf{B}}] = diag(\Sigma[\hat{\mathbf{B}}]) = \frac{\sigma^2}{M}(1, 0.5, 0.5, 0.5)^T$. So the variance of the estimates of the three parameter effects are equal and half that of the constant in the regression model. Whether that is a good result will be examined next.

## 4 FRACTIONAL FACTORIAL DESIGNS

While the OFAT design has an intuitive appeal, it is not the best design. While the simple example above doesn't require it, it should be apparent that OFAT has no possibility of estimating interactions should such a regression model be sought. But even for main effects-only models, an alternative design of the same size can be found which has better properties in terms of bias and precision.

For the simple example above, this alternative design only requires replacing the first design point (our baseline scenario) with $\mathbf{x}_1 = (1, 1, 1, 1)$. The JFORCE simulation returned $\bar{y_1} = 0.341$ over the 100 replications. The matrix equations for the point estimates, bias and precision can be used to see what effect this simple change has. Now the estimated regression model becomes $\hat{y}(\mathbf{x}, \hat{\boldsymbol{\beta}}) = 0.270 + 0.029x_1 - 0.022x_2 + 0.062x_3$ which suggests quite different effects of the Blue force characteristics. Here, the largest predicted effect is when Blue turns its CEC system on (fraction of Blue jets remaining increasing by $12.4\%$) and the effect of the number of

Blue jets is in the opposite direction from that predicted by OFAT. For bias, the alias matrix is:

$$A = \left( \begin{bmatrix} 4 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

so for example $E[\hat{\beta}_1] = \beta_1 + \beta_{23}$. Here the estimate for the effect of the number of Blue jets is equal to the true effect plus only the effect of the two-way interaction between the Blue jet speed and the Blue CEC system. Compared with the aliasing associated with the OFAT design, which had three confounding effects, this is clearly better. The same is true for the other regression coefficients.

Regarding precision, this alternative design yields $\Sigma[\hat{\mathbf{B}}] = \frac{\sigma^2}{4M}\mathbf{I}$ (note: $\mathbf{X}^T\mathbf{X}$ is a diagonal matrix and easily invertible), thus $var[\hat{B_j}^{OFAT}] = 2 * var[\hat{B_j}^{ALT}]$, $j = 1, 2, 3$. Compared with the OFAT design, confidence intervals for the regression coefficients with this alternative design should be around $40\%$ smaller. For the simple example above, the $95\%$ confidence intervals based on the OFAT design (the population variance can be estimated from the sample variance, so $\sigma^2 \approx \sum_{i=1}^{4} s_i^2/4 = 0.0412$) suggest that only the Blue jet speed significantly affects the fraction of Blue jets remaining.

However, the smaller confidence intervals associated with the alternative design (with $\sigma^2 \approx 0.0473$) actually allows concluding that all three parameters are significant. This *variance inflation* property of OFAT designs risks making more *false negatives* (i.e., misidentifying a significant effect, therefore reducing the power of the test) than is necessary.

Hopefully, this simple example is sufficient to convince the reader of the sub-optimal properties (increased bias and reduced precision) of the intuitively popular design choice of OFAT, and in fact, the relative reduction in precision gets worse as the number of parameters grows (Montgomery (2012, chapter 5.2)). What then is this alternative design, and how can it be constructed in the general $q$ parameter setting?

The above alternative design is actually a two-level *fractional factorial* design. One characteristic of these designs is their *resolution* which denotes their ability to reduce the bias in the estimated regression coefficients (the higher the resolution the better). Resolution $V$ fractional factorials (for reduced bias), augmented with centre points (for modelling non-linearity) are popular designs. Thankfully, many text-books explain fractional factorial designs, and one of the most popular is Montgomery (2012). Unfortunately, the popularity of this text, and of commonly used regression software packages, may be where the second common pitfall often arises.

## 5 ANALYSIS OF VARIANCE BASED REGRESSION ('CLASSIC' DOE)

The issue is that both Montgomery (2012) and common regression software (e.g., Minitab and JMP) seem to generally assume that the simulation responses at the design points are iid, as this allows the analysis of the regression coefficients to be conducted using common (and simpler) *Analysis of Variance* (ANOVA) procedures. While this does simplify the calculations required, as noted above the linearity of the OLS regression does allow the covariance matrix to be explicitly written (see Section 3.2).

So for the design and analysis of simulation experiments, are the iid assumptions likely to be violated, and what effect does this have on the regression coefficient confidence intervals?

Considering *independence* first, for simulations that employ *common random numbers* (CRN) the assumption of independence of the simulation's responses at the design points is not met (by design). CRN can be an effective *variance reduction technique* (VRT) that assists in multiple comparison statistical tests of alternative options, is helpful in the debugging phase of simulation scenario development, and is the default setting of the combat simulation used in the author's branch of DST. While ANOVA may still be applicable (CRN are a form of blocking, which can be added as an explicit parameter) it would only address one of the two assumptions.

As for the assumption of *identically distributed* simulation responses at each of the design points, Law (2007) discusses examples where the ratios of largest to smallest variance exceed an order of magnitude. For the simple JFORCE example, CRN do cause covariance between the responses at different design points and the response variance is not constant across the design space.

The resultant estimated covariance matrices for the regression coefficients using the fractional factorial design, and assuming iid or not, can be calculated and the variances extracted from the diagonals. This produces

$var[\hat{\mathbf{B}}_{iid}] = (1.18 \times 10^{-4}, 1.18 \times 10^{-4}, 1.18 \times 10^{-4}, 1.18 \times 10^{-4})^T$ and $var[\hat{\mathbf{B}}] = (1.28 \times 10^{-4}, 1.33 \times 10^{-4}, 1.17 \times 10^{-4}, 9.00 \times 10^{-5})^T$ which means that $var[\hat{\mathbf{B}}_{iid}]/var[\hat{\mathbf{B}}] = (0.92, 0.89, 1.01, 1.31)^T$.

Note that the estimated variance for each regression coefficient is constant, meaning that each confidence interval will have the same (half) width - this is also known as Fisher's Least Significant Difference. Thus the confidence intervals for the regression coefficients would be either under- or over-estimated if one simply used the common iid assumptions.

While this doesn't produce a 'gotcha' moment for this simple main effects-only example (i.e., all simulation parameters are classified as significant in both cases), it should be noted that in a larger regression model incorporating interactions between parameters, one two-way interaction (between the number of Blue Jets and the Red Force CEC system) was incorrectly classified as being a significant influence on the fraction of Blue Jets remaining when the iid assumptions were used (a false positive).

## 6 KLEIJNEN-LAW REGRESSION ('MODERN' DASE)

Hopefully by now the reader will have been convinced of two things - that OFAT designs should be replaced by a proper DOE, and that iid simulation responses need not be assumed. While Montgomery (2012) is perhaps the seminal text on what could be described as *classical* DOE, it is perhaps the lesser known Kleijnen (2015) which is the seminal text on *simulation* DOE, where the above remedies to violations of the iid assumptions (and others) are described. However, recent research by one of the authors of Gill et al. (2018) regarding Professor Kleijnen's text is worth repeating here, as it may assist analysts in following the procedures contained within.

First, it is possible that the analyst might wish to use a different number of replications at each design point ($M_i \neq M$, $i = 1, \ldots, N$) perhaps motivated by the differing variability noted above. While Kleijnen (2015) treats the cases of constant and non-constant number of replications separately, the OLS normal equations can in fact be generalised to $\mathbf{X}^T\mathbf{M}\mathbf{X}\hat{\boldsymbol{\beta}}^{OLS} = \mathbf{X}^T\mathbf{M}\overline{\mathbf{y}}$ where $\mathbf{M}$ is an $N \times N$ diagonal matrix with entries $M_{ii} = M_i$.

Second, when CRN are used (which does require a constant number of replications), in an effort to avoid having to estimate the full $N \times N$ covariance matrix $\Sigma[\mathbf{Y}]$, Kleijnen (2015) proposes a remedy initially suggested in a seminal text on simulation modelling (Law (2007)). The very simple idea is to compute a point estimate $\hat{\boldsymbol{\beta}}_r = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_r$ for each replication $r = 1, \ldots, M$, where $\mathbf{y}_r = (y_{1r}, y_{2r}, \ldots, y_{Nr})^T$ from which the sample means and sample variances are used to construct the associated confidence intervals. However, it turns out that due to the linearity of the regression coefficient estimators, this 'alternative' approach is in fact identical to that described above.

Third, the other assumption often challenged by combat simulation response data is that of normality. When this is the case, an approach suggested by Kleijnen (2015) is to use *jackknifing*. There, the $r$-th jackknifed *pseudovalue* (a weighted difference of the OLS estimators based on the simulation response averaged over all $M$ replications and the simulation response averaged over all replications excluding the $r$-th) is computed for each replication, and the sample means and variances of these pseudovalues are used to construct the confidence intervals. However it is relatively easy to prove that this jackknifing is also identical to that described above.

Finally, I have not yet commented on how to determine how accurate the fitted OLS regression model is. One might be able to use the estimated confidence intervals for the fitted regression coefficients as a guide (i.e., how close they include the value zero perhaps). But there is a better way, by using the so-called *lack-of-fit F-statistic*.

In the iid case, a ratio of two different estimates of the population variance $\sigma^2$ is used, one based on the fitted regression model (numerator) and one which doesn't (denominator). If the regression is a poor fit, the numerator will increasingly overestimate the population variance, i.e., larger values of the ratio. A statistical assessment of the regression fit can then be made against the critical value from the $F$-distribution with $N - q$ and $N(\sum_{i=1}^{N} M_i - 1)$ degrees of freedom.

In Gill et al. (2018) I claimed that Kleijnen (2015) was incorrect, in that his numerator 'represented the sum of weighted squared average residuals, and as such, risks some residuals cancelling each other out in the calculation of the average residual at each design point and therefore underestimating the Mean Squared Residual (MSR)' and 'risks suggesting an adequate regression when it may not be so'.

Technically, the first part of the claim regarding the MSR is correct. However, it is not the MSR that

should be the numerator. As Montgomery (2012) points out, one can show that: $\sum_{i=1}^{N}\sum_{r=1}^{M_i}\left(y_{ir}-\hat{y}_i\right)^2 = \sum_{i=1}^{N}\sum_{r=1}^{M_i}\left(y_{ir}-\bar{y}_i\right)^2 + \sum_{i=1}^{N} M_i\left(\bar{y}_i-\hat{y}_i\right)^2$ and while the LHS is the (correct) MSR, it is the ratio of the (correctly scaled into *Mean Squares*) terms on the RHS, which measure *pure-error* and *lack-of-fit*, which is the correct lack-of-fit $F$-statistic (as both approximate the population variance). So, the correct equation (as given in (2.30) of Kleijnen (2015)) is:

$$F_{N-q,N(\sum_{i=1}^{N} M_i-1)} = \frac{\sum_{i=1}^{N}\sum_{r=1}^{M_i}(\bar{y}_i-\hat{y}_i)^2/(N-q)}{\sum_{i=1}^{N}\sum_{r=1}^{M_i}(y_{ir}-\overline{y}_i)^2/(N(\sum_{i=1}^{N} M_i-1))}.$$

## 7 CONCLUSIONS

This paper set out to highlight two of the more common pitfalls analysts might face when conducting a Sensitivity Analysis of stochastic simulations. The aim was to convince the reader to resist the temptation to use OFAT designs and to be cautious when using DOE software that rely on iid assumptions.

A simple example using a combat simulation in development by DST Group was hopefully sufficient to demonstrate the negative implications in terms of bias or precision of failing to do so. It was shown that the OFAT design contained more bias than an equivalent-sized fractional factorial design, and suffered more false negatives. When using the fractional factorial design, the iid assumptions were shown to either under-estimate or over-estimate the size of the regression coefficient confidence intervals, potentially causing a false positive.

The first pitfall (OFAT design) should be avoided if one reads just about any text on DOE. However, one of the classic texts on DOE (Montgomery (2012)), as well as some DOE software packages, espouse the use of traditional ANOVA, thus making avoiding the second pitfall (iid assumptions) less easy.

The simulation focussed text on DOE (Kleijnen (2015)) and classic text focussed on simulation (Law (2007)), along with the author's recent modest contribution (Gill et al. (2018)), potentially offers a useful path forward, in particular, the explicit mathematical formulation for the characterisation of the bias and precision of estimated regression coefficients as functions of a general design and without the typical simplifying assumptions.

## ACKNOWLEDGMENTS

The author thanks Professor Jack P.C. Kleijnen for fruitful discussions surrounding regression analysis; Kevin Clark and Jessica Penfold from DST Group for the JFORCE CRN script and computing the various regression coefficient confidence intervals, respectively, and the reviewers for their constructive comments.

## REFERENCES

Au, T. A., P. J. Hoek, and E. H. S. Lo (2018). Combat analysis of joint force options using agent-based simulation. In *2018 Military Communications and Information Systems Conference (MilCIS)*, pp. 1–7.

Bettonvil, B. and J. P. Kleijnen (1997). Searching for important factors in simulation models with many factors: Sequential bifurcation. *European Journal of Operational Research 96*(1), 180 – 194.

Dunn, P. and G. Smyth (2018). *Generalized Linear Models With Examples in R*. Springer Texts in Statistics. Springer New York.

Gill, A., D. Grieger, M. Wong, and W. Chau (2018). Combat simulation analytics: Regression analysis, multiple comparisons and ranking sensitivity. In *Proceedings of the 2018 Winter Simulation Conference*, WSC '18, Piscataway, NJ, USA, pp. 3789–3800. IEEE Press.

Kleijnen, J. (2015). *Design and Analysis of Simulation Experiments* (2nd ed.). New York, USA: Springer.

Law, A. (2007). *Simulation Modeling and Analysis* (4th ed.). Boston, USA: McGraw-Hill.

Montgomery, D. (2012). *Design and Analysis of Experiments, 8th Edition*. John Wiley & Sons, Incorporated.

Myers, R., D. Montgomery, and C. Anderson-Cook (2016). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley Series in Probability and Statistics. Wiley.