FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE D COIMBRA

Paulo Francisco Constantino Valente

Data-Driven Quality by Design for Complex Generic Drug Products

Thesis submitted to the University of Coimbra
for the degree of Master in Biomedical Engineering,
specialization in Clinical Informatics and Bioinformatics

Supervisors:
Prof. Dr. Marco Seabra dos Reis
Prof. Dr. Cláudia Sousa Silva

September, 2019

This work was developped in collaboration with:

Process Chemometrics Laboratory
Process Systems Engineering Group
Chemical Process Engineering and Forest Products Research Centre
Faculty of Sciences and Technology of Univeristy of Coimbra

Bluepharma - Indústria Farmacêutica

Resumo

A indústria farmacêutica é uma das atividades mais inovadoras e regulamentadas, sendo a distribuição dos componentes terapêuticos no corpo humano, com a qualidade desejada, um dos grandes focos da investigação nesta área. Para garantir que a qualidade final dos medicamentos seja uma preocupação em todas as etapas do seu desenvolvimento, as entidades reguladoras têm incentivado as empresas a adotarem princípios de qualidade pelo design, o que promove mais conhecimento sobre o processo e permite reduzir os recursos necessários, o que, em última instância, torna os cuidados de saúde mais acessíveis para todos.

Neste trabalho, são analisados dados experimentais do processo de fabrico de um medicamento genérico complexo, a ser atualmente desenvolvido pela Bluepharma - Indústria Farmacêutica. Estes produtos são conhecidos por implicarem um esforço adicional, pois o seu desenvolvimento envolve tarefas mais complicadas do que os medicamentos convencionais. Além disso, neste problema, pretende-se que seis respostas diferentes, discretas e contínuas, sejam otimizadas simultaneamente.

Um novo método para identificar efeitos ativos em experiências de triagem é proposto, envolvendo o uso de regressão passo-a-passo com a implementação de hereditariedade de efeitos, modelos lineares generalizados e validação através do critério de informação de Akaike corrigido. Esta abordagem é mais simples do que as sugeridas na literatura com objectivos semelhantes e apresenta resultados muito melhores do que as técnicas padrão utilizadas normalmente na indústria farmacêutica.

Além disso, alguns outros procedimentos são realizados para extrair informações importantes dos dados disponíveis, como o estudo dos melhores níveis de cada fator para cada resposta, otimização de várias respostas simultaneamente, análise dos fatores não controlados e a criação de modelos preditivos. A combinação de todos estes métodos permite uma maior compreensão do processo de desenvolvimento e fornece novas técnicas auxiliares, ajudando a atingir as caraterísticas pretendidas do produto muito mais eficientemente.

Palavras chave: qualidade farmacêutica pelo design, medicamentos genéricos complexos, experiências de triagem, hereditariedade de efeitos, otimização simultânea de múltiplas respostas

Abstract

The pharmaceutical industry is one of the most innovative and regulated activities, and the delivery of the therapeutic components with the desired quality is one of the biggest concerns in pharma R&D. In order to ensure that the final quality of the drug products is a focus in all stages of the development, the regulatory agencies have encouraged the companies to adopt quality by design principles, which promotes more knowledge about the process and allows to reduce the required resources, and ultimately to make the healthcare more affordable for everyone.

In this work, it is analyzed the experimental data of the production process of a complex generic drug, being currently developed by Bluepharma - Indústria Farmacêutica. These products are known for implying an additional effort as their development comprises harder tasks compared to conventional drugs. Besides, in this problem, six different responses, both discrete and continuous ones, are expected to be simultaneously optimized.

A new method to identify active effects in screening experiments is proposed, involving the use of stepwise regression with the enforcement of effects heredity, generalized linear models and corrected Akaike information criterion validation. This approach is simpler than the same-purpose ones suggested in the literature and it was found to perform much better than the standard techniques executed usually in the pharmaceutical industry.

Besides, some other procedures are considered to retrieve important information from the available data, such as the study of the best levels of each factor for each response, optimization of multiple responses simultaneously, analysis of non-controlled factors, and creation of predictive models. The combination of all these methods provides a better understanding of the development process and make available new auxiliary techniques, aiding to achieve the targets much more efficiently.

Keywords: pharmaceutical quality by design, complex generic drug products, screening experiments, effects heredity, multiple response optimization

À Coimbra das lições, dos sonhos e tradições,
dos doutores e das canções.
Que continues a ser a fonte dos amores.

Agradecimentos

Começo por expressar um sincero e grande obrigado às pessoas que me acompanharam e auxiliaram na realização deste trabalho. Ao professor Marco Reis, cujas reconhecidas competências intelectuais e clareza de ideias ajudaram a delinear os detalhes que tanta diferença fazem. À doutora Cláudia Silva, pela expressa vontade em pôr à disposição os meios necessários e possíveis, e por acreditar no potencial deste trabalho. À Ana Sofia Lourenço, pela paciência para esclarecer as minhas muitas questões, que permitiram uma maior compreensão do problema. A vossa pronta disponibilidade em ajudar no que fosse necessário e as nossas discussões foram, sem dúvida, uma grande contribuição para o sucesso deste trabalho.

Gostaria ainda de deixar uma palavra de apreço à Bluepharma, pela oportunidade de ter contribuido para um projeto do "mundo real". A vontade da empresa em melhorar a forma como os seus dados são analisados mostrou o quão frutífera pode ser a sinergia entre os mundos académico e industrial. Se, por um lado, esta colaboração levou ao meu desenvolvimento de novas competências, por outro, mostrou o quão esta ligação (por vezes escassa) é benéfica para o avanço da ciência.

Sendo este trabalho o culminar de um caminho de cinco anos, não poderia deixar de dar um forte agradecimento a todos os colegas e amigos com quem tive a felicidade de partilhar esta etapa. Sem dúvida, fizeram destes anos uma experiência incrível e única, a nível profissional e sobretudo a nível pessoal. Ao BEST Coimbra, obrigado por me fazer compreender ainda melhor a filosofia *work hard play harder*, com a qual cada vez mais me identifico.

Por fim, um grande obrigado à minha família. Em especial ao meu irmão, pelo exemplo de superação de desafios e vontade de querer sempre mais. E ainda mais especial aos meus pais, pelo esforço e apoio incondicional, por celebrarem cada pequena conquista minha com tanto ou mais entusiasmo do que eu, e por me possibilitarem e incentivarem a aproveitar as oportunidades que me fizeram viver os últimos anos tão intensamente. Agora podem descansar um pouco, por enquanto...

Contents

List of Tables

List of Figures

Glossary

AIC Akaike Information Criterion
AICc Corrected Akaike Information Criterion
ANOVA Analysis of Variance
API Active Pharmaceutical Ingredient

CI Confidence Interval
CMA Critical Material Attribute
CMIM Conditional Mutual Information Maximization
CPP Critical Process Parameter
CQA Critical Quality Attribute
CV Cross Validation

DoE Design of Experiments
D-P D'Agostino-Pearson
DS Design Space

FDA Food and Drug Administration
FFD Full Factorial Design
FP Formulation Parameter

GLM Generalized Linear Model
GoF Goodness of Fit

H0 Null Hypothesis
HNPP Half Normal Probability Plot

K-S Kolmogorov–Smirnov

LARS Least Angle Regression
LASSO Least Absolute Shrinkage and Selection Operator
LR Logistic Regression
LRR Logistic Ridge Regression

MI Mutual Information
ML Maximum Likelihood
MLE Maximum Likelihood Estimation
mRMR Minimum Redundancy Maximum Relevance

NPP Normal Probability Plot

OLS Ordinary Least Squares

PDF Probability Density Function
PI Prediction Interval
PP Process Parameter

QbD Quality by Design
QTTP Quality Target Product Profile

R&D Research and Development
RLD Reference Listed Drug
RSM Response Surface Methodology

SCAD Smoothly Clipped Absolute Deviation
SSVM Stochastic Search Variable Selection
SU Symmetrical Uncertainty
SVM Support Vector Machine
S-W Shapiro-Wilk

TPP Target Product Profile

1

Introduction

1.1 Contextualization

The innovation in pharmaceutical industry is driven mainly by two areas of research and development (R&D): drug discovery and drug delivery. The first one is focused on new active pharmaceutical ingredients (APIs), i.e. new chemical entities responsible for the therapeutic effect. However, in order to ensure that it will reach the target site, the drug product contains along with the API, the excipients, pharmaceutically inert substances that contribute to achieve the attributes included in the target product profile, such as, pharmacokinetic profile, stability, patient compliance, etc. Drug delivery R&D aims at designing and formulating better drug delivery systems able to improve the safety/efficacy ratio of drug products.

The quality of the final drug product is a major focus of drug development in either case. In 2002, the Food and Drug Administration (FDA), the United States federal agency responsible for drug approval, launched a new initiative, *Pharmaceutical Current Good Manufacturing Practices (cGMPs) for the 21st century: A Risk-Based Approach* [1], whose purpose was to modernize the pharmaceutical development and manufacturing in order to improve the quality of pharmaceutical products. This document, along with some others released by the FDA [2, 3] and by the International Conference on Harmonisation [4–6], represented a shift in the pharmaceutical industry philosophy. This new ideology is consistent with the quality by design (QbD) concept that was first developed by Dr. Joseph M. Juran [7].

The QbD is a systematic approach that moves from the more standard experience-based methodology to a more scientific and risk-based approach. These principles promote then a higher understanding of the product and manufacturing process by the industry from the start, building quality into the product instead of testing it (quality by testing) [8]. Therefore, QbD became a way to aid both the pharmaceutical companies and the regulatory authorities, assuring drug product quality and

facilitating the process of approval.

One of the main objectives of QbD is to find the subspace of input parameters (the variables that experimenters can manipulate) which had proven to guarantee that the final goals are met, the so called design space (DS) [6]. Basically, the DS defines the ranges of critical material attributes and process parameters that ensure that the final drug product meets consistently the desired quality requirements.

1.2 Motivation

Movements within a submitted and approved design space do not require a new regulatory approval, only changes beyond it need a regulatory post-approval [6], as they imply deviations that have not been proven before to maintain the output quality. Some insights about the vital importance of a good definition of the DS during the product's process development in pharmaceutical QbD can then be considered.

It greatly facilitates the management of process operations and leads to the increase of knowledge about the product and the process. This results in a more efficient exploitation of the available resources and in reducing the costs associated with quality failures and post-approval changes, during the development of the product. Consequently, a lower product's price for the final consumer is expected, which leads to more affordable healthcare and make the products more competitive than the existing ones – a particularly important factor in specific pharmaceutical market segments such as the generic drug products one.

The development of more recent analytical equipments led to an increase in the amount of recorded data, providing more opportunities to extract useful information from it. The pharmaceutical industry has followed the trend to look for the available data in order to improve its processes R&D, especially since that quality by design was considered a reference methodology. The understanding of the relations between the input and output variables has receiving increasing attention.

Even so, the pharmaceutical industry is still in its early stages of applying data analysis techniques to support decisions when compared to other industries. Most of the times, the companies do not have a specific team with the appropriate data science background and rely on statistical software to perform simple statistical analysis. However, sometimes more advanced methods are necessary to deal with more complex formulations and manufacturing processes. Besides, few studies have been published exploring novel methodologies in this area.

Therefore, the informative potential of the pharmaceutical data remains vastly unexplored, which offers a good opportunity for the study and application of machine learning and advanced statistics for the improvement of pharma processes.

1.3 Objectives and Contributions

In this work, the general goal is to explore the available data from the R&D of a generic complex drug and retrieve important information that can aid the development of the process, improving the way how data is analyzed in pharmaceutical industry. The work development resulted from a step-by-step progress, i.e. some new objectives and challenges were being addressed as some others were being completed, contributing for that the discussions with the Bluepharma experimenters.

The initial goal was to identify important effects when the response is binary. In fact, binary responses are rarely considered in this type of problems. Once a suitable method was obtained, it was verified that it could be easily adapted to handle continuous outcomes as well, and so it was extended to those responses and compared with the traditional pharmaceutical approach that was already been performed by the experimenters.

The identification of active effects, which corresponds to find the critical parameters required to achieve the target responses, is indeed the main focus of the development stage where data come from. However, some other possible objectives were addressed, such as: evaluate non-controlled factors; build scale-independent models; establish a relationship between intermediary responses and the final ones, designing control points in each operation unit of the process, i.e. inline process control; create models to guide future experiments, which includes the definition of the best levels of the factors for each response and for all the responses simultaneously, and the prediction of the outcome values of future trial combinations. Some of these tasks were successfully completed, while others were not accomplished because the approach is not possible at all or because the amount of data is too limited.

This work resulted in several contributions, both for the real-world problem being discussed in this document and for other similar contexts. Regarding the experiments analysis case, the main contribution is:

- Development of a new procedure to identify active effects in screening experiments, which incorporates a modified stepwise regression to enforce effects heredity, the generalized linear model and the corrected Akaike information criterion validation. It has the advantages that it can be used for both discrete

and continuous problems and for several different data structures. Furthermore, it is a very interpretable method which can be implemented and used in a software of common practice by the experimenters, with little effort.

For the specific problem of the development of this complex generic drug product by Bluepharma, the main contributions are:

- The use of the aforementioned method, which allows identifying important interactions (taking their main effects in consideration) while the standard pharmaceutical method (used by the experimenters) only allowed to identify main effects. Besides, it also allows to use different response distributions;
- Getting better and more suitable models than the ones obtained using the traditional approaches;
- The recommendation of the best levels of each one of the selected factors, for each response, and the suggestion of a combination of values that theoretically ensures that all the target responses are achieved simultaneously;
- The study of the non-controlled factors and identification of a possible important variable, which allowed the experimenters to find new paths to meet the project goals more consistently;
- The creation of predictive models to aid future experiments.

In general, all those achievements contributed to a larger knowledge about the process and they are expected to guide the development to accomplish more effectively the target product quality.

1.4 Document Structure

The document is organized to provide a sequential contextualization and learning of the problem. In the **chapter 2 - Background Concepts**, several basic insights are introduced, such as the implementation of quality by design methodology in the pharmaceutical industry and some elemental knowledge about screening experimental design. It constitutes a very useful compendium for non-experimenters. The **chapter 3 - State of the Art**, presents an extensive review about several methodologies proposed in the literature to identify active effects in screening experiments, from more standard and simple approaches to more novel and complex ones.

In the **chapter 4 - Problem Definition and Dataset**, some useful information about the specific case study is provided and the initial analysis of the problem is

performed. The **chapter 5 - Methodologies**, defines the pipeline used to achieve the stated objectives and a further description of all approaches is contemplated, in order to better understand how and why they were implemented. It constitutes also a very useful compendium for non-data scientists.

In the **chapter 6 - Results and Discussion**, the work results are shown and interpreted and the whole process is discussed, contemplating achievements, limitations and also a comparison between the proposed procedure and the standard one. Finally, the **chapter 7 - Conclusions**, addresses some final remarks and some considerations about future work.

2

Background Concepts

In order to better understand the problem being studied and the methods further used, it is useful to get some insights about the quality by design (QbD) approach in the pharmaceutical industry and in the generic drug products pharma segment. In a QbD methodology, data being analyzed results normally from experimental trials. So, some knowledge about experimentation and design of experiments more specifically is considered as well.

Many investigators in machine learning area only deal with observational data, so it is important to focus on some central aspects of experimental data. Experimentation is one of the most used activities in several industries. It allows to investigate how purposeful changes in the settings of input variables in a system affect the output, helping to get more and useful information of the system and how it can be improved. These systems are mostly processes; nowadays, experiments are commonly used to process modelling and its optimization.

2.1 Pharmaceutical Quality by Design

In the pharmaceutical area, the QbD approach starts with the identification of the target product profile (TPP), beginning then with the final goals in mind. The features of TPP may include description, indications and contraindications, dosage and route of administration forms and adverse reactions, among others [9]. The TPP is then a prospective summary of the medicinal product attributes for the intended commercial product based on all customers and end users needs. It must also meet the demands of payers and government agencies. Therefore, it constitutes a guide for product development.

Based on it, a quality target product profile (QTPP) is established, where are also identified the potential critical quality attributes (CQAs) of a drug product.

A CQA is a physical, chemical or biological property or characteristic that should have values between an appropriate limit, range or distribution in order to assure that the desired quality defined by the QTPP is met [6]. The quality attributes of a drug product can be drug content, content uniformity, dissolution ratio, stability, etc. They can be extremely relevant to achieve the target quality (becoming a CQA) or not (non-CQA). Their criticality is evaluated through a risk assessment, based on the severity of something goes wrong, i.e. how harmful for the patient is if the product is outside of the acceptable range for that attribute, and the uncertainty level of the knowledge [10, 11].

Once the CQAs are initially identified, the objective is to set up the group of input parameters that have an impact on the desired quality attributes. These parameters may be related with components of the formulation, critical material attributes (CMAs), or process related, critical process parameters (CPPs). The CPPs refers to parameters whose variability impacts the CQA and then should be controlled or monitored to ensure that the target quality is achieved consistently [6]. Contrary to CQAs and CPPs, CMAs are not defined by the International Conference on Harmonisation. However they have been extensively used with a close definition to the CQAs (physical, chemical or biological property or characteristic that should place between an appropriate limit, range or distribution to ensure the desired quality) but they are referred to input materials such as the drug substance or other excipients that are part of drug product composition [10]. The process parameters may include features like blending, speed, temperature and pressure; the material attributes can be, for example, particle size distribution, polymer grades, specific surface area.

Again, it is necessary a methodology to determine if these material and process attributes are critical or not. This is normally a data-driven approach, performed through the analysis of experimental trials. In order to detect the complexities and interactions when several input parameters are varied across its defined ranges, multivariate techniques are usually applied to determine which ones are critical ones and how they impact the CQAs.

A pharmaceutical manufacturing process is often composed by a series of unit operations, which are discrete activities that comprise physical and/or chemical transformations such as mixing, drying, filtration, evaporation or dilution. The output of each unit operation becomes the input of the next one and the analysis of the process can be performed on each individual unit operation or on the combination of all unit operations of the manufacturing process. In fact, a given CQA may be a quality

attribute of a specific unit operation (intermediate CQA) or may be a quality attribute of the final medicinal product, resulting then from the joint-transformations of all units [10]. The relationship between input and output variables is shown in figure 2.1 for a single unit operation.

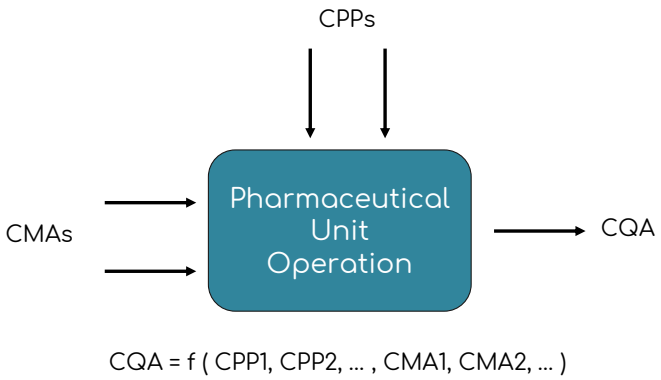


Figure 2.1: Relationship between CPP, CMA and CQA. Adapted from Yu et al. [10].

This relationship between input and output variables is known as the knowledge space, which is expected to provide useful information about the design space (DS). The DS establish then the multidimensional combination and interaction of input variables (CMAs and CPPs) that should be respected to ensure that the CQAs have values within the target limits [6].

However, the companies are encouraged by the regulatory agencies to work in a narrower region inside the DS, which is a more optimized space around the target - the control space. This region is also referred to as the normal operating range, while the DS corresponds to the proven acceptable range. The relationship between knowledge, design and control spaces is represented in figure 2.2.

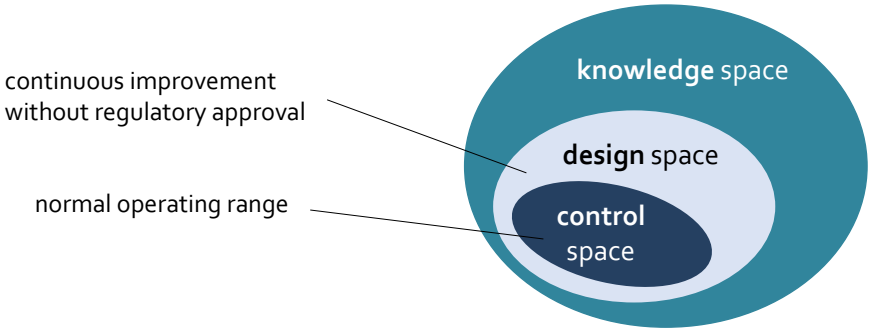


Figure 2.2: Relationship between pharmaceutical quality by design spaces.

2.2 Generic Drug Products

A generic drug product, as defined by FDA [12], is a pharmaceutical product that is therapeutically equivalent to an already approved medicinal product, referred to as the reference listed drug (RLD) or branded product. It is expected that the clinical effect and the safety profile to be the equivalent when they are administered to patients under the same conditions. In order to be considered as such, the generic product must present two characteristics: pharmaceutical equivalence and bioequivalence.

A product is pharmaceutically equivalent to the branded one if it is identical in terms of dosage form, route of administration and active pharmaceutical ingredient (API) [12]; basically, the TPP should be the same for both products. The bioequivalence refers to the absence of a significant difference in the rate and extent of absorption of the API of both generic and RLD products, under the same conditions [12], as demonstrated in one or more clinical trials.

Once the performance is expected to be equivalent in terms of TPP, the CQAs of the generic to be produced will then be the ones of the reference drug. However, some characteristics such as manufacturing process, formulation or excipient may be different from the original ones. Thus, in this type of drug products, the main focus is on obtaining a set of critical material and process parameters, and define a design space from it.

A special case of generic drug products are the complex ones. As it name suggests, it is a generic drug that has some complex step during its development (may be a complex API, formulation, dosage form or route of delivery, for example), such as defined by FDA [13]. The development of these products tends then to be more technically and scientifically challenging and resources-consuming than in typical generic products, which leads to a much higher development risk. Positively, the complex generic drug products tend to offer the opportunity to gain competitive advantages.

2.3 Basic Concepts of Experiments

In experiments, the studied inputs variables are called factors. These variables are controlled by the experimenter and they are expected to be independent variables. The output variables are called responses (or outcomes) and they correspond to the variables that are measured from the change of factor values and so they are

dependent variables.

In order to study the effect of a given factor on the response variables, two or more values of the factor are usually used. These values are called factor levels or settings. Each factor can be quantitative (continuous range of values) or qualitative (discrete number of values). On the one hand, there is more flexibility in choosing the (number of) levels of quantitative factors than qualitative ones. On the other hand, the levels of quantitative factor must be chosen with caution: they should have values far enough so an effect can be detected but they should belong to a range considered as acceptable due to chemical and physical properties [14]. Besides, when choosing the levels of a factor, other constraints such as the associated cost may also need to be taken into consideration [14]. All these intervals should then be chosen to assure the feasibility of the experimental trials.

From the independent variables that are studied in a given experiment, only some of them present a statistically significant impact on the system response. These are referred to as active factors [15]. When evaluating the effects of a factor, we may consider the effects of both main factors and interactions. The main effect is then the effect of one of the independent variables on the dependent one, without taking into consideration the effects of the remaining factors [15].

However, sometimes the impact of a change in the levels of one of the factors on the response depends on the value of another independent variable (main factors may influence each other) [16]. In fact, this joint effect is different from the sum of the individual effects of those factors (*the whole is greater than the sum of its parts* [attributed to Aristotle]). Often this type of effects is worth to study and it is said to be an interaction effect. An interaction is modeled by the simple product of two or more main factors. For example, an interaction that results from the joint effect of two variables is called as a two-factor interaction. Similarly, an interaction resulting from the combined effect of three variables is a three-factor interaction and so on.

The two-factor interactions can be visually inspected through the so called interaction plots. In these graphs, both variables (that compose the interaction) are plotted simultaneously and it is analyzed how the relationship between each factor and the response varies when the level of the other factor is changed. Examples of interactions plots are displayed in figure 2.3 for arbitrary interactions AB, resulting from factors A and B.

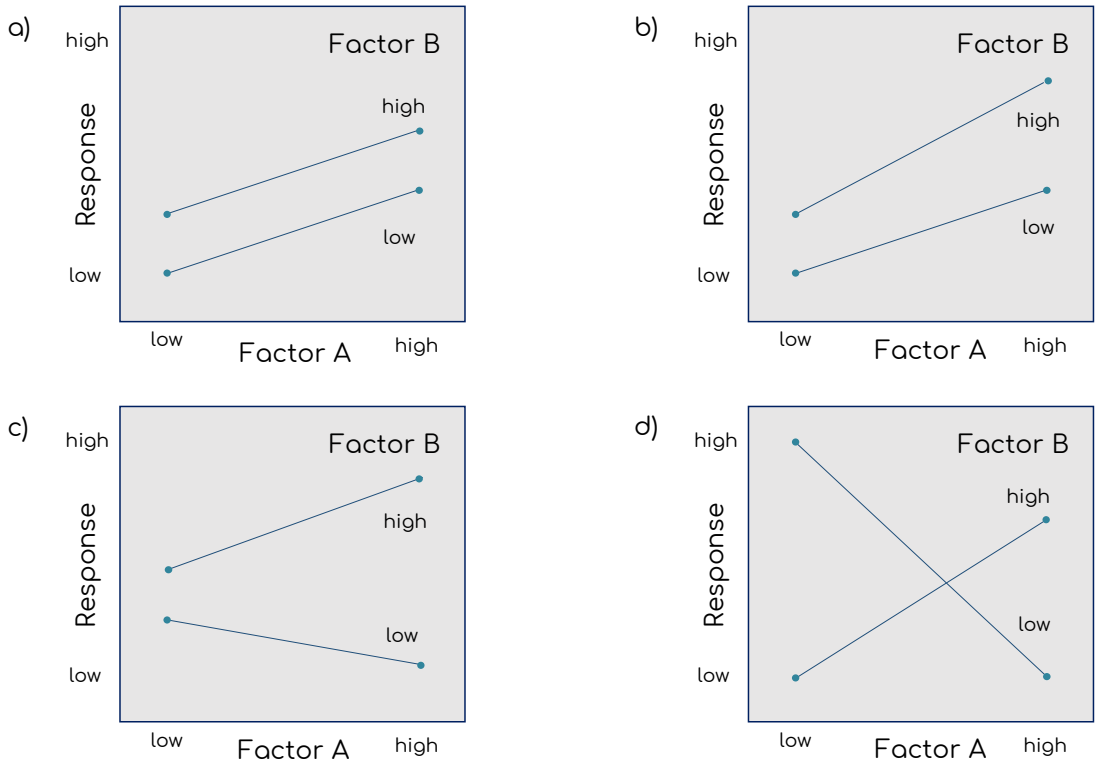


Figure 2.3: Different types of interaction plots.

The presence or absence of an interaction effect is advised by the parallelism degree of the two blue lines represented in figure 2.3: parallel lines suggest that no interaction occurs, and the more the non-parallel the lines are, the stronger the interaction is expected to be [16].

For figure 2.3, in the plot a) the lines are parallel and so there is no interaction. For the remaining plots, the non-parallel lines suggest the presence of an interaction: a small one in b), a moderated one in c) and a strong one in d); indeed, for the most important interactions, the lines are expected to cross each other. For the plots where an interaction exists, mainly the bottom ones, it is possible to observe that the value of the response for a given factor depends on the other factor. However, these assumptions should always be confirmed with statistical analysis.

2.4 Design of Experiments

The design of experiments (DoE), also referred to as experimental design, is a planned set of experiments which aim is to obtain the maximum amount of information in the smallest number of experimental runs possible. The general concept is to change all the relevant independent variables, the factors, simultaneously through a

group of trials in order to cover the area of interest [17]. The collected data is then interpreted using mathematical models and statistical methods, allowing to retrieve meaningful and valid conclusions from it.

The DoE is a more effective approach than the old one-factor-at-a-time: it can be used to study interactions effects and it requires fewer experimental trials to obtain the same level of statistical power [18]. In fact, DoE has been called the most cost-effective method for optimization [19].

Two stages are usually performed when analyzing the relevant factors for a given response in experimental designs: screening and response surface methodology (RSM). The screening experiments are performed at the beginning of a DoE analysis. The objective is to test a large number of prospective variables and get information of which ones of those factors are the most important ones, i.e. the ones that are more likely to have a significant impact in the response. After it, with the selected factors, a second type of experiments, the RSM one, is typically used to optimize the performance of the process and/or the composition of the product [10, 20].

2.4.1 Screening Designs

In a screening design, for each one of the variables considered in the DoE, a maximum and a minimum values are attributed, based on pre-knowledge or constraints. Therefore, in these type of designs, usually only two levels are considered (for continuous variables), the high level, coded as +1, and the low level, coded as -1 [16]. For RSM designs, more than two levels are typically used [21]. An experimental design can be seen as a table where each row corresponds to one experimental trial (commonly designed as a sample in machine learning) and each column corresponds to one experimental factor (feature in machine learning). The values in the columns indicate the levels (high or low) of the factors.

Often, center points are also included. These points correspond to trials where all the factors are at their mean level (coded as 0) and they can be used to check the linearity (detect curvatures) of the process in a simple way [19]. It is also recommended to perform some replications, i.e. repeat some trials in order to get higher statistical power [14, 19]; however, this is not possible many times due to resources limitation.

When addressing experimental designs, two types can be contemplated: the full factorial design (FFD) and the reduced ones. In a FFD, all the possible combinations of factor levels are performed. The table 2.1 represents an example of a full factorial

table with 3 factors and 2 center points. For a two-level (high and low) FFD, if there are k factors, the total number of experiments is 2^k; so, the number of necessary trials grows exponentially with the increase of the number of variables. Thus, this may not be feasible when we have a significant number of variables because it would consume a lot of resources, and it has been considered that the cost of screening runs must be at most 25% of the DoE's total budget [19].

In fact, in these situations, a reduced design may be a more appropriate option. Here, a lower number of experimental trials are performed. The trials are chosen to maximize the variation (amount of different information) with a more economic design. These designs allow then a good compromise between cost and information.

Figure 2.4 shows an example of both full factorial and reduced factorial designs for an experiment with 3 factors. The corresponding FFD table is represented in table 2.1, where the trial number corresponds to the same vertex number in the figure.

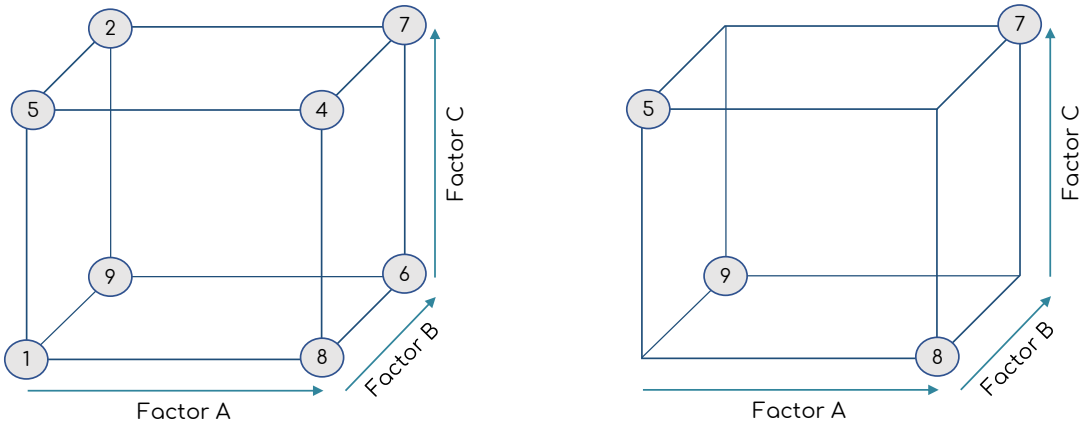


Figure 2.4: Full factorial (left) and fractional factorial (right) designs with 3 factors. Adapted from Elazazy [20].

Table 2.1: Example of a full factorial design with 3 factors and 2 center points.

Run	Factor A	Factor B	Factor C	Pattern
1	-1	-1	-1	—
2	-1	1	1	-++
3	0	0	0	000
4	1	-1	1	+-+
5	-1	-1	1	−+
6	1	1	-1	++-
7	1	1	1	+++
8	1	-1	-1	+–
9	-1	1	-1	-+-
10	0	0	0	000

2.4.1.1 Reduced Factorial Designs

The more traditional reduced designs are the regular and the non-regular orthogonal two-level fractional factorial designs. More recently, a new category, the optimal designs, has been proposed. The orthogonal designs have the desired properties that the estimated effects are totally independent of each other. However, a specific number of trials is required, specifically a power of 2 for regular and a multiple of 4 for non-regular orthogonal designs. On the other side, the optimal design can be used for any number of runs [22].

Besides, for traditional designs, a high resolution may be required. The resolution refers to the number of terms which may be estimated in the regression equation without aliasing: the higher the resolution, the more terms can be evaluated [21]. If the resolution is not high enough, some effects will be totally confounded (or aliased) with each other. To be completelly confounded basically means that the correlation between two effects is equal to 1 and so it is not possible to distinguish those effects.

For instance, if we want to estimate the main effects and the two-factor interactions, neglecting higher-order interactions (which is the most common procedure, as explained in the next section about effect principles), a resolution of level V or higher is necessary in order to those effects not be confounded [21,23]. For a process with a significant number of factors, to consider that resolution type means that a lot of experimental trials must be performed, which may not be possible due to budget constraints as stated before: for example, 10 and 15 factors would require a minimum of 128 and 256 runs, respectively [21].

The optimal experimental designs are non-orthogonal and computer-generated designs. Once they are non-orthogonal, the estimates of the effects will be partially

correlated, but that correlation is small: the effects will never be totally confounded even when second-order interactions are considered and so it is possible to estimate these effects if they are present; the detection of important interactions compensates the variance inflation that is introduced by the non-orthogonality [22].

The optimal designs are, regardless of the number of runs, the ones that maximize a given function of the information matrix, accordingly to some criterion, which is typically the D-criterion or the I-criterion [22].

2.5 Effects Principles: Hierarchy, Sparsity and Heredity

When studying factorial experiments, three fundamental principles are commonly addressed: hierarchy, sparsity and heredity. They are very relevant for a successful screening and they were well described and discussed by Wu and Hamada [14].

The first principle, the effects hierarchy, suggests that lower-order effects are more likely to be important than higher-order ones, i.e. the main effects tend to be the largest ones on average, then the two-factor interactions, then three-factor interactions, and so on.

The second principle, the effect sparsity, states that only a small number of effects are expected to be important, and it is sometimes referred to as the Pareto principle in experimental design, based on the separation of *the vital few from the trivial many* concept developed by the already mentioned Dr. Joseph Juran.

The last principle, the effects heredity, is related to the relationship between an interaction and its parent factors. More specifically, it indicates that an interaction can be active only if one (weak heredity) or both (strong heredity) main factors are also active. For example, assuming a given interaction AB: according to the weak heredity principle, at least one of its parents (factor A or factor B) should be active in order the interaction to be active as well; following the strong heredity principle, both factors A and B must be active. This principle has the advantage that a model is more easily interpretable when the main factors are considered (interactions are more difficult to interpret, and higher the order harder is that task).

The hierarchy and heredity principles are visually represented in figure 2.5, where the effects of an arbitrary example are shown. In the same figure, the box size of each effect is equivalent to its importance. It is possible to observe that more significant main effects lead to more meaningful interactions (heredity) and that, on

average, the lower the order, the bigger the impact of the effects (hierarchy).

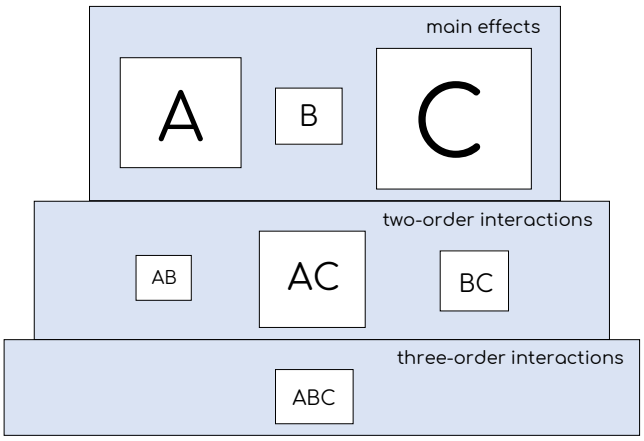


Figure 2.5: Visual representation of the hierarchy and heredity principles. Adapted from Li et al. [15].

Some studies, which took into account a significant amount of different datasets from several engineering fields, were performed in order to analyze empirically these principles, such as Li et al. [15] (113 datasets) and Bergquist et al. [24] (22 datasets) and they had shown an empirical evidence for the support of the aformentioned ideas. Even it is not guaranteed that these principles are true for any particular situation, their assumption seems to be reasonable.

These principles can be especially important in screening experiments. Once they are used to analyze the impact of a large number of factors with a small number of runs, they are based on the sparsity and hierarchy principles. Besides, it is considered that the three principles may improve the power of the analysis of non-replicated experiments (which are common in the screening phase) and that effects heredity can be advantageous when we investigate data from complex aliasing patterns, allowing to identify possible important interactions without be necessary to resort to the high-resolution designs [24].

3

State of the Art

Several methodologies have been proposed to identify active effects in screening experiments, from some decades ago. In this chapter, an extensive review of such techniques is contemplated, from the more standard to the recent ones. A special case of screening experiments is the supersaturated design, which is an experimental design where the number of runs is smaller than the number of main effects. This type of design was not referenced in the previous chapter because it is not the situation of the problem being studied: there are more trials than factors. However, a large amount of studies about the detection of important effects in screening stages is related to supersaturated designs, and so those researches are also reviewed because they present approaches whose principles may also be applied to the remaining screening designs. In the end, some applications related to the pharmaceutical industry will also be presented. In order to have some chronological insights, the years of the proposed methodologies are also indicated.

3.1 Analysis of Variance - ANOVA

The standard method for the detection of significant effects in factorial experiments is the analysis of variance, typically referred to as ANOVA. ANOVA is designed to evaluate a given continuous response based on one or more categorical predictor variables. Therefore, it is the primordial technique to experimental studies because these usually compare levels of treatment (categorical independent variables).

In fact, ANOVA is being extensively used to analyze factorial experiments in several industries. Some case studies include synthesis of lactulose from whey permeate [25], improvement of wire bonding quality [26], development of a nanomanufacturing system for recycling of welding rod residuals [27] or reduction of energy consumption [28]. In the last example, a least square model was used, which is equivalent to an ANOVA approach when two-level independent variables are used. If the predictors

have more than two categories, least square estimation and ANOVA can still be equivalent, if dummy variables are created.

ANOVA can be used to estimate the main effects and all the interaction effects, but replication is required for that. For unreplicated experiments, there are no degrees of freedom to estimate the error, which implies that there is no residual sum of squares and thus it is not possible to compute the ANOVA F-test [14]. In these cases, two procedures can be followed: 1) to consider only some interactions; 2) to remove unimportant factors.

In order to overcome this inconvenient, several methods were proposed to identify active effects in designs without replication. Hamada and Balakrishnan [29] made a very interesting review and comparison of several studies performed before 1998, more specifically the studies carried by Daniel (1959) [30], Holms and Berrettoni (1969) [31], Zahn (1975) [32], Seheult and Tukey (1982) [33], Box and Meyer (1986) [34], Johnson and Tukey (1987) [35], Voss (1988) [36], Benski (1989) [37], Lenth (1989) [38], Bissell (1989) [39], Berk and Picard (1991) [40], Bissell (1992) [41], Le and Zamar (1992) [42], Juan and Pena (1992) [43], Loh (1992) [44], Dong (1993) [45], Schneider, Kasperksi and Weissfeld (1993) [46], and Venter and Steel (1996) [47].

From all those studies, only the ones of Daniel, Box and Meyer, and Lenth will be further discussed because they are the ones which receive more attention in the literature, the remaining ones are barely referenced.

3.2 Normal and Half-Normal Probability Plots

One of the most commonly used methods to aid ANOVA in the detection of important effects is a graphical technique, the probability plot of effects, proposed by Daniel (1959) [30]. This approach consists of plotting the factor estimates (that result from a least squares estimation) on a normal probability plot (NPP) or half-normal probability plot (HNPP): the order values of the estimates are plotted against their corresponding coordinates on the normal or the half-normal probability scales, respectively [14]. The basic idea is that the inactive effects fall along a straight line while the significant ones fall off the same line.

The big difference between both is that in the HNPP approach, the values represented are the absolute effects. This is considered as advantageous comparing with NPP, because the active effects appear in the upper right corner, overcoming the visually misleading that sometimes the normal plots causes [14].

Figure 3.1 shows an example of the analysis of important effects through the observation of the half-normal probability plot. In the example, 3 factors (A, B and C) are considered and all the main effects and interactions are analyzed on the HNPP. It is possible to observe that 3 effects (B, C and BC) fall apart of the line and thus they are suggested to be active effects.

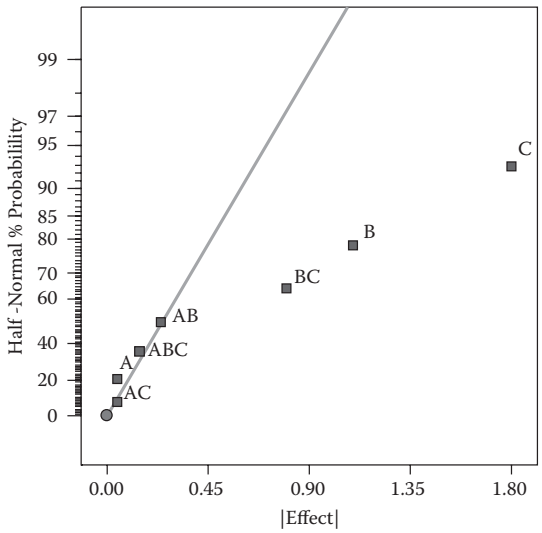


Figure 3.1: Example of half-normal probability plot analysis. Adapted from Anderson et al. [48].

The limitation of this method is that it only allows to visually judge the importance of the effects, no statistical values are addressed. So, typically, once the important effects are visually suggested, the remaining ones are removed and a statistical analysis is performed. For example, for the unreplicated problem, the removal of the effects that are no detected through the HNPP analysis turns the problem into a replicated one and therefore a simple ANOVA test is used to obtain the significance values of those effects.

Once they rely on the analysis of plots, these types of approaches may depend heavily on the skills of the people who are interpreting it.

The case study of shrinkage porosity in permanent mold casting is one example of the application of this method in the industry [49].

3.3 Lenth's Method

The method proposed by Lenth (1989) [38] is said to be simple, without too many assumptions, and to perform generally well [14, 15]. It is based on the computation

of a pseudo standard error (PSE) given by

$$PSE = 1.5 \times \underset{|\beta_i|<2.50 \cdot s_0}{median} |\beta_i| \quad , \quad s_0 = 1.5 \times median\,|\beta_i| \tag{3.1}$$

where β_i corresponds to the estimated effects and s_0 is a standard error parameter. For each effect, a t-like statistic is obtained with

$$t_{PSE,i} = \frac{\beta_i}{PSE} \tag{3.2}$$

and the effect β_i is significant if the $|t_{PSE,i}|$ value is bigger than a critical value [14].

3.4 Bayesian Models

In bayesian methodologies, the probability of a given effect to be active, referred to as posterior probability, depends on the prior probability, which is computed through pre-knowledge.

Box and Meyer (1986) [34] were the first ones to propose a bayesian approach to analyze experimental designs, more specifically fractional factorial ones. In their work, two different parameters were used: α, which is the probability of an effect to be declared as active, and k, which is computed as the ratio of the mean squared significant effets over the mean squared inert effects. The authors computed these values through the analysis of 10 examples of fractional factorial examples.

The α corresponds to the prior probability. In the work of Box and Meyer, the α was set to 0.2 for all the effects, but a different value can be attributed to each effect, based on prior information. A prior probability of 0.2 means that the probability of an effect to be active is 20%, which is in according to the effect sparsity principle. In fact, this concept was first presented in this work.

The remaining effect principles, the hierarchy and heredity ones, were not incorporated in the algorithm of Box and Meyer, but they were introduced in bayesian procedures by Chipman (1996) [50]. In this work, the posterior probabilities were obtained using the stochastic search variable selection (SSVS) algorithm of George and McCulloch [51], which was modified to use hierarchical priors. The hierarchical priors account for the dependence between related predictors, i.e. the prior probability of the two-way (and higher-order) interactions depends on if its parents are active or not. More specifically, the probability that an interaction AB is active

$P(\delta_{AB} = 1 \mid \delta_A,\delta_B)$ may take four different values:

$$P(\delta_{AB} = 1 \mid \delta_A, \delta_B) = \begin{cases} p_{00} & \text{if} \quad \{\delta_A, \delta_B\} = \{0,0\} \\ p_{01} & \text{if} \quad \{\delta_A, \delta_B\} = \{0,1\} \\ p_{10} & \text{if} \quad \{\delta_A, \delta_B\} = \{1,0\} \\ p_{11} & \text{if} \quad \{\delta_A, \delta_B\} = \{1,1\} \end{cases} \tag{3.3}$$

where δ_A and δ_B are equal to 1 if they are active or equal to 0 if they are inert. The conditional probabilities (p_{00}, p_{01}, p_{10}, p_{11}) are thus (0, 0, 0, p) for the strong heredity, and (0, p1, p1, p2) for the weak heredity principles. The authors also suggested the relaxed versions of both strong and weak heredity which attributes a small probability instead of a zero one.

In this paper, the prior probabilty for a main effects was set to 0.5 (bigger than in the study of Box and Meyer), and the interaction effect conditional probabilities were set to: (0.00, 0.00, 0.00, 0.50) for the strong heredity; (0.00, 0.01, 0.01, 0.50) for the relaxed strong heredity; (0.00, 0.25, 0.25, 0.50) for the weak heredity; (0.01, 0.25, 0.25, 0.50) for the relaxed weak heredity.

Chipman, Hamada and Wu (1997) [52] followed the relaxed weak heredity approach of Chipman but used a prior probability of 0.25 for main effects and conditional probabilities of (0.01, 0.10, 0.10, 0.25) for the interactions. They evaluated their proposed procedure using fractional factorial and supersaturated designs, between others.

Beattie, Fong and Lin (2002) [53] proposed a two-stage Bayesian model selection strategy where they first applied the SVSS approach and then employed the intrinsic Bayes factor method of Berger and Pericchi [54], to further select active effects from the ones selected in the first stage.

Bergquist, Vanhatalo and Nordenvaad (2011) [24] proposed a three-steps method where the sparsity, hierarchy, and heredity principles are successively added, by adjusting the prior probabilities. The goal was to compare the posterior probabilities for each effect in each one of those three rounds. The hierarchy and heredity principles were considered following the ideas of Chipman, but they were incorporated into the less parameterized method of Box and Meyer; in fact, in the algorithms of Chipman, several parameters need to be specified.

More specifically: in the first round, a prior of 0.2 was attributed to all effects (sparsity); in the second round, it was considered a prior of 0.5 for main effects,

0.1 for two-factor interactions and 0.01 for higher-order interactions (sparsity and hierarchy); in the third round, it was given a prior of 0.5 for main effects, 0.3 for two-factor interactions with strong heredity, 0.02 for two-factor interactions with weak or no heredity and 0.01 for higher-order interactions (sparsity, hierarchy and heredity).

The authors studied their approach in full and fractional factorial designs with a resolution of at least IV. In each round, a factor was considered active if the posterior probability was higher than 0.5.

The Bayesian approaches present a very interesting way to incorporate the effect principles, but several model parameter must be specified before the analysis. Furthermore, the use of Bayesian methods by practitioners is limited because usually they are not available in the computational packages [24].

3.5 All-Subsets and Stepwise Approaches

If the number of effects (both main effects and interactions) is larger than the number of trials, no least squares estimation can be computed and so a variable selection procedure is required. All-subset and stepwise-version selection techniques are commonly used procedures.

Hamada and Wu (1992) [55] proposed a 3-steps procedure which takes into account the effects heredity principle to analyze fractional factorial designs. First, they applied a standard technique on the main effects, such as ANOVA or half-normal plot, to identify active first-order effects. Second, they expanded the model including all the two-factor interactions whose at least one term is a significant main effect, and applied a forward stepwise regression analysis on that larger model. Third, they created a new model composed by the significant effects identified in the previous step and all the remaining main effects, and they did again the forward stepwise selection. They repeated steps 2 and 3 until the selected model stops changing.

Lin (1993) [56] and Westfall, Young and Lin (1998) [57] used forward stepwise regression procedures to select important effects in supersaturated designs. Abraham, Chipman and Vijayan (1999) [58] proposed the all-subsets regression as a better approach (in comparison to the forward one) to select active effects in supersaturated experiments.

In the all-subsets selection, all the combinations of variables are tested, and thus it is a very time-consuming method, which may be infeasible when a large number of

independent variables are being considered. So, in those cases, forward stepwise selection or a combination of forward and backward stepwise regression, known simply as the stepwise regression, are typically preferable. Pinault (1988) [59] used stepwise regression to analyze orthogonal designs and Lu and Wu (2004) [60] applied a modified stepwise selection combined with staged dimensionality reduction to study supersaturated designs.

3.6 Penalized Least Squares Methods

A more recent and alternative approach is the use of penalized least squares methods, which shrink the estimates to zero and so they can be used as variable selection procedures. The most well-known one is the least absolute shrinkage and selection operator, or LASSO.

Xing, Wan and Zhu (2013) [61] proposed a LASSO procedure for supersaturated problems where the value of the penalty was chosen based on a self-voting using ordinary least squares estimation. Mohammed (2018) [62] proposed a robust LASSO (Huber loss function with LASSO) to analyze factorial experiments whose response follows an epsilon skew Laplace distribution.

Yuan, Joseph and Lin (2007) [63] proposed a modified least angle regression (LARS) [64], which is very close to the LASSO. The method was modified by the authors to incorporate the heredity principle, both in the weak and strong versions; for the weak procedure, the main effect chosen to enter the model along with the interaction was the one with the highest predictive score. The methodology was used in a supersaturated design example and the authors showed that the enforce of effects heredity leads to a better performance and it may decrease the ambiguity of the aliased effects when compared to the ordinary LARS. The authors also highlighted that even the LARS has a close connection with the LASSO, it is not clear how their approach can be adapted to the LASSO algorithm.

Choi, Li and Zhu (2010) [65] proposed a LASSO variation that automatically enforces the heredity constraint, where an interaction term only is considered if the corresponding main terms are already included in the model (strong heredity): this implies that if a given main effect is shrunk to zero, the corresponding interactions will be set to zero as well. The authors applied the algorithm to a design of experiments problem in a simulation study and it showed better performance than the standard LASSO.

Noguchi, Ojima and Yasui (2012) [66] proposed a procedure similar to the previous

one, but they suggested the use of weak heredity instead of the strong one: for a given interaction if at least one of corresponding main effects has an estimate different of zero, then the interaction should have as well. The authors analyzed the performance of their algorithm using data from a fractional factorial design and a supersaturated design and they showed to obtain better results than LASSO in its ordinary and strong heredity versions, and than the more standard forward selection. However, the authors emphasized that both weak and strong versions of the heredity principle should be considered because there is no *a-priori* reason to choose one of them.

Jang and Cook (2017) [67] proposed the use of the LASSO plot as a graphical tool for detection and ranking of the important effects of an unreplicated factorial experiment. The authors compared it with the standard half-normal plots and the LASSO approach showed to be more robust to different variance-covariance structures and to the presence of outliers.

Although the LASSO is the most well-known and used penalized least squares method for variable selection, other penalty types can be considered, such as the smoothly clipped absolute deviation (SCAD) and the Dantzig selector.

Li and Lin (2012) [68] extended the concept of the nonconcave penalized likelihood variable selection, proposed by Fan and Li [69], and they suggested to screen active effects in supersaturated designs using a SCAD penalty. They compared it then with the aforementioned bayesian approaches of Chipman et al. and Beattie et al., and the SCAD procedure showed to perform better; however, only one dataset example was used to establish the comparison. The same authors presented then, in another article [70], a two-stages procedure where they first applied a stepwise variable selection to the full model and next applied the SCAD method to the selected variables.

Phoa, Pan and Xu (2009) [71] suggested the use of Dantzig selector, proposed by Candes and Tao [72], to search for active effects in supersaturated designs. They recommended identifying the important effects through the analysis of the profile plots created varying the regularization parameter. They also proposed an automatic variable selection procedure by choosing the tuning parameter based on a model selection criterion: a modified version of Akaike information criterion for supersaturated design.

3.7 Some Other Procedures

Drosou and Koukouvinos (2019) [73] proposed a support vector machine with recursive feature elimination algorithm adapted to analyze supersaturated designs with continuous response. The magnitude of the weight of the variables is considered to iteratively remove non-important factors, i.e. the variables with a small predictive effect.

Zhang, Zhang and Liu (2009) [74] proposed a partial least squares selection method for detect active effects in supersaturated designs, based on the variable importance in projection of each factor.

Wolters and Bingham (2011) [75] proposed a simulated annealing search algorithm coupled with the corresponding visualization methods to analyze non-regular screening experiments. Their method is intentionally non-convergent in order to generate a large set of good models, instead of a single best solution; however, they also suggested the use of an entropy criterion for an automated selection, i.e. to choose the single best model. The procedure is based on sparsity principle and it also respects the heredity one, dropping variables if they violate the weak version of this principle.

3.8 Information Criteria for Binary Responses

Although some of the aforementioned studies state that generalized linear models can be used in their work, which means not-normally distributed responses can be considered, none of them, for the best of our knowledge, use examples with Bernoulli distributed responses, i.e. binary ones.

Balakrishnan, Koukouvinos and Parpoula (2011) [76] proposed a method for searching active effects in supersaturated designs for a binary response, using the symmetrical uncertainty (SU) measure combined with an information gain measure. They first performed an information theoretical approach based on entropy (Shannon, Rényi, Tsallis and Havrda–Charvát entropies were tested) and selected the half variables with the highest information gain values. Then, they computed the SU between each variable and the response and selected the ones whose SU value was at least as large as the SU median of all values. The features selected in both procedures were then considered as significant ones. The proposed algorithm was studied using several supersaturated examples and only main effects were considered.

Drosou, Koukouvinos and Lappa (2017) [77] proposed a two-step approach to detect active effects in two-level supersaturated designs for a response with Bernoulli dis-

tribution. In the first phase, they computed the mutual information (MI) between each independent variable and the response and they retained the factors whose MI value was greater than a given threshold; based on simulations, the authors found the geometric mean of the MI values as the best threshold. In the second phase, the so-called Tabular Cusum was determined: for the features selected in the previous step, the maximum likelihood estimation and the corresponding coefficients β_I of each factor were computed and taking those β_I's as input, the statistics C_I^+ and C_I^- were calculated; if any of those two statistics exceeded a given decision interval then the I-factor was considered as significant.

The authors applied simulations for several models and for several supersaturated designs. However, only main effects were analyzed. They also compared their algorithm with other three variables selection methods: LASSO, conditional mutual information maximization (CMIM) [78], and the minimal redundancy maximal relevance (mRMR) [79]. The CMIM technique is based on the MI between the factors and the output class, but conditional to the features already picked. The mRMR algorithm looks for the maximization of the MI between the features and the response and as well the minimization of the MI between the selected features. The performance of the methods was analyzed through the geometric mean values between type I and type II errors and the algorithm proposed by Drosou et al. outperformed the remaining methods.

3.9 Some Pharmaceutical Case Studies

Rege, Gawel and Kou (2002) [80] used statistical experimental design to identify the critical input variables that affect the content uniformity and loading of active agent coated on tablets in an Accela-Cota machine. A fractional factorial screening design with 16 runs (3 replicates each combination) was conducted to study the influence of 6 different independent variables in 2 continuous responses. The effects of the input parameters were analyzed by a linear model and the ones with a p-value lower than 0.05 were considered as significant.

Zahel et al. (2017) [81] presented a workflow for the criticality assessment in a biopharmaceutical process. One stage of that workflow is the identification of the design of experiments' factors that have an impact on the critical quality attributes of the process. This procedure is performed using stepwise regression, where variables with a p-value lower than 0.05 enter in the model and the ones with a p-value larger than 0.1 are removed from the model.

3.10 Final Remarks

As stated by Choi et al. [65], the literature on experimental design has a lot of research about the construction of efficient designs, but do not invest the same attention about methodologies to analyze such designs, and the more traditional analysis techniques still dominate. In fact, even several methodologies have been reviewed in this chapter, they are mostly applied in simulations or benchmark examples. When we look for industry case studies, the standard methods, such as the ANOVA and least squares estimations or at most the stepwise procedures, are the ones which are still applied.

It is possible to observe that the same happens in the pharmaceutical industry. The adoption of quality by design (QbD) principles provided a way to develop new data analysis strategies. However, the lack of works in the literature about the screening phase suggests that few studies have been performed in this area. On the other side, the further steps of pharma QbD (optimization after detection of important effects) seems to have that focus, being easy to find application of several new methodologies, including some more complex ones such as artificial neural networks.

Besides, it was verified that a lot of pharmaceutical case studies state that they selected the critical process parameters and critical material attributes based on a given software analysis but they do not specify which methods were used, which gives the idea that the experimenters relied on the software and not on the techniques itself.

On the one hand, the identification of important effects in screening experiments must result from the analysis of very interpretable models, which may explain why standard approaches prevail. One the other hand, the detection of active effects is often not so hard to complete. In these cases, the traditional methods are suitable because they provide an easy way to perform the objectives. However, in more complex problems, such the one being studied in this document, a different analysis may be required.

Relatively to effects heredity enforcement, it is possible to conclude that some approaches were already proposed to deal with it, but they considered it in different ways. While some methods first select the interaction and then along with it the corresponding main effects are also included in the model, other procedures only select the interactions if the corresponding main effects are already in the model; the majority of the examples are related to the last group. So, some algorithm looks first for the interactions while other ones look first for the main effects, which present

two different ways to incorporate this principle in the models. In this document, the first methodology will be referred to as back-enforcement of heredity and the second one will be referred to as fore-enforcement of heredity.

4

Problem Definition and Dataset

The available data is property of Bluepharma - Indústria Farmacêutica, and it is related to a complex generic drug product being currently developed by the company. The manufacturing process of the drug product is divided into several stages (unit operations), namely 5, and in each one of them some variables, material attributes or process parameters, can be manipulated. The material attributes will be considered as formulation parameters in this document from now on. The response variables, i.e. the critical quality attributes (CQAs), are recorded at the end of the final stage.

Data comprises 32 experimental trials, from which 24 were design of experiment (DoE) runs and the remaining 8 were validation runs. For the identification of the active effects, only the DoE trials were considered once they correspond to a planned screening experimental set which was conceived to retrieve the maximum information possible. The validation trials contained information about very specific combinations whose some values were outside of the ranges used in the DoE, and so analyzing them together with the original runs would lead to misleading results. These extra trials were used for further analysis of the problem.

4.1 Variables of the Process

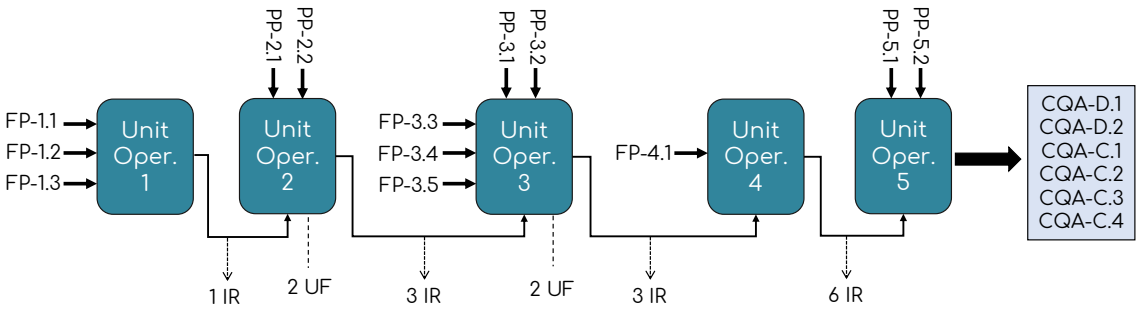


Figure 4.1: Manufacturing process of the complex generic drug product.

Figure 4.1 shows a process chart representing the unit operations of the entire manufacturing process of the product being developed.

The input variables are given by a two letters and two numbers code. The letters give information about the factor type, namely FP if it is a formulation variable or PP if it is a process parameter. The first number (1 to 5) gives information about which of the 5 phases of the process each factor is related with. So, for example, the factor FP-1.1 is a formulation parameter that is manipulable in the first stage of the process. The last character is just a number to create an unique reference for each one of the parameters from the same stage. The factor FP-1.1 is a categorical (binary) one while the remaining ones are all continuous.

The output variables of the process are the product CQAs. There are 2 discrete responses (more specifically binary ones), represented by CQA-D, and 4 continuous responses, represented by CQA-C. The table 4.1 shows some important information about each response. It presents the target for each CQA and the acceptable interval for the continuous responses, based on the reference product.

Table 4.1: Information about the critical quality attributes.

Response	**Type**	**Target**	**Acceptable interval**
CQA-D.1	Binary	Positive class	-
CQA-D.2	Binary	Positive class	-
CQA-C.1	Continuous	Match interval	[24 - 31]
CQA-C.2	Continuous	Minimize	[0 - 0.6]
CQA-C.3	Continuous	Minimize	[0 - 1.6]
CQA-C.4	Continuous	Maximize	[95 - 105] %

In addition to the factors and the CQAs, some other variables are recorded in each unit operation, uncontrolled factors (UF in the process chart) or intermediary responses (IR in the process chart). The integer before UF or IR in figure 4.1 represents the number of uncontrolled factors or intermediary responses, respectively, in that stage.

The uncontrolled factors correspond to independent variables which values were not manipulated during the experiments, i.e. they were not considered in the experimental design, but even so they were measured. A non-controlled factor may be either a variable that can not be controlled at all or a variable that is not expected to impact the responses.

The intermediary responses are dependent variables measured between each unit operation. Their values are not considered critical to ensure that the final product has the desired quality but they can be related to the final CQAs, and therefore they can be used to predict or at least have an idea about the final outcomes in early stages of the process, i.e. to be used for inline process control.

4.2 Preliminary Considerations About the Dataset

The 24 screening trials were performed based on an I-optimal design of experiments. As a reference, in order to complete a full factorial experiment, 8192 runs would be necessary, and for an orthogonal fractional factorial experiment with resolution V, 256 would be required [21], which is extremely costly and impossible to perform in practice.

Although screening designs are mainly used for finding large main effects, interactions are common, especially two-factor ones, as discussed before. If one or more interactions are important and are not included in the model, their effect may bias the estimates of the main effects, and so we not only not detect the true effects but also the estimated ones will be influenced by those interactions.

The optimal DoE was generated considering an *a-prior* model containing only the main effects. If interactions were considered when building that design, a very large number of trials would be required, which was not feasible. However, as explained in Background Concepts chapter, when using an optimal design there is no total confounding and then it is possible to detect interaction effects even that some correlation is introduced.

Following the effect hierarchy principle, it is most likely main effects and two-factor interactions to be important. Besides, three-factor and higher-order interactions are rare and 24 trials for 13 factors may not be suitable to detect them. So, we may focus on the low-order effects, say main effects and two-factor interactions, assuming that higher-order interactions are negligible.

The number of two-way interactions is given by

$$\frac{\text{number of factors} \times (\text{number of factors} - 1)}{2} \tag{4.1}$$

which corresponds to 78 interactions when the number of factors is 13. So, in the total, 91 predictors were considered.

The nominal factor is binary and so it was coded as 0 or 1. Primarily, all the factors were standardized and then each interaction was created multiplying the standardized values of the corresponding parent factors.

4.3 Preliminary Analysis of the Responses Data

For the binary CQAs, the classes distribution is: for the response CQA-D.1, there are 11 positive outcomes and 13 negatives ones; for the response CQA-D.2, there are 12 positive outcomes and 12 negative ones. This means that the classes are perfectly balanced for one response and almost balanced for the other one.

For the continuous CQAs, in order to get more insights about the distribution of the data points, it was displayed the boxplots of the responses' values, which are represented in figures 4.2 to 4.5. For each response, the top sub-figure represents the boxplot built from the DoE data and its outliers, and the bottom sub-figure is a zoomed version of the latter one containing the boxplot and all the experimental values that fall inside it. The points corresponding to validation trials are also represented.

The blue points are related to the DoE trials (the ones used to built the boxplot). The validation trials were also plotted to study the intra-variability of the process, i.e. these runs correspond to sets of replicates which allows studying the variance of the outcome when the same combination of factors values is repeated. Additionally, the target interval for each response is also represented to get an idea of how close or far the performed trials are from the desired outcome.

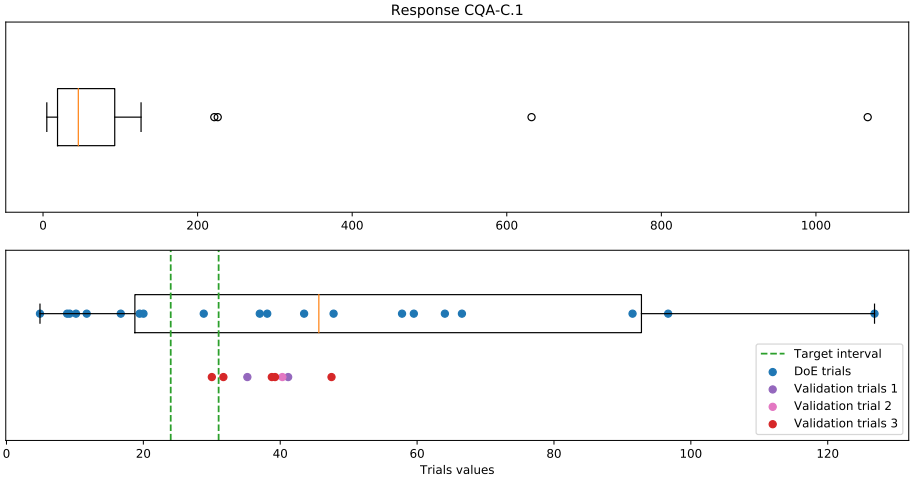


Figure 4.2: Trials' values for the response CQA-C.1.

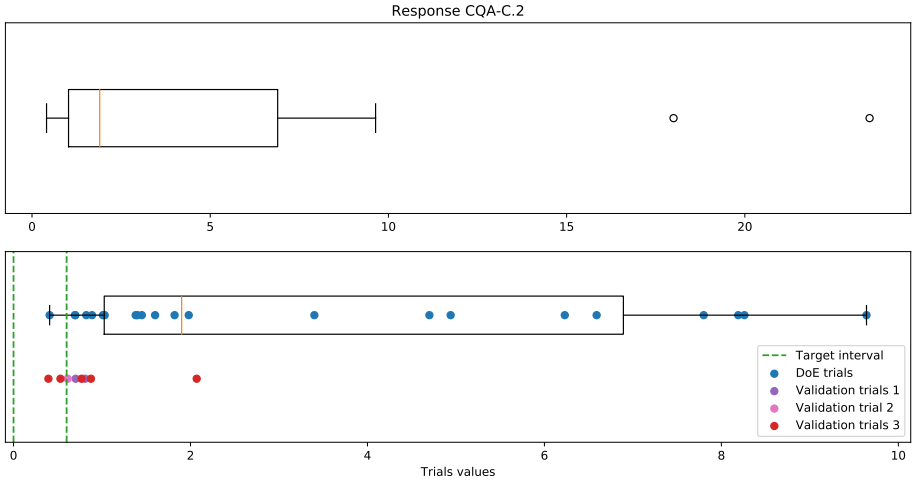


Figure 4.3: Trials' values for the response CQA-C.2.

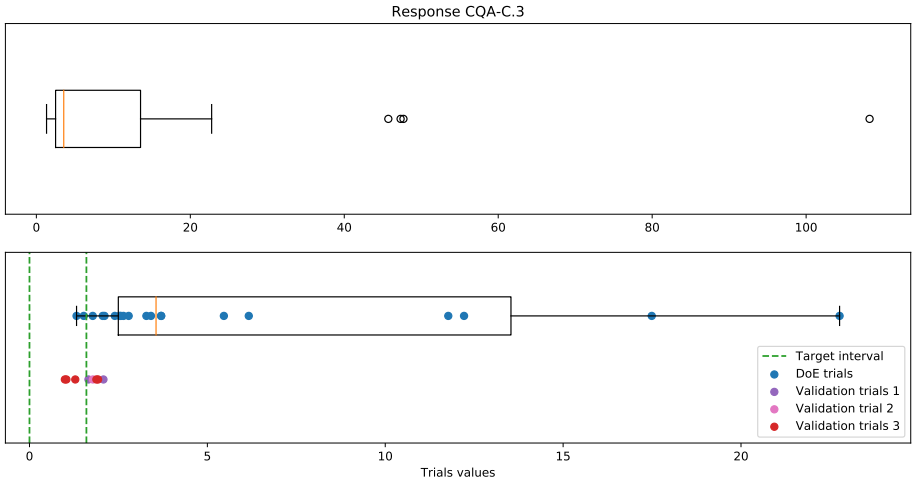


Figure 4.4: Trials' values for the response CQA-C.3.

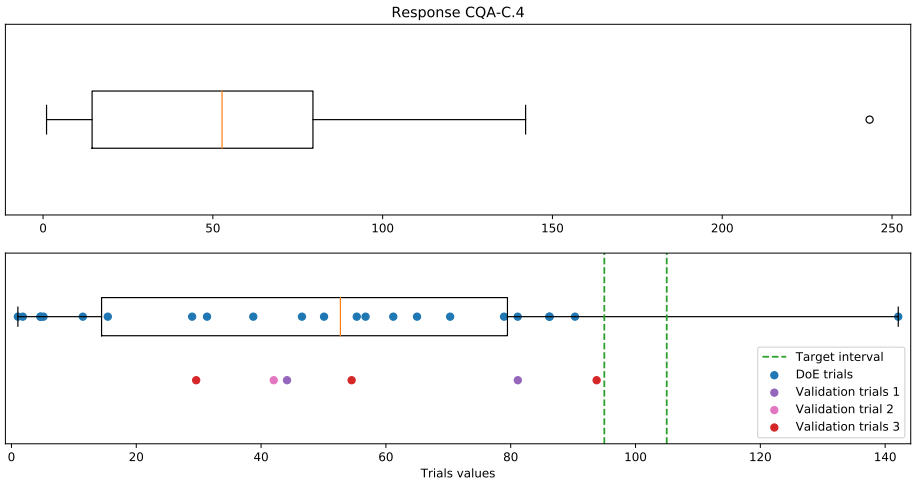


Figure 4.5: Trials' values for the response CQA-C.4.

For the response CQA-C.4, figure 4.5, two of the runs had output values outside of the physically possible range, i.e. higher than 105%. Furthermore, it is possible to observe that there is a very substantial intra-variability for the replicates. In fact, it was confirmed by the experimenters that the quantification method of this response is not the best one and thus the uncertainty associated with the measured values and the consequent models is high. Therefore, the analysis of this response will not be considered in this document.

For the remaining responses, figures 4.2-4.4, it is possible to inspect that the boxplots are the typical ones of populations with a right-skewed distribution, i.e. the right whisker is much longer than the left one and the median value moves towards the left side of the box.

5

Methodologies

The main goal of this work is the identification of active effects, the factors which have an important impact in each one of the responses (critical quality attributes), and therefore correspond to the critical input parameters. Once interactions are often important, both main effects and interaction effects were considered. With the addition of second-order interactions, the number of total independent variables exceeds the number of experimental trials. So, a feature selection procedure is required.

Additionally, a further analysis of the problem was also performed to get some more useful insights and make available techniques that can be used in the future, which includes the optimization of several responses simultaneously, the study of non-controlled factors and the creation of predictive models.

This chapter is divided into three parts. The section 5.1 presents an overview of the followed pipeline and a summarize of the used methods. The section 5.2 gives a further theoretical explanation of those methods, which aids to understand the reasons why each approach was followed and how it is incorporated in the pipeline. The section 5.3 presents the software used in this work.

5.1 Proposed Pipeline

5.1.1 Identification of Active Effects

The screening experiments are performed to identifying important effects rather than for prediction, therefore it is desirable to obtain not only a model with a good fit but also one with a meaningful interpretation. Besides, the number of trials is small, so due to some partial aliasing, interactions may dominate over main effects in data analysis, which is not expected in reality. Thus, the proposed procedure respect the effect principles, which leads to models that are more interpretable and

agree more with the theoretical knowledge.

While the effects hierarchy is imposed by considering only main effects and two-factor interactions, effects heredity must be enforced during the process of model selection. As stated before, a variable selection approach is necessary. Besides, it is preferable a method whose relations between main effects and interactions are easy to model so the enforcement of heredity can be implemented.

Taking these assumptions into consideration, a stepwise regression selection procedure (composed of both backward and forward steps) with back-enforcement of heredity was applied. An exhaustive explanation of how the effects heredity was introduced in stepwise regression in this work will be provided in section 5.2.3. Such procedure was implemented using the "Pruned Forward Selection" method and the "Enforce Effect Heredity" option available in JMP Pro [82]. However, to the best of our knowledge, there are no examples of its application in the literature.

To map the independent variables into the responses, regression-derived methods were then applied. For the binary outcomes, logistic regression (LR) was considered. For the continuous responses, as observed in the previous chapter, they have a right-skewed distribution. Even though ordinary least squares (OLS) - the standard linear regression approach - can be used for non-normal data, for very skewed distributions some of the assumptions, such as normally distributed errors and homogeneity of variance, do not typically hold. So, in these cases, it is often preferable to use generalized linear models (GLM) that can be applied to other data distributions. In fact, the LR is a special case of GLM. So, for continuous responses, GLM was considered. However, the standard OLS was also applied in order to establish some comparisons.

As validation method, to guide the stepwise regression variable selection and to select the best models, it was used the corrected Akaike information criterion (AICc) - lower the AICc, the better the model. However, due to the underlying uncertainties of models selection, all the models with an AICc difference of less than 2 in relation to the lowest AICc value were considered, i.e. the ones with

$$\Delta\text{AICc} = \text{AICc}(model) - \text{AICc}(minimal) \leq 2 \tag{5.1}$$

Burnham and Anderson [83] suggested that there is empirical support for the models with $\Delta\text{AICc} \leq 2$, i.e. there is no substantial evidence that those models are worse than the one with lower AICc value. The authors also state that this interval is

invariant for different scales, so this assumption was used as a rule of thumb.

Figure 5.1 shows the example of the selection path by the stepwise selection, with the AICc value as the validation method. The model with the minimal value of AICc is taken as reference and then all the model with an AICc difference lower than 2 are selected as well.

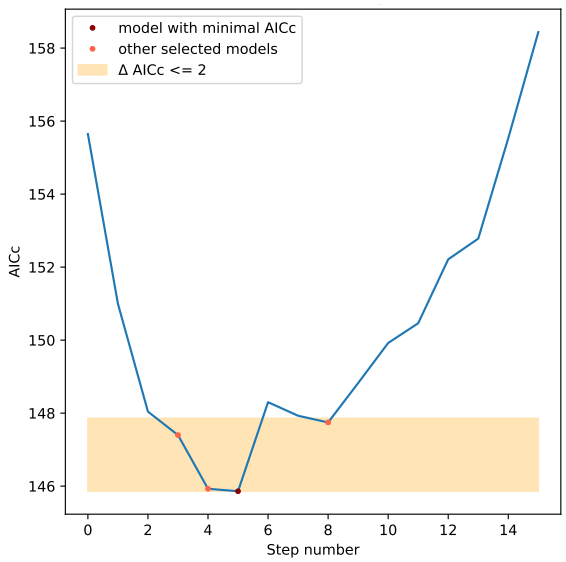


Figure 5.1: Example of the path of stepwise selection of the important effects with AICc validation, for one of the responses.

From the list of selected models, the final chosen model was the one with higher R^2, when OLS was used, or generalized R^2, when LR or GLM was used. This criterion was considered for two main reasons. First, the model with higher R^2 value will be the one with a better fit from the set of models that are likely to be equally good accordingly to the AICc criterion. Second, the R^2 typically increases when it is added a variable with some extra importance; in a standard machine learning problem, the model with lower number of predictors is often preferable, but in this problem the main goal is to detect active effects, so the main concern is to not miss any potential active effect, even it is found to not be important in posterior analysis.

In fact, in screening experiments, type II errors (considering an active factor as an inactive one) are more problematic than type I errors (considering an inactive factor as active one), so it is better to keep a larger number of factors in a first step, even if they result in redundant cost in follow-up experiments [57].

5.1.1.1 Investigation of Distribution of Continuous Responses

Several different distributions can be modeled by the GLM methodology. In order to evaluate which one is more suitable to the responses data, four positive right-skewed distributions were fitted to the data: lognormal, exponential, gamma and Weibull. The fit was performed using maximum likelihood estimation and the goodness of fit was then evaluated through the Kolmogorov-Smirnov statistical test and the AICc value.

It can be already disclosed that the lognormal distribution was confirmed to be the best-fitted one for all continuous reponses. Data with a lognormal distribution or a close-lognormal one should have a normal or a close-normal distribution when a log-transform is applied. Therefore, the $\log_{10}$ of the values of those 3 responses was computed and the normality of the transformed data was evaluated using the Shapiro-Wilk and D'Agostino-Pearson.

Figure 5.2 shows a schematic representation of the workflow used to identify important effects.

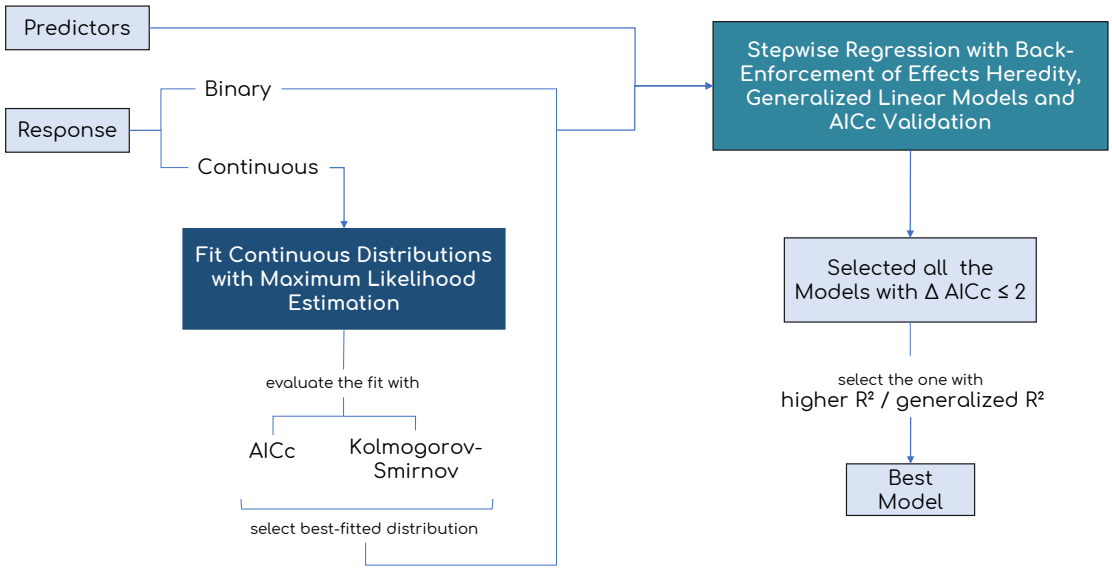


Figure 5.2: Workflow used for identification of active effects.

5.1.2 Recommended Levels for Important Factors

Once the active effects were identified, it was studied the levels which important factors should have to achieve the desired response values, through the analysis of three components: Wald statistical test p-values, parameters estimates and interaction plots.

For the main effects, the first step was to verify the Wald test result which evaluates if a given effect is significant or not in the selected model. If the main effect is significant according to the Wald test, then it recommended level can be analyzed through its coefficient estimate obtained using the generalized linear model. A positive coefficient means that the response increases when the higher value of the factor is used, and it decreases for the lower value. So, in order to maximize the response, if the coefficient is positive then the high level is suggested, if it is negative then the low level is recommended. Otherwise, in order to minimize the response, the opposite happens.

If the main effect is not significant according to the Wald test, it means that the effect only entered in the model because the factor is a parent of one or more selected two-factor interactions. In this case, the corresponding interaction plots must be analyzed.

5.1.3 Multiple Response Optimization

The study of the recommended levels is an individual analysis for each response. However, five different responses are being considered and it is very likely that those suggested levels may be contradictory in some situations, e.g. a given factor can affect different responses in opposite ways (for one of the CQAs a high level may be recommended and for other CQA a low level may be better).

So, in these cases, a multiresponse optimization procedure is often executed. However, it is an approach performed on data from more advanced steps of the development process specially designed for optimization, typically response surface methodology experiments, and not in the screening stage. In this work, a multiobjective optimization is implemented to be used as a *proof of concept* of the proposed procedure.

For that purpose, desirability functions were used: from the regression formulas obtained previously, an individual desirability (a value in the interval [0,1]) is computed for each response and after that an overall desirability was calculated as the geometric mean of the individual desirabilities. The goal is then to maximize the overall desirability and find possible combinations of independent variables' values which ensure that all the CQAs targets are met simultaneously.

5.1.4 Predictive Models

Although the main goal was to detect important effects and then recommend the best values to achieve the target outcome, it may be advantageous to have models that can predict (with the best precision possible) the future outcomes of a given new combination of values. The regression formulas obtained before can be used exactly for that purpose. In fact, those models were obtained using the AICc value as validation criterion, which incorporates a way to control both underfitting and overfitting, and therefore assures they are good generalizations.

However, it is possible to implement other algorithms for prediction than the ones used for variable selection. For the binary responses, besides the logistic regression, two other methods were considered: logistic ridge regression (LRR) and support vector machine (SVM) with a linear kernel. Once the important variables were selected while identifying the important effects and it already includes interactions, non-linear methods are not necessary. In order to evaluate the predictive ability, a 500 times repeated 10-fold-stratified cross-validation was used as validation procedure, because the application of AICc to SVM is not direct. For the LRR and SVM methods, the penalization parameter was varied through grid search in the interval [-6, 6] in the logarithmic scale, i.e. in the interval $[2^{-6}, 2^{6}]$, with unitary exponential steps.

For the continuous responses, at this stage of the process development, instead of the single mean value of the outcome for a new trial, it is more interesting to know the interval where the response is expected to fall, i.e. the prediction interval (PI). However, the PI is traditionally formulated based on the normality assumption of the errors distribution, which as stated before it is not usually valid for the non-normal generalized linear models.

As the distribution of the responses is lognormal, the standard approach is to log-transform the response, compute the PI in the transformed scale and then back-transform it to the original scale, i.e. if a $\log_{10}x$ transformation is applied, then the back-transformation is 10^{x}. The transformation of the response is not as suitable as the use of GLMs, even both approaches are close to each other, but there is no standard way to compute the PI in the GLM approach. A possible method could be to calculate bootstrapping PIs. However, once we can compute the predicted mean using both transformation and GLM procedures, it is possible to have an idea of the differences between the values obtained with each one. Therefore, using that difference as a reference, we can calculate the PI using the transformation method

and have a total understanding of the results.

Furthermore, as we are more interested in the prediction intervals, no other methods as linear ridge regression or support vector regression were considered because they would only optimize the mean value. Besides, considering the limited available information and the intrinsic data intra-variability, those methods are not expected to bring extra useful information at this stage of the process.

5.2 Theoretical Background

5.2.1 Probability Distributions

In tables 5.1 (continuous distributions) and 5.2 (binomial distribution), it is displayed information about the probability distributions contemplated in this work.

Table 5.1: Continuous distributions considered in this work.

Distribution	**Parameters**	**Probability Density Function**	**Population Mean**
Normal	$\mu \in \mathbb{R}$ (location) $\sigma > 0$ (scale)	$f(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	μ
Lognormal	$\mu \in \mathbb{R}$ (location) $\sigma > 0$ (scale)	$f(x \mid \mu,\sigma^2) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{(log(x)-\mu)^2}{2\sigma^2}}$	$e^{\mu+\frac{\sigma^2}{2}}$
Exponential	$\beta > 0$ (scale)	$f(x \mid \beta) = \frac{1}{\beta} e^{-\frac{x}{\beta}}$	β
Gamma	$k > 0$ (shape) $\theta > 0$ (scale)	$f(x \mid k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-\frac{x}{\theta}}$	$k\theta$
Weibull	$k > 0$ (shape) $\lambda > 0$ (scale)	$f(x \mid k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{x-1} e^{-(\frac{x}{\lambda})^k}$	$\lambda\Gamma\left(1+\frac{1}{k}\right)$

Table 5.2: Binomial distribution considered in this work.

Parameters	**Probability Mass Function**	**Population Mean**
$n \in \mathbb{N}$ (number of trials) $p \in [0,1]$ (success probability of each trial)	$f(x \mid n,p) = \binom{n}{k} p^k (1-p)^{n-k}$	np

Some notes related to these tables: x represents the data being evaluated; Γ is the gamma function; for the lognormal distribution, the location and scale parameters of x are related to the logarithm of x and not to x itself; the Bernoulli distribution is a special case of the binomial distribution, for n=1.

5.2.2 Maximum Likelihood Estimation

The maximum likelihood estimation (MLE) is a statistical method for estimating the parameters of a given distribution by maximizing the likelihood function $L(\theta \mid x)$, which is a function that describes the probability of obtaining the observed data (x), for all values of the parameters (θ) within the parameters space (Ω) [84].

The likelihood function is given by

$$L(\theta \mid x) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{5.2}$$

where $f(x_i \mid \theta)$ is the probability density function (PDF) of the sample. For example, observing the normal PDF in table 5.1, it is possible to state that if we are trying to fit the data to the gaussian distribution, two parameters are being considered, mean (μ) and standard deviation (σ), and the parameters space is defined by

$$\Omega = (\mu,\sigma) : -\infty < \mu < \infty \quad \text{and} \quad 0 < \sigma < \infty \tag{5.3}$$

The maximation of the likelihood function is then equivalent to find the values of these parameters that best explain the observed data.

5.2.3 Stepwise Regression with Enforcement of Heredity

The stepwise regression is a technique that can be used to select a group of important variables in a regression procedure. In this feature selection method, the statistical significance related to entering or removing of a given variable is analyzed.

The method is composed of a mixture of two methods: forward selection and backward selection. It starts with a model containing only the intercept. Then, in the first step, the variable with the most significant effect is added to the model [82]. After that, in the following steps, the algorithm considers three different possibilities:

1. Forward selection: from the variables that are not in the model, add the one whose effect is the most significant one (usually based on the score test [85]).
2. Backward selection: from the variables that are in the model, remove the one whose effect is the least significant one (usually based on Wald test [85]).
3. Do both backward and forward selections in a single step.

In order to choose which one of the above actions is performed at each step, the algorithm takes into consideration a given validation method, which was AICc in this work. The best final model is the one that grants the best solution taking that validation method.

When interactions are considered, the effects heredity was imposed using the "Enforce Effect Heredity" option of JMP Pro [82].

In a forward action, if the most significant effect is an interaction and one or both of its parent factors are not included in the model, a compound effect is created, containing the interaction and any other inactive effect necessary to satisfy the heredity principle. If this compound effect is the most significant one, then all the effects of the compound are added to the model in a single step.

This is a very interesting approach because it allows to add an interaction to the model even if both parents are not active effects (it means that the interaction has a very strong effect), but the probabilites of it happening are reduced as its parents must also enter in the model.

In a backward action, if a parent of one or more interactions is the less significant effect, another compound is created, containing the main effects and its interaction(s).

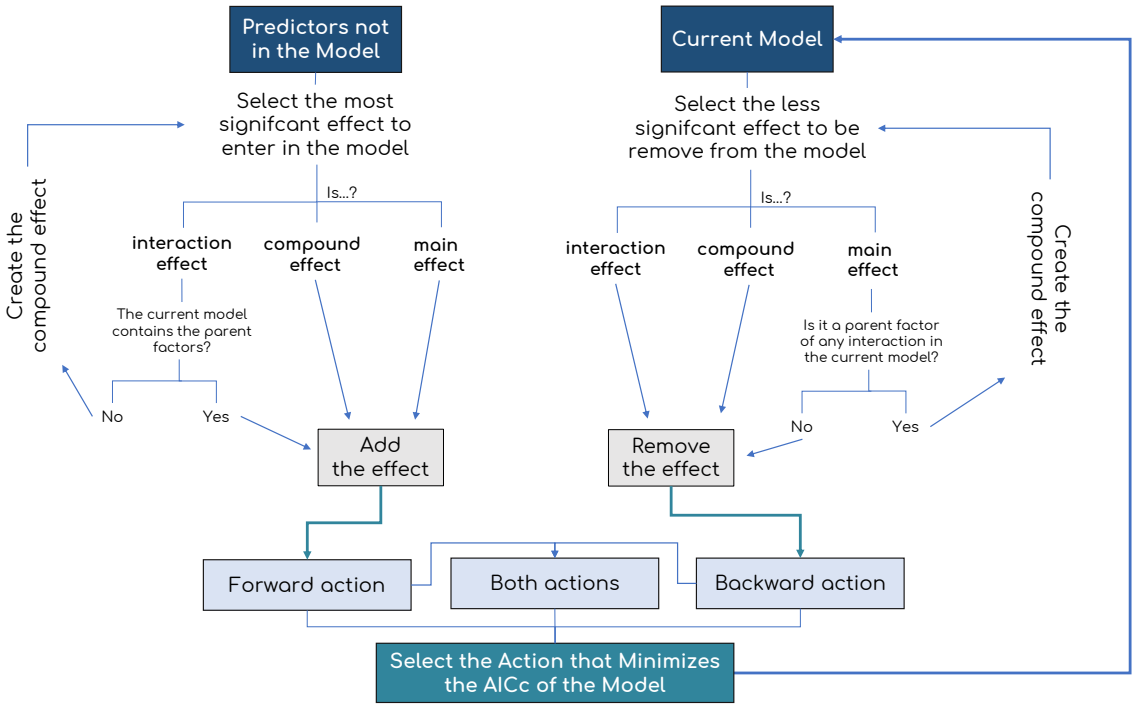


Figure 5.3: Schematic representation of the stepwise procedure with back-enforcement of heredity.

Figure 5.3 shows a schematic representation of the procedure used to select the effects to be removed or to enter in the model at each step.

It is important to highlight that the best of the three possible actions is taken even if the AICc value of the new model is higher than the AICc of the current model, in order to enlarge the search space. The search will stop when the AICc does not decrease after 10 steps, comparing to the minimal value already obtained, as it is possible to observe in figure 5.1.

5.2.4 AICc: corrected Akaike Information Criterion

The Akaike information criterion (AIC) [86], estimates the relative quality of statistical models based on the principle of maximum likelihood (ML) and it is given by

$$\text{AIC} = -2logL(\theta) + 2k \tag{5.4}$$

where $L(\theta)$ is the ML function for the model and k is the number of estimated parameters in the model.

Observing equation 5.4, it is easy to understand that the first term tends to get lower as more parameters are added to the model while the second terms always increase as more parameters are considered. This can be seen as a trade-off where AIC looks simultaneously for both the goodness of fit (GoF) of the model to the data and the simplicity of the model.

Using AIC is then possible to evaluate the fitting of the data (in terms of the balance between underfitting and overfitting) using all points simultaneously, without the need to split the data into groups as it happens with the cross-validation or bootstrapping techniques, for example. In a problem like the one being studied, this characteristic is very suitable because of two main reasons: 1) the sample size is small; 2) for an optimal experimental design, the trials are planned to contribute with the biggest amount of information possible and so it is favorable to consider all the trials together. The split of the data would mean that a given feature-space would be neglected in each group. So, the AIC can be seen as a better validation method for variable selection than the other two techniques.

It is important to notice that, when evaluating the models, the AIC values itself are not important but the relative values, i.e. the comparison of those values between different models. Given a set of models built from the same data, the one with a

lower AIC value is typically considered as the one with best fitting [86].

It was also verified that AIC may not perform well when the number of parameters (predictors) is small in relation to the sample size [87,88]. In order to overcome this issue, a second-order variant of AIC with small-sample adjustment was proposed [87,89], known as corrected Akaike Information Criterion (AICc). In the AICc, the term that penalizes the model complexity is multiplied by a correction factor:

$$\text{AICc} = -2logL(\theta) + 2k \cdot \frac{n}{n-k-1} \tag{5.5}$$

which is equivalent to

$$\text{AICc} = \text{AIC} + \frac{2k(k+1)}{n-k-1} \tag{5.6}$$

where n is the sample size. Due to the small sample size nature of the problem being studied, AICc was used.

In the context of the problem, the AICc value was used as validation method to select the action to take at each step of the stepwise regression approach (either backward, forward or both actions) and to evaluate the GoF of skewed distributions to the data of continuous responses.

5.2.5 Statistical Tests

5.2.5.1 Kolmogorov-Smirnov

The Kolmogorov-Smirnov (K-S) is a non-parametric statistical test that can be used as a GoF measure to determine if whether two distributions differ, or whether a given empirical distribution differs from a hypothesized distribution [90]. In this case, the hypothesized distributions are the right-skewed distribution considered. More specifically, it statistic quantifies the distance between the empirical distribution function (EDF) of the data sample being evaluated and the cumulative distribution function (CDF) of that reference distribution:

$$\text{K-S}_{\text{statistic}} = \sup_x |EDF(x) - CDF(x)| \tag{5.7}$$

where $\sup_x$ is the supremum of the set of distances.

For the K-S test, the null hypothesis is then that the data follow a specified distribution; the alternative hypothesis is that the data do not follow that distribution.

Even K-S is often used to test the normality of the sample, it is a general-purpose test, i.e. it can be used to test any distribution, not being personalized for Gaussian distribution as it happens with the Shapiro-Wilk and D'Agostino-Pearson tests. In fact, it has been noted that the K-S test has low statistical power and it should not be considered to test normality [91].

5.2.5.2 Shapiro-Wilk and D'Agostino-Pearson

The Shapiro-Wilk (S-W) [92] and the D'Agostino-Pearson (D-P) [93, 94] are statistical tests used to evaluate if a given population is normally distributed.

The D-P test, also known as D'Agostino K^2, calculates the kurtosis and skewness of data and combine that statistics to produce an omnibus [1] test which determines how far the asymmetry and shape of the data distribution diverge from the values expected from a gaussian distribution:

$$\text{D-P}_{\text{statistic}} = \sqrt{Z_1(g_1)^2 + Z_2(g_2)^2} \tag{5.8}$$

where g_1 and g_2 are the sample skewness and kurtosis, respectively, and Z_1 and Z_2 are transformations of those measures.

The S-W test is a more commonly used test to perform normality assessment, but has a less interpretable procedure compared to the D-P test. It statistics results from some calculations which envolve rearrange of the data in ascending order and some tabled coefficients:

$$\text{S-W}_{\text{statistic}} = \frac{\left[\sum_{i=1}^{m} a_i (x_{n+1-i} - x_i)\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{5.9}$$

where x_i are the ordered data values and a_i are weights from the the Shapiro-Wilk table based on the value of n, the sample size. If n is even, $m = n/2$, while if n is odd, $m = (n–1)/2$.

For both the Shapiro-Wilk and D'Agostino-Pearson tests, the null hypothesis is that the data population is normally distributed; the alternative hypothesis is that data do not follow a gaussian distribution.

[1] It is considered an omnibus test because it is able to detect deviations from normality due to either skewness or kurtosis.

5.2.5.3 Wald

The Wald test [95, 96] is a hypothesis test which evaluates how far an estimated parameter $\hat{\beta}$ is from a given value β_0 under the null hypothesis $H_0 : \hat{\beta} = \beta_0$:

$$\text{Wald}_{\text{statistic}} = \frac{\hat{\beta} - \beta_0}{SE(\beta)} \tag{5.10}$$

where SE is the standard error. It can be extended multiple parameters to simultaneously compare the MLE estimators with the hypothesized value:

$$\text{Wald}_{\text{statistic}} = (\hat{\beta} - \beta_0)'[cov(\hat{\beta})]^{-1}(\hat{\beta} - \beta_0) \tag{5.11}$$

where $cov(\hat{\beta})$ is the variance-covariance matrix which is equivalent to the inverse of the information matrix.

In regression problems, it can be used to access the significance of the explanatory variables used to build a given model. In this case, it evaluates the estimated parameters in reference to the null hypothesis (H0) that states the parameters have a zero value, $\beta_0 = 0$, and so the predictors are not statistically significant [97]. Therefore, if the H0 is rejected for a given parameter estimate, that predictor will be statistically significant, which in the context of the problem corresponds to be an active effect.

5.2.5.4 Significance Level (Alpha)

For the four aforementioned tests, in order to reject or not the null hypothesis, a reference value of 0.05 was used for the alpha value, i.e. if the p-value of the test is higher than α, we can not reject the H0.

5.2.6 Generalized Linear Model

The generalized linear model (GLM) concept was developed by Nelder and Wedderburn [98] and further discussed by McCullagh and Nelder [99]. This approach is an extension to the standard linear regression.

A GLM is composed of three main components. The first one is the *random component*, which is the distribution probability of the dependent variable (response), i.e. normal, lognormal, Weibull, beta, exponential, binomial, etc. The second one is the

systematic component, which corresponds to a parameterized function of predictors:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} = \beta X^T \tag{5.12}$$

where β_0 is the intercept, which corresponds to the mean value of the outcome when all X=0 and the β_j are the coefficients for the x_j variables, calculated from the data.

The third element of GLM is the link function, which specifies the link between the two previous components. It describes the linear relationship between the mean of the response Y, $E(Y) = \mu_i$, and the linear predictors:

$$f(\mu_i) = \eta_i = \beta X^T \tag{5.13}$$

In fact, in GLM, instead of using Y as the outcome, it is used a function of its mean. The β_j parameters can be obtained using maximum likelihood estimation.

5.2.6.1 Binary Responses

Binary dependent variables belongs to the binomial family. In GLM regression, the link function for binomial dependent variables can be either a logit or a probit link function, being the first one the more used one. The logit link is given by

$$f(\mu_i) = logit(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) \tag{5.14}$$

In this case, the mean of the Y values is equivalent to the probability π, which is the success probabibility (Y=1), i.e. the probability of the positive class. Then, GLM for binomial distributions with a logit link is equivalent to the more well-known term logistic regression:

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_k x_{ik}}}{1+e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+...+\beta_k x_{ik}}} \tag{5.15}$$

5.2.6.2 Continuous Responses

For continuous responses, several link functions may be used, depending on the data distribution. For the distributions being considered, the log identity link $f(\mu_i) = log(\mu_i)$ is used for the gamma distribution, and the identity link $f(\mu_i) = \mu_i$ is used for the remaining ones, i.e. normal, log-normal, Weibull and exponential [82]. In

the identity link, the mean is model directly, as no function was used. The mean of these distributions are the ones represented in table 5.1.

Under the assumption that the errors are normally distributed, which is typically used when the response is gaussian, the MLE results coincide with the ordinary least squares (OLS) algorithm, i.e. to maximize the likelihood function is equivalent to minimize the squared residuals.

5.2.7 Confidence and Prediction Intervals

The confidence and the prediction intervals correspond to the uncertainty in the estimation of the mean response and in the prediction of the response of a new observation in a regression problem, respectively. While the confidence interval (CI) refers to $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$, the prediction interval (PI) refers to $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon_i$, which means that the PI accounts for an extra variability and therefore it will be always wider than the CI.

Figure 5.4 shows an arbitrary example of a linear regression fit for a single predictor. The best-fit line is the blue one, the CI is the grey shadow and the PI is defined by the dashed red lines.

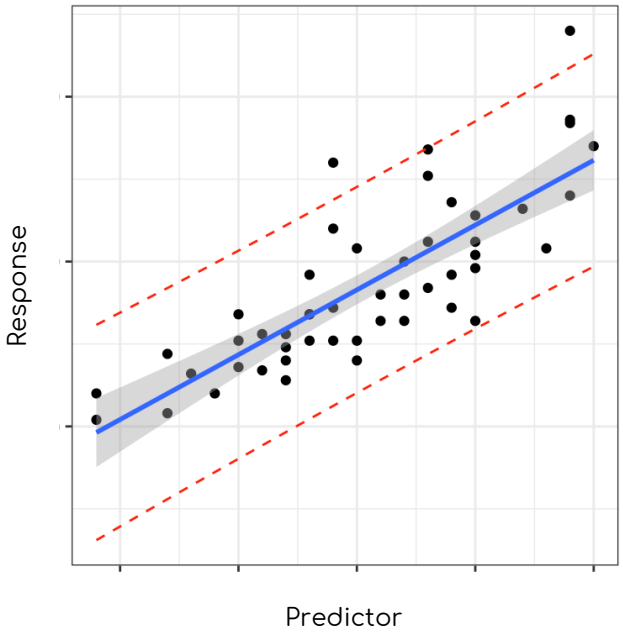


Figure 5.4: Simple linear regression with the best-fit line and the corresponding confidence and prediction intervals. Adapted from `http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/`.

For a multiple linear regression, the (1-α)·100% prediction interval is computed as follows:

$$\text{PI} = y_0 \pm t^*_{\alpha/2,n-2} \cdot s_{y \cdot x} \cdot \sqrt{1 + x_0(X'X)^{-1}x'_0} \tag{5.16}$$

where $t^*_{\alpha/2,n-2}$ is the z-value for the chosen α level (which is equal to 1.96 for $\alpha = 0.05$, i.e. a PI of 95%), x_0 is the row-vector with the predictors values of the new sample, y_0 is the predicted mean value for that x_0, X is the matrix with the predictors' values used to compute the best-fit line, and $s_{y \cdot x}$ is the standard deviation of the residuals, given by

$$s_{y \cdot x} = \sqrt{\frac{\sum_{i=1}^{n}(residual_i)^2}{n-k}} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-k}} \tag{5.17}$$

where n is the sample size and k is the number of parameters (the intercept and the selected effects).

5.2.8 Logistic Ridge Regression

The ridge regression is a penalized least squares method that besides the minimization of the sum of squared residuals (performed in simple OLS), imposes a penalty to the regression coefficients β, shrinking the less important parameters towards zero [100]:

$$\hat{\beta}_{\text{Ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{d} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{5.18}$$

where the first term is the standard least squares loss function, being y the n-dimensional outcome vector and X the $n \times p$ design matrix, and the second term is the additional penalization imposed by ridge. The β is the p-dimensional vector of the estimated parameters and λ is the (non-negative) tuning parameter that controls the degree of regularization, i.e. the larger the λ value, the greater the shrinkage. The penalization term is also known as L2-norm.

Similarly, the logistic ridge regression is defined as follows:

$$\hat{\beta}_{\text{Log.Ridge}} = \arg\min_{\beta} \left\{ \sum_{i=1}^{d} \left[y_i \left(\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j \right) + log \left(1 + e^{\beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j} \right) \right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\} \tag{5.19}$$

The method is said to provide a good balance between overfitting and underfitting and to deal well with multicollinearity problems [100].

5.2.9 Support Vector Machine

Support Vector Machine (SVM), introduced by Vapnik and Chervonenkis, is a supervised machine learning algorithm that can be used for both regression and classification problems, but it is mostly used for classification tasks, in which it was used in this document.

The goal of the method is to find the hyperplane in the n-dimensional space (where n is the number of predictors) that best classifies the data points, i.e. that best differentiate the two classes of a binary problem. This hyperplane is the one that maximizes the distance between the nearest points of each class: those points are known as support vectors and the distance is referred to as margin.

The hyperplane corresponds then to a decision boundary, where data points falling on opposite sides of it are attributed to different classes. Figure 5.5 shows a graphical representation of the SVM mechanism for two predictors. The margin is defined as $\frac{2}{||w||}$, where w is the weight vector of the figure.

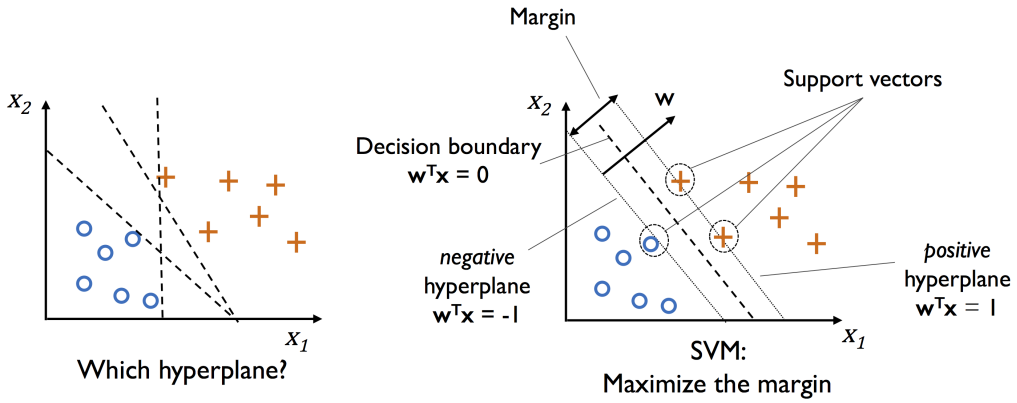


Figure 5.5: Hyperplanes and support vectors in SVM. Extracted from Raschka et al. [101].

For non-perfectly separable problems, the width of the margin is controlled by the C parameter, as represented in figure 5.6: a larger value of C leads to a smaller margin, and a smaller value of C to a larger margin. This parameter states the trade-off between to set a larger margin and to lower the misclassification rate (how many points are misclassified by the model).

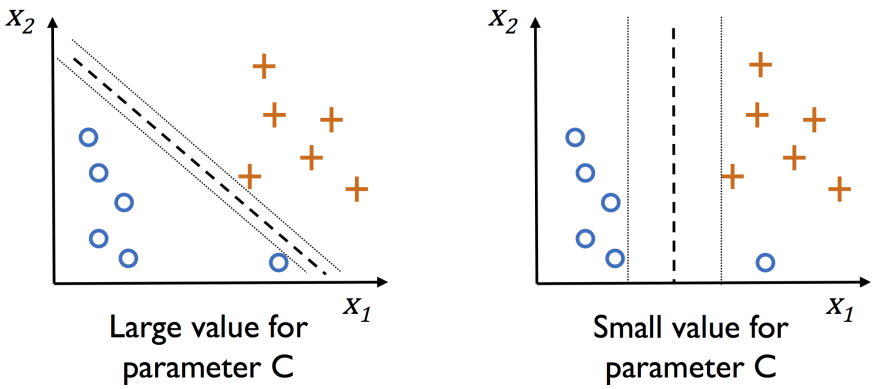


Figure 5.6: Trade-off in the choose of the C value. Extracted from Raschka et al. [101].

The SVM technique can be also extended to non-linear problems through the use of a kernel function, but it will not be discussed in this document as only linear SVM is used.

5.2.10 Repeated k-Fold-Stratified Cross Validation

In k-fold cross-validation (CV), the dataset is splitted into k approximately equal-sized groups of samples, named as folds. For each one of those k subsets, the samples are predicted using the model fitted (trained) considering the remaining k-1 subsets. Therefore, each sample is validated exactly once and it is used to fit the data k-1 times. For each fold, a prediction error is estimated and a total prediction error is obtained averaging the errors of all k folds, giving then information about the general prediction ability of the model.

For a high k value, the bias tends to be low because a big part of the total dataset is used to train the model, so it will be closer to the real one. However, the variance tends to increase because a smaller number of data points is used to validate the model each time, mainly when outliers are present. For a low k value, due to opposite reasons, the bias tends to be higher and the variance tends to be lower. Generally, k=5 or k=10 are recommended as good bias-variance compromise [102, 103]. Figure 5.7 represents the 10-fold CV procedure, where E are the errors.

Figure 5.7: Representation of 10-fold cross-validation. Extracted from `http://karlrosaen.com/ml/learning-log/2016-06-20/`.

Due to the small sample size nature of the dataset and the characteristics of the optimal experimental design, the estimated error of a k-fold CV would rely too much on the samples used in each one of the folds. To overcome this issue, two approaches were taken into account: stratification and repetition.

Using stratification, the proportion of each class in the full dataset is approximately maintained in each one of the folds. One the one hand, this ensures that each fold is a good representative of the problem, which is suitable in such small dataset. One the other hand, it guarantees that some metrics used to evaluate the models (sensibility and specificity) can be calculated, which would not be possible if, for example, all the samples in the test set belonged to a unique class.

Using repetition, the CV is repeated N times with different and random distribution of the samples in the folds, in each run. In the context of the problem being studied, different folds may lead to a very different performance, and this approach leads to a more stable and robust procedure for performance estimation.

More specifically, in order to get a trustful result, it was performed 500 repetitions of 10-fold-stratified cross-validation.

5.2.11 Performance Measures

5.2.11.1 R^2 and Generalized R^2

The coefficient of determination, most known as R-squared or simple R^2, represents the proportion of the variance in the dependent variable that is explained by the

independent variable(s), in a regression model. It can be used as a goodness of fit and it is given by

$$\mathrm{R}^2 = 1 - \frac{\text{Explained Variance}}{\text{Total Variance}} = 1 - \frac{SS_{total}}{SS_{residuals}} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)}{\sum_{i=1}^{n}(Y_i - \bar{Y})} \tag{5.20}$$

where SS_{total} is the total sum of squares and $SS_{residuals}$ is the sum of squares of the residuals.

Although the R^2 can have negative values in some cases (when the fit is worse than the mean of the data), it ranges normally from 0 to 1, where a higher value typically indicates a better fit to the outcome: a R^2 equal to 1 means that the fit is perfect.

For regression models others than ordinary least squares, a normalized pseudo-R^2, which will be referred to as generalized R-squared, was used [82]. This adjusted pseudo-R^2, also known as the Nagelkerke R^2 or the Craig and Uhler R^2, is a normalized version of Cox and Snell's pseudo-R^2 [104].

The generalized R^2 compares the likelihood of the fitted model (L_M) to the likelihood of the intercept-only model (L_0) and it is scaled to have a maximum value of 1. For a binomial response distribution, the generalized R^2 is given by

$$\text{Generalized } \mathrm{R}^2 = \frac{1 - \left(\frac{L_0}{L_M}\right)^{2/n}}{1 - (L_0)^{2/n}} \tag{5.21}$$

and for the remaining distributions it is defined as

$$\text{Generalized } \mathrm{R}^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n} \tag{5.22}$$

where n is the sample size.

As the traditional R^2, it ranges from 0 to 1, and for normal responses it is simplified to the standard R-square.

It is very important to highlight that once the generalized R^2 value is based on the intercept-only model, we must not compare those values for models built on different response distributions.

5.2.11.2 Accuracy, Sensitivity and Specificity

The performance of a given classifier is normally evaluated based on a confusion matrix. This matrix represents the distribution of the true versus the predicted classes of validation or test samples, and it is composed by four different values: true positive (TP), positive samples correctly classified; true negative (TN), negative samples correctly classified; false positive (FP), negative samples wrongly classified; false negative (FN), positive samples wrongly classified. Recapping, in this problem, the positive class is the target response and the negative class is the not desired one.

From those values, several evaluation metrics can be calculated, namely accuracy, sensitivity and specificity.

The sensitivity represents how well the classifier correctly predicts the positive cases:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{5.23}$$

The specificity represents how well the classifier correctly predicts the negative cases:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5.24}$$

The accuracy is a summary of the former two metrics and represents the general ability of the classifier to correctly predict a random sample:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.25}$$

Due to the binary and balanced nature of the dataset, the combination of the referred metrics gives a robust analysis of the problem.

5.2.12 Desirability Functions

The desirability function approach, firstly proposed by Harington [105], is one of the most simple and widely used methods in the industry to simultaneous optimization of several different responses. The basic idea is to transform the model prediction equation of each response into an individual desirability on a scale [0,1] and then combine the individual desirabilities into a single value, denoted by overall desirability, using the geometric mean.

So, if a given experiment is associated to n responses $Y = (y_1, ..., y_n)$, to each fitted

response $\hat{y}_i$ will be assigned a desirability function $0 \leq d_i \leq 1$, where d_i=0 represents a completely undesirable value of y_i and d_i=1 a completely desirable one. The overall desirability D is then given by

$$D = \left(\prod_{i=1}^{n} d_i \right)^{1/n} = (d_1 \times ... \times d_n)^{1/n} \tag{5.26}$$

which can be extended to give different importances to each response, originating a weighted geometric mean:

$$D = \left(\prod_{i=1}^{n} d_i^{w_i} \right)^{1/\sum_{i=1}^{n} w_i} = d_1^{w_1} \times ... \times d_n^{w_n} \tag{5.27}$$

The geometric mean has the desired property that if any response is completely undesirable (d_i=0), then the overall desirability will be also inadmissible (D=0).

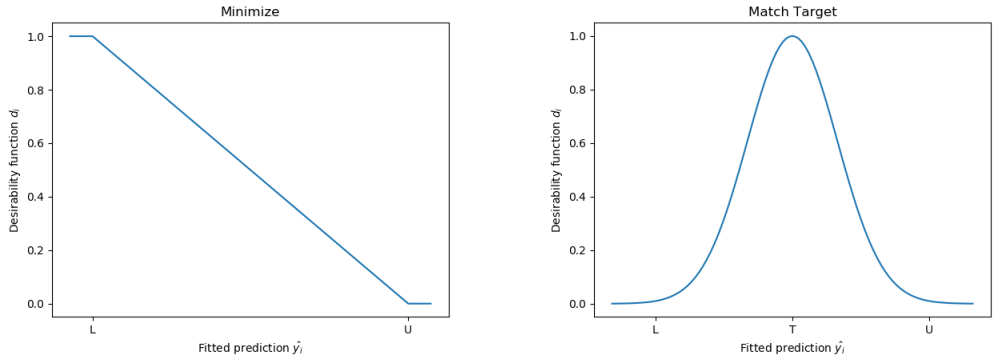
For the individual desirabilities, several different functions can be used. For the continuous responses, the most standard ones are the functions proposed by Derringer and Suich [106]. In the CQAs whose goal was to minimize the response, their transformation with a unitary scale was applied:

$$d_i^{minimize} = \begin{cases} 0 & \text{if} \quad \hat{y}_i > U \\ \frac{\hat{y}_i - L}{U - L} & \text{if} \quad L \leq \hat{y}_i \leq U \\ 1 & \text{if} \quad \hat{y}_i < L \end{cases} \tag{5.28}$$

For the CQAs whose objective was to achieve a given target, a smoother transformation than their suggestion was considered:

$$d_i^{target} = \frac{\text{PDF}_{\text{normal}}\left(\hat{y}_i \mid \mu = T, \sigma = \frac{-\Delta^2}{2 \cdot ln(0.01)}\right)}{\text{PDF}_{\text{normal}}\left(T \mid \mu = T, \sigma = \frac{-\Delta^2}{2 \cdot ln(0.01)}\right)} \quad , \quad \Delta = L - T = U - T \tag{5.29}$$

where $\text{PDF}_{\text{normal}}$ is the probability density function of the normal distribution, L is the lower bound, U is the upper bound, and T is the target. These transformations are represented in figure 5.8. The upper and lower bounds for each response were defined from the intervals described in table 4.1, and the target one was establish as the media value of the two bounds, as shown in table 5.3.



(a) Minimize. **(b)** Match target.

Figure 5.8: Individual desirabilities functions for continuous responses.

Table 5.3: Control points of the desirability functions.

Response	**Goal**	**Lower bound** L	**Upper bound** U	**Target** T
CQA-C.1	Match target	24	31	27.5
CQA-C.2	Minimize	0	0.6	-
CQA-C.3	Minimize	0	1.6	-

For the binary responses, the probability of obtaining the positive class was considered as the individual desirability because it is already a value in the range [0,1], and so it can be directly used without any transformation. For those responses, the desired value was set to 1, i.e. the goal is to maximize the response.

In this work, the same importance was given to each CQA. In order to determine the factor settings (i.e. the values of the predictors) that maximize the overall desirability function, an optimization procedure is required. In this case, the gradient descent algorithm is used.

5.3 Software

The effects selection through stepwise regression, generalized linear models and AICc validation was implemented using the JMP Pro 14 (Generalized Regression personality [82]) and the JMP Scripting Language; the desirabilities functions were also performed using this software. The fitting of continuous distributions, the statistical analysis, the classifiers and prediction intervals construction, and all the data visualization were implemented in Python.

6

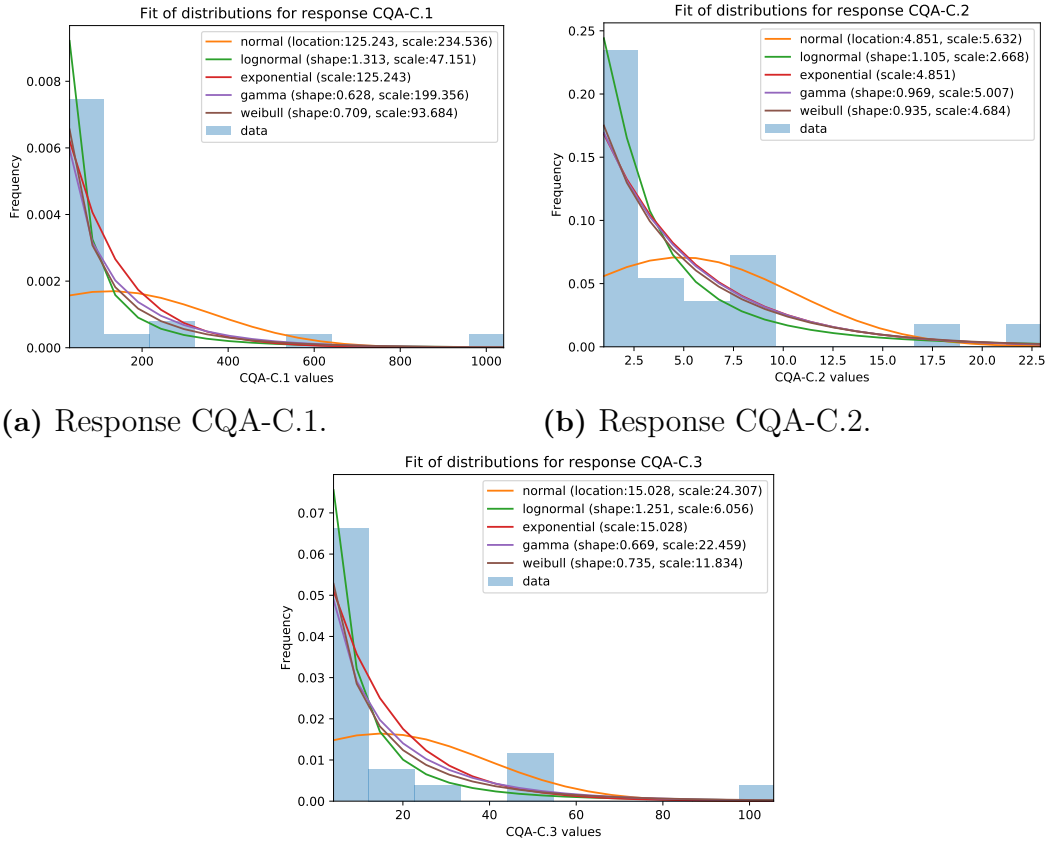
Results and Discussion

In this chapter, all the results of the aforementioned procedures are presented and discussed. For the identification of important effects, the models built using only main effects will be also referred to as first-order models and the ones that also considered two-factor interactions will also be designed as second-order models. Only the best selected models, the ones with higher R^2 from the models selected through ΔAICc criterion, will be discussed; the remaining ones may be consulted in the Appendices.

6.1 Distribution of Continuous Responses

Figure 6.1 shows a graphical representation of the fit of the four positive skewed distributions to each one of the continuous responses. The normal distribution was also considered to establish a comparison. The best parameters (location, shape and scale) obtained using the maximum likelihood estimation for each distribution are also present in each sub-figure.

The goodness of fit values of the fitted distributions are shown in tables 6.1 (response CQA-C.1), 6.2 (response CQA-C.2) and 6.3 (response CQA-C.3). Both corrected Akaike Information Criterion (AICc) value and Kolmogorov-Smirnov (K-S) p-value are presented. For each response, the best values, i.e. higher p-value for K-S test and lower value for AICc, are in bold.



(a) Response CQA-C.1.

(b) Response CQA-C.2.

(c) Response CQA-C.3.

Figure 6.1: Fit of the distributions to the continous responses.

Table 6.1: Goodness of fit for the response CQA-C.1.

	Normal	**Lognormal**	**Exponential**	**Gamma**	**Weibull**
K-S p-value	0.006	**0.955**	0.023	0.192	0.490
AICc value	334.646	**270.716**	282.034	280.259	277.401

Table 6.2: Goodness of fit for the response CQA-C.2.

	Normal	**Lognormal**	**Exponential**	**Gamma**	**Weibull**
K-S p-value	0.115	**0.636**	0.224	0.256	0.368
AICc value	155.646	**124.581**	125.359	128.359	128.172

Table 6.3: Goodness of fit for the response CQA-C.3.

	Normal	**Lognormal**	**Exponential**	**Gamma**	**Weibull**
K-S p-value	0.016	**0.116**	0.002	0.046	0.115
AICc value	255.837	**169.863**	180.257	179.630	177.440

From table 6.1, related to response CQA-C.1, the K-S null hypothesis (H0) is rejected for both normal and exponential distributions. For the 3 remaining distributions, the lognormal is the one whose H0 is not rejected with more confidence (very high confidence for instance) and it is also the one with an AICc value significantly lower than the other distributions.

About table 6.2, related to response CQA-C.2, the K-S H0 is not rejected for all distributions. Even so, the lognormal one is the one whose non-rejection is more trustful and also the one with a lower AICc value.

From table 6.3, related to response CQA-C.3, only for the Weibull and lognormal distributions the K-S H0 is not rejected, even that for these distributions the p-value is substantially small. Similarly to the first response, the AICc value is significantly lower for the lognormal distribution.

Taking into account these observations, which result from the evaluation of two different goodness of fit measures, the lognormal distribution seems to be the most suitable one to represent all the three responses data. The results of Shapiro-Wilk (S-W) and D'Agostino-Pearson (D-P) statistical tests for the analysis of normality of the $\log_{10}$ transformation are shown in table 6.4.

Table 6.4: The p-values for the normality statistical tests after the log-transformation of the responses data.

Response	**S-W p-value**	**D-P p-value**
CQA-C.1	0.625	0.453
CQA-C.2	0.254	0.265
CQA-C.3	0.007	0.160

The results of table 6.4 agree with the previous observations. The data of response CQA-C.1 have clear evidence of following a lognormal distribution while there are some doubts about the response CQA-C.3. In fact, for this response, the H0 of normality is rejected for the Shapiro-Wilk test. Even so, the D'Agostino-Pearson test result supports the idea of normality of the log-transformation of the data.

Therefore, the most suitable approach seems to be the use of GLM for the lognormal response, which will be referred to as lognormal-GLM, from now on. However, in order to analyze the benefits of the proposed approach, the standard least squares method without any type of transformation will be applied as well.

6.2 Important Effects in Binary Responses

6.2.1 Detection of Active Effects for Binary Responses

The information about the best selected models for the two binary responses is presented in tables 6.5 (response CQA-D.1) and 6.6 (response CQA-D.2). The measures of fit, AICc and generalized R^2 values, are shown for each model. The proposed procedure is represented in the tables by 'SR BEH LR AICc', which stands for 'Stepwise Regression with Back-Enforcement of Heredity, Logistic Regression and AICc validation'.

Table 6.5: Best selected models for response CQA-D.1.

Selection Method	**Effects type**	**Selected effects**	**AICc value**	**Gener. R^2**
SR BEH LR AICc	**1st order**	FP-1.1, PP-2.2, FP-3.5, FP-2.1, FP-5.1	22.41	0.91
	2nd order	FP-1.1, PP-2.2, FP-3.5, FP-2.1, FP-5.1	22.41	0.91

Table 6.6: Best selected models for response CQA-D.2.

Selection Method	**Effects type**	**Selected effects**	**AICc value**	**Gener. R^2**
SR BEH LR AICc	1st order	FP-3.4, FP-4.1	34.55	0.292
	2nd order	PP-2.2, PP-3.1, FP-4.1, PP-2.2*PP-3.1, PP-2.2*FP-4,1	**26.93**	0.828

Analyzing the table 6.5, it is possible to observe that both first-order and second-order models are equal, which means that for this response, even when two-factor interactions are considered, only main effects are selected. This is an important consideration because it shows that the methodology only select interactions if they are indeed more important than the main effects, otherwise only main factors are selected.

About table 6.6, it is clear the benefits of taking into consideration the two-factor interactions, as a much better model is obtained for this case, i.e. the AICc difference value between first and second-order models is very significant. Besides, the objective measure, the generalized R^2, is very low for the first-order model and it has a very good value for the second-order one.

Therefore, the models selected as the ones that better describe the dependent variables are then the second-order one for the latter response and the only one that

was shown for the former response, the ones in bold.

6.2.2 Recommended Levels for Response CQA-D.1

The parameters estimates obtained for the selected effects for this response are shown in table 6.7.

Table 6.7: Estimates and p-values for selected effects for response CQA-D-1.

Effect	Estimated coefficient	Wald p-value
PP-2.2	10.9177	<0.0001
FP-3.5	-11.5785	<0.0001
FP-1.1	3.4058	<0.0001
PP-5.1	-6.7781	<0.0001
PP-2.1	-6.9034	<0.0001

From table 6.7, there are only main effects and so all of them are selected because they are statistically significant. Taking into consideration that the target response is the positive class, which can be seen as response maximization, it is possible to observe that the high level is suggested for the factors PP-2.2 and FP-1.1, and the low level is recommended for the other ones, the factors PP-2.1, FP-3.5 and PP-5.1.

6.2.3 Recommended Levels for CQA-D.2

The parameters estimates obtained for the selected effects for this response are shown in table 6.8 and the interaction plots of the selected two-factor effects are displayed in figure 6.6.

Table 6.8: Estimates and p-values for selected effects for response CQA-D.2.

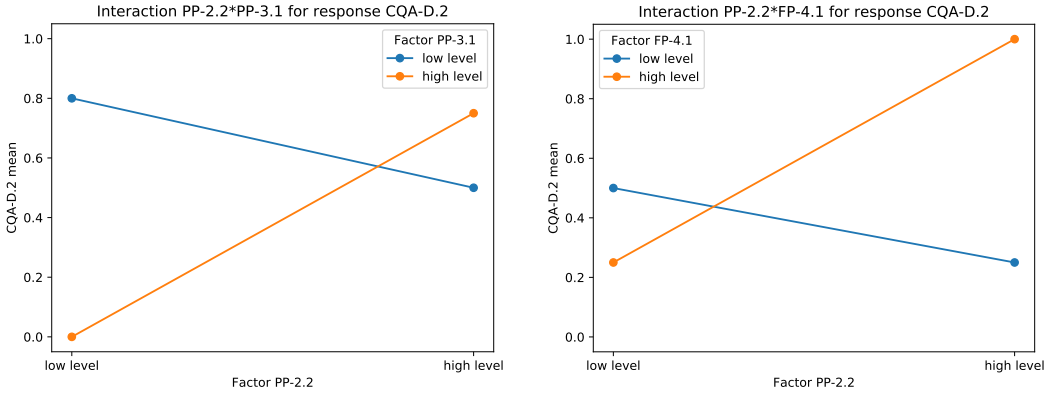
Effect	Estimated coefficient	Wald p-value
PP-2.2	6.5963	<0.0001
FP-4.1	8.5315	<0.0001
PP-3.1	-0.4319	0.3472
PP-2.2*FP-4.1	8.3962	<0.0001
PP-2.2*PP-3.1	1.3790	0.0012

From table 6.8, it is possible to observe that the main effects PP-2.2 and FP-4.1 are statistically significant, while the effect PP-3.1 only enter in the model because it belongs to one of the selected interactions.

Taking into consideration the estimates, and knowing that the goal is to have a positive response, i.e. to maximize the response, it is suggested a high level for

both PP-2.2 and FP-4.1 factors. The analysis of their interaction is not required because their levels are already defined by their individual impact, but studying the sub-figure 6.2b it is clear that when there is an interaction of the parameters at their high level the response is equal to one, which supports the aforementioned idea.

In contrast, the suggested level for the factor PP-3.1 derives from the analysis of its interaction. From subfigure 6.2a, it is possible to observe that there are two combinations that maximize the response: 1) both parent factors at their high level or 2) both parent factors at their low level. The other parent of the interaction is the factor PP-2.2, which is individually significant and so its suggested level must be followed, which is the high one. So, the first option is the most suitable one and the factor PP-3.1 should be at the high level too.



(a) Interaction PP-2.2*FP-3.1. **(b)** Interaction PP-2.2*FP-4.1.

Figure 6.2: Interaction plots of selected interactions for response CQA-D.2.

6.3 Important Effects in Continuous Responses

6.3.1 Detection of Active Effects for Continuous Responses

For the continuous responses, the lognormal distribution was the one with best fit to the data of the three outcomes being considered, and so lognormal-GLM seems to be the best approach to model the independent variables into the dependent one. However, in order to analyze the benefits of the suggested procedure, the stepwise regression with back-enforcement of heredity using the ordinary least squares (OLS), without any type of transformation, was applied as well. Those procedures will be referred to in the tables as 'SR BEH lognorm-GLM AICc', which stands for 'Stepwise Regression with Back-Enforcement of Heredity, lognormal Generalized Linear Model and AICc validation', and 'SR BEH OLS AICc', which stands for

'Stepwise Regression with Back-Enforcement of Heredity, Ordinary Least Squares and AICc validation'.

As stated before, the Bluepharma experimenters had already performed an analysis on the continuous responses, using a standard approach commonly executed in the pharmaceutical industry: combination of least squares estimations and ANOVA measures, considering only the main effects and without any type of response transformation. Those results will also be displayed to establish some comparisons between the traditional technique and the proposed procedure, and will be referred to in the tables as 'Standard Approach'.

Tables 6.9 (response CQA-C.1), 6.10 (response CQA-C.2) and 6.11 (response CQA-C.3) show the information about the selected effects considering all the approaches aforementioned. The same measures of binary responses are shown: in this case, Gen. R^2 (generalized R^2) is referring to the lognormal-GLM method and R^2 is related to the OLS methods (both OLS with stepwise procedure and the standard analysis).

The results of tables 6.9-6.11 show again the importance of considering the interaction effects. There is a general improvement of both AICc values (they decrease) and generalized R^2 or R^2 (they increase) when we go from the first-order models (only main effects) to the second-order ones (both main effects and interactions), for both OLS and lognormal-GLM approaches.

Table 6.9: Selected models for response CQA-C.1.

Selection Method	**Effects type**	**Selected effects**	**AICc value**	**Gen. R^2 or R^2**
SR BEH lognorm-GLM AICc	1st order	FP-1.3, PP-3.1, FP-4.1, FP-1.2	266.20	0.51
	2nd order	FP-1.3, PP-2.2, FP-3.4, PP-5.1, FP-1.3*PP-5.1, PP-2.2*FP-3.4	**265.73**	0.66
SR BEH OLS AICc	1st order	FP-1.3, FP-4.1, FP-3.3	333.33	0.34
	2nd order	FP-1.3, PP-2.2, FP-3.4, PP-5.1, FP-1.3*PP-5.1, FP-1.3*PP-2.1, PP-2.1*PP-5.1, FP-3.4*PP-5.1	320.36	0.85
Standard Approach		FP-1.3, FP-4.1, FP-3.3	333.33	0.34

Table 6.10: Selected models for response CQA-C.2.

Selection Method	Effects type	Selected effects	AICc value	Gen. R^2 or R^2
SR BEH lognorm-GLM AICc	1st order	FP-1.3, PP-2.1, FP-3.1, FP-3.3, FP-4.1	116.31	0.64
	2nd order	FP-1.3, PP-2.1, FP-3.3, FP-3.5, FP-4.1, PP-5.2, FP-3.3*PP-5.2	**113.80**	0.79
SR BEH OLS AICc	1st order	FP-1.1, PP-2.1, FP-3.3, FP-3.5, FP-4.1	145.74	0.67
	2nd order	FP-1.1, PP-2.1, FP-3.5, FP-4.1, FP-1.1*FP-2,1, FP-1.1*FP-3.5, FP-2.1*FP-3.5, FP-2.1*FP-4.1, FP-3.5*FP-4.1	132.82	0.93
Standard Approach		PP-2.1, FP-3.3, FP-3.5, FP-4.1	145.92	0.60

Table 6.11: Selected models for response CQA-C.3.

Selection Method	Effects type	Selected effects	AICc value	Gen. R^2 or R^2
SR BEH lognorm-GLM AICc	1st order	PP-3.1, FP-4.1, PP-5.2	168.55	0.34
	2nd order	PP-2.1, PP-2.2, PP-3.2, FP-3.4, PP-5.1, PP-2.2*PP-3.2, PP-2.2*FP-3.4, PP-2.1*PP-2.2	**153.05**	0.87
SR BEH OLS AICc	1st order	FP-2.1, FP-3.1, FP-3.5, FP-4.1	225.57	0.41
	2nd order	PP-2.2, PP-5.1, PP-2.2*PP-5.1	222.21	0.40
Standard Approach		PP-2.1, FP-3.3, FP-3.5, FP-4.1	225.57	0.41

It is possible to note that the standard approach gives similar results to the first-order stepwise OLS (it gave equal models for two responses and an almost equal one to the other response).

The analysis of the same tables also allows to observe that the AICc values are significantly lower for the method using the lognormal-GLM, and then it is possible to conclude that it is a much better approach than the standard OLS and, consecutively, it is much more suitable than the methodology followed typically in the pharmaceutical industry. It is important to remember that the R^2 values of lognormal-GLM and OLS must not be directly compared because they are based on the intercept-only model; however, the AICc value can be considered for that purpose as it was used in other works [107, 108].

An interesting observation is that not only the AICc values are lower for the lognormal-GLM but also the selected models agree more with the effect principles, more specifically with the hierarchy one, mainly for the two first responses. For the response

CQA-C.1, using lognormal-GLM are selected 4 main effects and 2 interactions while using OLS are selected 4 main effects and 4 interactions. For the response CQA-C.2, are chosen 6 main effects and 1 interaction using lognormal-GLM, while using OLS are selected 4 main effects and 5 interactions. So, the proposed procedure not only gives a better practical solution but also a better theoretical one.

These considerations already present some meaningful conclusions. In order to get some more insights, it was displayed the predicted versus real value points of those responses using the best models, i.e. the second-order ones of both OLS and lognormal-GLM stepwise approaches, presented in the previous tables.

Figures 6.3 (response CQA-C.1), 6.4 (response CQA-C.2) and 6.5 (response CQA-C.3) show the comparison of those fits. The left block of sub-figures is related to the OLS fitted model and the right block is related to the lognormal-GLM fitted model. The top sub-figures show all the predicted versus real points while the bottom sub-figures show a zoomed version of the top ones, in the area closer to the target values. For the zoomed plots, it is also represented the confidence intervals for each point, in vertical bars.

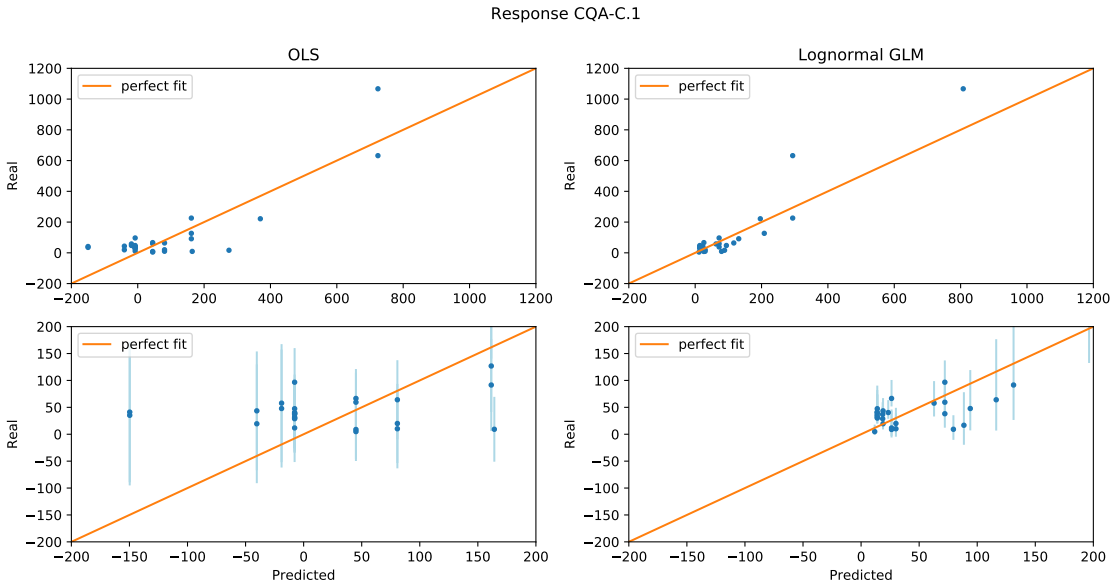


Figure 6.3: Fit for response CQA-C.1, using both OLS and lognormal-GLM methods. Top sub-figures: zoom out (all points); bottom sub-figures: zoom in (target points).

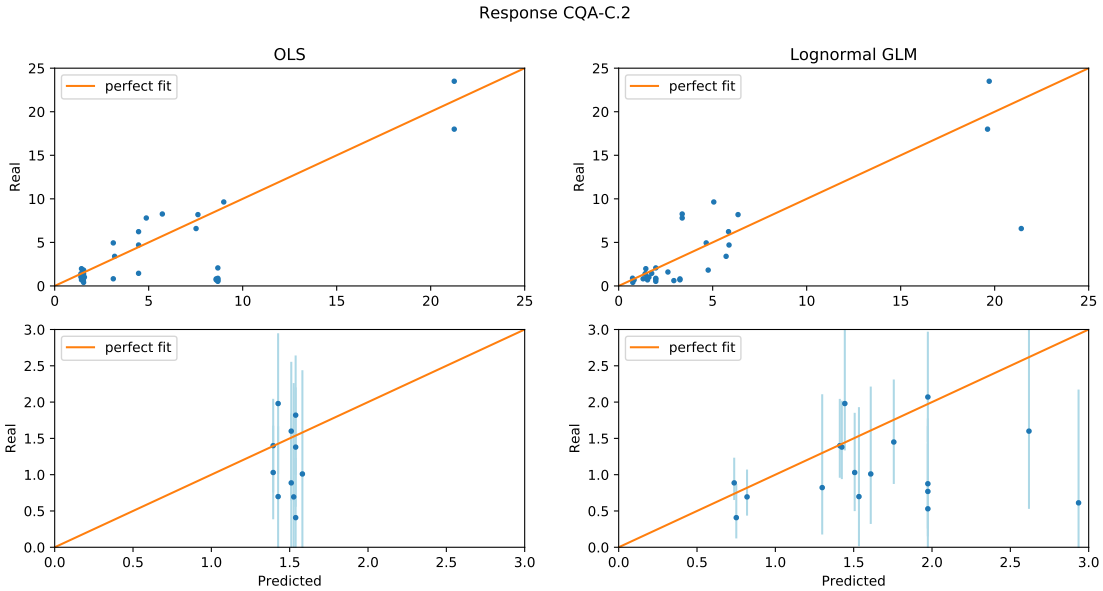


Figure 6.4: Fit for response CQA-C.2, using both OLS and lognormal-GLM methods. Top sub-figures: zoom out (all points); bottom sub-figures: zoom in (target points).

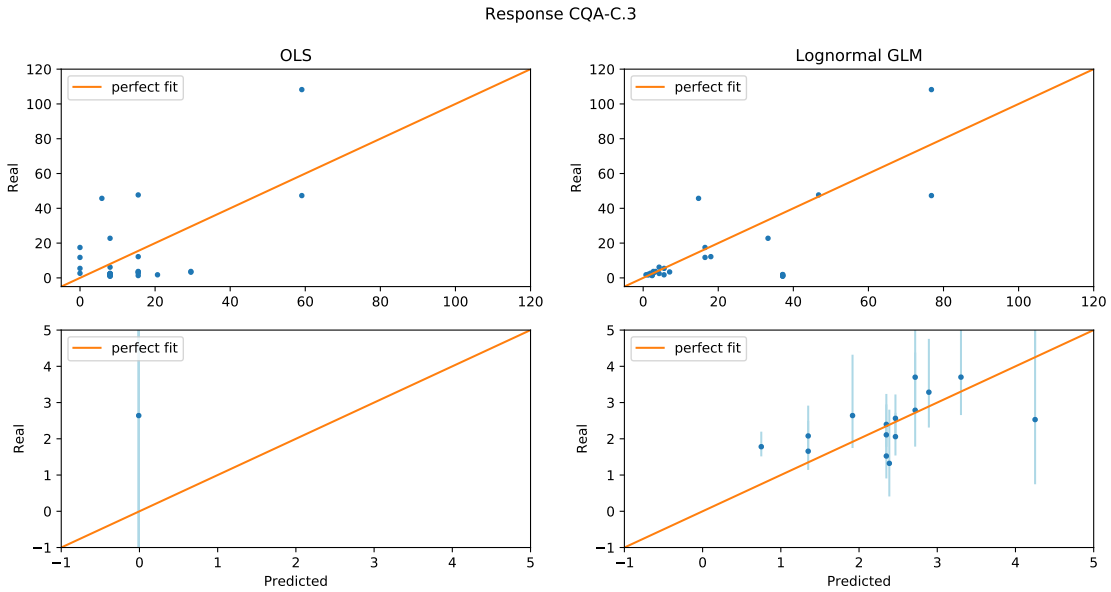


Figure 6.5: Fit for response CQA-C.3, using both OLS and lognormal-GLM methods. Top sub-figures: zoom out (all points); bottom sub-figures: zoom in (target points).

For all of the responses, it is possible to observe that more than the fit to be better in general, it is especially much better in the points that are close to the desired values: inside the interval [0, 50] for response CQA-C.1, and lowest values for responses CQA-C.2 e CQA-C.3. This is mainly observed for the first response but is also

evident for the other two responses.

For the two first outcomes, the number of predictors is lower for the lognormal-GLM and even so the fit is better. For the last response, the number of predictors is lower in the OLS method so a direct comparison should not be performed, as it is expected to get a better fit as more predictors are considered. However, it is possible to see that when the OLS is used, the predicted values are far away from the real ones for the points close to the desired interval (only one point appears in the zoomed plot).

Another important remark is that the confidence intervals are much lower for the points close to the target when the lognormal-GLM method is used. It informs that not only the fit is better, but also the precision and certainty of it.

These observations are very relevant for the problem being studied as it shows that when lognormal modeling is considered, it is obtained a model whose predictors explain much better the response values closer to the target ones than if a standard method is used. In this type of problem, where the main goal is to achieve the combinations of values that lead to the desired output, this is extremely important because it suggests that the target can be achieved much more efficiently, which is a major concern in pharmaceutical R&D.

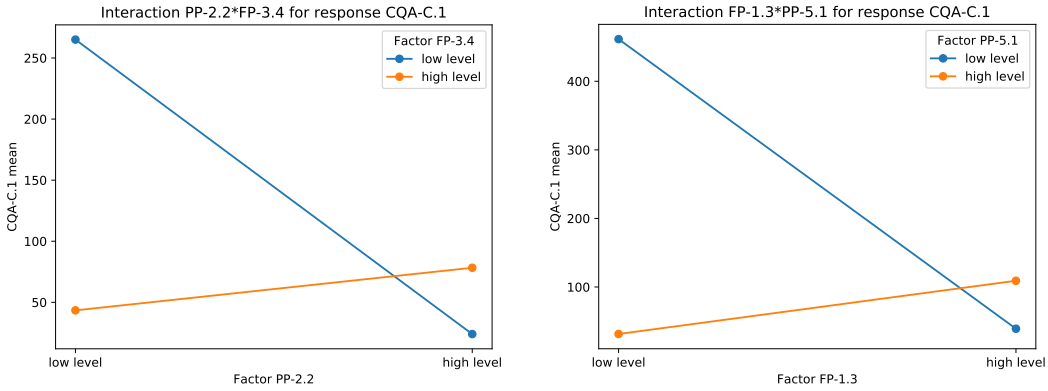
The models selected as the ones that better describe the dependent variables are then the second-order lognormal-GLM ones for the three continuous responses, the ones in bold in the tables.

6.3.2 Recommended Levels for Response CQA-C.1

The parameters estimates obtained for the selected effects for this response, using the lognormal-GLM, are shown in table 6.12 and the interaction plots of the selected two-factor effects are displayed in figure 6.6.

Table 6.12: Estimates and p-values for selected effects for response CQA-C.1.

Effect	**Estimated coefficient**	**Wald p-value**
FP-1.3	-0.5653	<0.0001
FP-3.4	-0.3525	0.0096
PP-2.2	-0.1255	0.3400
PP-5.1	0.0104	0.9433
FP-1.3*PP-5.1	0.8319	<0.0001
PP-2.2*FP-3.4	0.4723	<0.0001



(a) Interaction PP-2.2*FP-3.4 **(b)** Interaction FP-1.3*PP-5.1

Figure 6.6: Interaction plots of selected interaction for response CQA-C.1.

From table 6.12, it is possible to observe that the main effects FP-1.3 and FP-3.4 are statistically significant, while PP-2.2 and PP-5.1 only enter in the model because they are parents of important interactions. From this table and the interactions plots of figure 6.6, it is visible how the effects influence the response. In this particular case, the goal is neither to minimize neither to maximize the response, but achieve a specific target interval. So, it is harder to indicate which is the desired level for each factor because several different combinations can lead to it.

One possible solution to overcome it is to invert the regression model and find the null space, i.e. the *collection* of combinations of factor values that are expected to lead to the desired response values. However, this is beyond the scope of this work. Besides, from those possible combinations, it would be infeasible to experiment all of them, and once several other responses are being evaluated, the levels of the selected factor may be chosen to simultaneously satisfy the remaining CQAs, which limits that null space. This approach will be further discussed.

6.3.3 Recommended Levels for Response CQA-C.2

The parameters estimates obtained for the selected effects for this response, using the lognormal-GLM, are shown in table 6.13 and the interaction plot of the selected two-factor effect is displayed in figure 6.7.

From table 6.13, it is possible to observe that all main effects are statistically significant. The goal is to minimize the response, and so the recommended levels are: high level for factors PP-2.1 and PP-5.2; low level for factors FP-1.3, FP-3.3, FP-3.5 and FP-4.1.

Table 6.13: Estimates and p-values for selected effects for response CQA-C.2.

Effect	Estimated coefficient	Wald p-value
FP-4.1	0.6697	<0.0001
FP-3.3	0.4249	<0.0001
PP-2.1	-0.3151	0.0008
FP-3.5	0.2637	0.0011
FP-1.3	0.3054	0.0034
PP-5.2	-0.1891	0.0290
FP-3.3*PP-5.2	0.4345	<0.0001

The parent factors of the selected interaction are both individually significant, so the evaluation of the interaction plot was not necessary. However, analyzing the figure 6.7, the minimal value for the response is obtained when the factor FP-3.3 is at the low level and the factor PP-5.2 is at the high level, which is consistent with the aforementioned ideas, as expected.

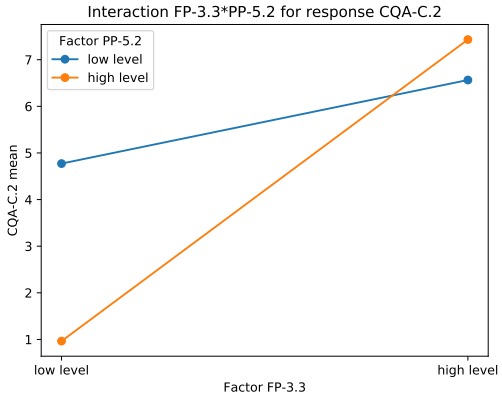


Figure 6.7: Interaction plots of selected interaction for response CQA-C.2.

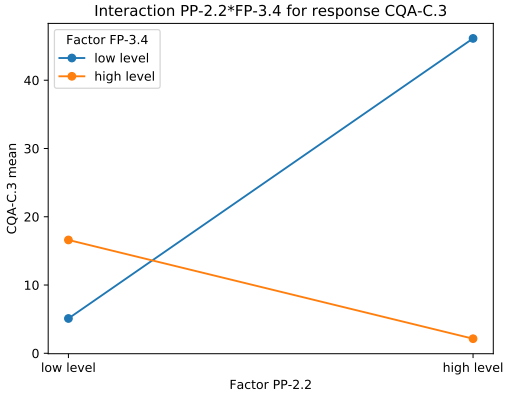
6.3.4 Recommended Levels for Response CQA-C.3

The parameters estimates obtained for the selected effects for this response, using the lognormal-GLM, are shown in table 6.14 and the interaction plots of the selected two-factor effects is displayed in figure 6.8.

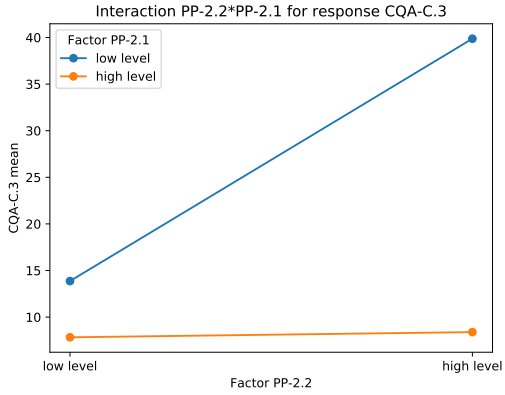
From table 6.14, it is possible to observe that the main effects PP-2.1, FP-3.4 and PP-5.1 are all statistically significant. Taking into consideration that the goal is to minimize the values of the response, a high level is recommended for all of them. The effects PP-2.2 and PP-3.2 are not significant which indicates that they only enter in the model because they are parents of important interactions, and therefore their recommended level will depend on those interactions, which are represented in figure 6.8.

Table 6.14: Estimates and p-values for selected effects for response CQA-C.3.

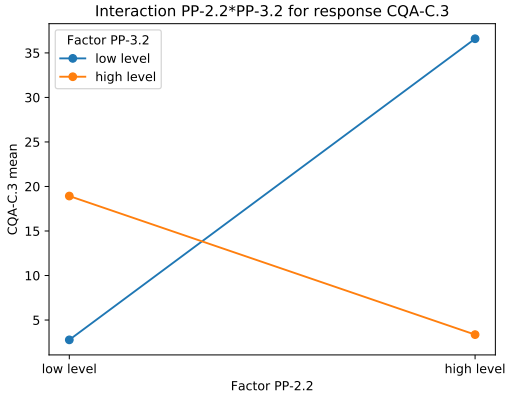
Effect	Estimated coefficient	Wald p-value
PP-2.1	-0.4497	<0.0001
FP-3.4	-0.3947	<0.0001
PP-5.1	-0.5653	<0.0001
PP-2.2	-0.1109	0.1339
PP-3.2	-0.0085	0.9005
PP-2.2*PP-3.2	-0.8912	<0.0001
PP-2.2*FP-3.4	-0.7767	<0.0001
PP-2.1*PP-2.2	-0.4673	<0.0001



(a) Interaction PP-2.2*PP-3.4.



(b) Interaction PP-2.2*PP-2.1.



(c) Interaction PP-2.2*PP-3.2.

Figure 6.8: Interaction plots of selected interactions for response CQA-C.3.

From sub-figure 6.8a, it is inspected that the interaction PP-2.2*FP-3.4 minimizes the response value for the pair of a high level of FP-3.4 and a high level of PP-2.2, which is feasible because as stated before the factor FP-3.4 should be at the high level. From sub-figure 6.8b, it is possible to see that, in order to have the minimum response value, as long as the PP-2.1 is at the high level, the level of PP-2.2 is not very significant. Finally, from sub-figure 6.8c, can be visualized that are two

possible approaches: 1) both parent factors at their high level and 2) both parent factors at their low level. As concluded for the first interaction, the factor PP-2.2 should be at the high level, and so the first option should be considered. Therefore, taking into consideration these observations, for this response, it is suggested the high level for both factors that are not significant by itself, PP-2.2 and PP-3.2, in order to approach the desired response.

6.4 Multiple Response Optimization

Table 6.15 presents a resume of the recommended levels for each response, identified in the previous sections. From this table, it is suggested that the factor FP-1.2 may be neglected in follow-up experiments. Besides, as expected it could happen when several responses are analyzed, some factors have contradictory levels, namely the factors PP-2.1, FP-4.1 and PP-5.1.

Table 6.15: Resume of the recommended levels of each factor for each response.

Factor	CQA-D.1	CQA-D.2	CQA-C.1	CQA-C.2	CQA-C.3
FP-1.1	high				
FP-1.2					
FP-1.3			*	low	
PP-2.1	low			high	high
PP-2.2	high	high	*		high
PP-3.1		high			
PP-3.2					high
FP-3.3				low	
FP-3.4			*		high
FP-3.5	low			low	
FP-4.1		high		low	
PP-5.1	low		*		high
PP-5.2				high	

The multi-response optimization procedure was applied to verify if it was possible at this stage of the process to obtain a combination of values which satisfy all the CQAs targets simultaneously. The nominal factor, FP-1.1, was identified for only one response and the recommended level was also considered by the experimenters as the suitable one in terms of safety. Therefore, its value was replaced by the suggested level value (1), so it can be considered in the analysis but not as a continuous factor, allowing the gradient descent algorithm to be freely used.

The results of the maximization of the overall desirability function of the obtained models are shown in table 6.16.

Table 6.16: Results of the maximization of the overall desirability.

Response	Goal	Target	Predicted value	Overall desirability
CQA-D.1	Positive class (1)		1	D= 0.713
CQA-D.2	Positive class (1)		0.908	
CQA-C.1	Match target	[24, 31]	27.521	
CQA-C.2	Minimize	[0, 0.6]	0.430	
CQA-C.3	Minimize	[0, 1.6]	0.462	

As it is possible to observe in table 6.16, it is possible to obtain a combination of values such that the predicted values for the responses are very satisfactory. For the continuous CQAs, all the values are inside the target interval, and for the binary CQAs, the predicted probabilities of obtain the target class are equal to 100% for one response and 90.8% for the other one.

It is important to highlight that the overall desirability value is dependent on the functions chosen to model each response and that the objective was just to find a good set of settings that met all the goals, and the desirability functions are merely mathematical methods to find an optimum value. In fact, it was impossible to obtain an overall desirability of 1 as it would imply that the predicted values of CQA-C.2 and CQA-C.3 are 0, which is physically impossible.

Besides, the optimization is not a screening procedure and it was applied here as a *proof of concept* of the proposed models. In order to establish a comparison with the models obtained by the company practitioners using the standard pharmaceutical methodology, the desirability functions were also applied to those models, but the overall desirability was equal to 0. This means that with the standard methods it is not possible, at least at this stage, to get a theoretical combination of values that satisfy all the CQAs, which is another evidence of the benefits of the proposed methodology and the consequent obtained models.

6.5 Uncontrolled Factors

After the identification of important effects from the controlled factors, the procedure (stepwise regression with back-enforcement of heredity, with logistic regression for binary responses and lognormal-GLM for continuous ones) was repeated extending the models to account for the uncontrolled factors. It was verified that it is obtained the exactly same models for all the responses, except to the binary CQA-D.2.

The results for this response are shown in table 6.17, where the uncontrolled factor

which was selected is represented by the coded name UF.

Table 6.17: Selected models for CQA-D.2 with uncontrolled factors.

Selection Method	Effects type	Selected effects	AICc value	Gener. R^2
SR BEH LR AICc	1st order	FP-3.4, FP-4.1, PP-2.1, PP-3.1, UF	33.44	0.67
	2nd order	PP-2.2, PP-3.1, FP-4.1, PP-2.2*FP-4.1, PP-3.1*UF	25.12	0.94

From table 6.17, it is possible to see that a given non-controlled independent variable was selected in both first and second-order models (being statistically significant), and it even participates in one selected interaction. Comparing the evaluation measures with the ones obtained only with the controlled factors, table 6.6, it is verified a decrease in the AICc value and an increase in the generalized R^2.

These observations suggest that there is a possibility that this variable may have a significant impact on the response and may be necessary to turn it in a controllable factor. In fact, these considerations were presented to the Bluepharma experimenters and it was a very important information to them, who, based on it, found a new path to the variability control between different batches and to achieve the target result more consistently.

6.6 Classification of Binary Responses

Three classifiers were considered to build binary prediction models: simple logistic regression (LR), logistic ridge regression (LRR) and support vector machine (SVM) with a linear kernel. They were modeled from the selected effects in section 6.2; the uncontrolled variable detected as possibly significant in response CQA-D.2 was not considered.

The results of the 500 times repeated 10-fold-stratified cross-validation are shown in tables 6.18, for response CQA-D.1, and 6.19, for response CQA-D.2. The values correspond to the mean ± standard deviation of the 500 replications.

For the LRR and SVM approaches, the presented values are the best ones from the models created varying the penalization hyper-parameter (λ in LRR and C in SVM) through the logarithmic grid search, and taking the accuracy values as the selection metric. All the grid search results are represented in Appendices.

Table 6.18: Results of the best classifiers, for response CQA-D.1.

Classifier	Penalization Parameter	Accuracy (%)	Sensitivity (%)	Specificity (%)
LR	-	83.17 ± 4.07	77.20 ± 7.17	90.22 ± 3.31
LRR	$\lambda = 0.03125$	92.87 ± 3.98	95.58 ± 5.72	89.98 ± 3.33
SVM	C = 16	**94.26 ± 3.32**	98.14 ± 3.98	89.94 ± 3.49

Table 6.19: Results of the best classifiers, for response CQA-D.2.

Classifier	Penalization Parameter	Accuracy (%)	Sensitivity (%)	Specificity (%)
LR	-	82.30 ± 3.98	83.64 ± 4.17	80.96 ± 6.72
LRR	$\lambda = 1$	**86.80 ± 3.08**	82.82 ± 4.63	90.78 ± 3.79
SVM	C = 0.125	80.50 ± 4.87	77.34 ± 3.98	83.66 ± 6.28

Even though the cross-validation may provide an underestimated evaluation of the models prediction ability, the results show that the classifiers have a very good performance, especially taking into consideration the small number of samples used to train them. It also supports the idea that the selected effects are indeed good predictors.

The consideration of the accuracy, sensitivity and specificity metrics is also relevant because they allow to analyze the binary models using measures that are much more common than AICc or generalized R^2. The performance of the classifiers suggests that they can be used to predict pretty well the response values of the binary outcomes in future trials, i.e. to know if a given combination of factors values is likely to achieve the target, mainly the ones that present best results: SVM for response CQA-D.1 and LRR for response CQA-D.2.

6.7 Validation Results

As it was mentioned before, data about some validation trials was also available and it was considered to further evaluate the selected methods, as an external validation source. However, those trials contain some factors with values outside of the design of experiment (DoE) ranges, which may be troublesome because the behavior of the process was not evaluated in those new intervals and some kind of extrapolation of the trained models will be required, which is very likely to lead to misleading results.

Therefore, only the trials whose all the selected factors for a given response have values inside the DoE ranges will be discussed. Besides, it is considered that the not selected factors are not only unimportant in the DoE ranges but also in general, so

they have no influence in a given response and they can be neglected.

For the binary responses, all the trials had values inside the DoE ranges for the selected factors of the response CQA-D.1. The real outcome of those trials was the positive class, which was the class predicted by the built classifiers.

For the continuous responses, one of the validation trials had values inside the DoE ranges for the selected factors of two responses, CQA-C.1 and CQA-C.2. The information about the observed (real) and the predicted values (mean and its 95% prediction interval) is shown in table 6.20. As a reference, it is represented in table 6.21, the values predicted for the same trials using the models which were obtained by the pharma company experimenters using the standard pharmaceutical methodology.

Table 6.20: Observed and predicted values for a validation trial, considering the proposed models.

Response	**Observed value**	**Predicted with GLM**	**Predicted with log-transformation**	
		Mean	Mean	Interval
CQA-C.1	40.31	23.39	17.47	[2.63 , 115.90]
CQA-C.3	1.78	0.75	0.68	[0.19 , 2.44]

Table 6.21: Observed and predicted values for a validation trial, considering the models obtained using the standard approach.

Response	**Observed value**	**Predicted with OLS**	
		Mean	Interval
CQA-C.1	40.31	-54.38	[-478.27 , 369.50]
CQA-C.3	1.78	-10.06	[-56.97 , 36.68]

Comparing the results of the tables 6.20 and 6.21, it is possible to observe big differences between the models obtained using the methodology proposed in this work and the ones obtained using the standard one. The first consideration is related to the predicted mean values and the observed ones. While the proposed models have means close to the real values, the standard methods' models obtain a value far away from the observed one. In fact, for those models, the predicted value is negative, which is physically impossible. This is another advantage of the lognormal procedure, it is implicit that only positive values may occur.

The second consideration is about the predicted intervals (PI). It is easily visible that the PIs are very much shorter in the proposed models. As expected, the PIs

are larger for response CQA-C.1 than response CQA-C.3 as a result of the larger interval and scale of values in DoE trials.

These observations confirm that the proposed models are more suitable than the ones obtained before in Bluepharma from the use of standard techniques.

About the remaining validation trials, i.e. the trials with values outside of the DoE ranges, it was verified that even some observed values were inside the prediction intervals, other ones were not. However, this was already expected because the extrapolation of the trained models to new ranges is not a correct approach. If the experimenters want to extend the ranges, then a new planned experiment must be performed to retrieve the appropriate information. In fact, this is one of the reasons why changes outside the design space require a new regulatory approval.

In order to complete the analysis of these results, two additional notes are addressed, both related to the results of table 6.20. The first one is that the prediction intervals are not symmetrical, which is a consequence of the skewed nature of lognormal distribution and the exponential computing of PI after back-transformation. The second one is that the predicted mean obtained with the log-transformation is smaller than the one obtained with lognormal-GLM, which is due to the fact that the arithmetic mean (calculated on non-transformed data) is more sensitive to the individual large values than the geometric mean (computed after data transformation) [109, 110].

7

Conclusions

This work provides an extensive analysis of the available data of a real-world problem being currently developed by a pharmaceutical company. The main focus at this stage of the development process was to get the factors that most impact each one of the responses being considered, i.e. the critical formulation and process parameters. A new methodology to obtain such models was proposed and it was shown to achieve better results than the standard techniques, which had been implemented before.

The methodology suggested in this study not only consider interactions but it can also be applied to both discrete and continuous responses, and it is even possible to use it with right-skewed distributions that may occur and do not respect the ordinary least squares assumptions. Particulary, it was considered a continuous response distribution which is much more suitable to this problem than the normal one, and it was verified the importance of consider the interactions.

In fact, some of the factors that are included in the second-order models, which are the ones who provide a better explanation of the responses, were not selected in the first-order models, and so their importance would be missed if we did not look to interactions as well. Besides, this is an advantage of consider a back-enforcement of heredity instead of a fore-enforcement one.

Furthermore, the procedure conforms to the effect principles, incorporating the heredity one in a simpler but very interpretable and efficient way when compared to the literature approaches that explicitly consider that principle: it requires fewer parameters and assumptions than the Bayesian models, it is easier to implement and interpret than penalized least squares methods and it is more complete than the forward selection approach of Hamada and Wu [55].

For the 5 responses being considered, it was obtained a total of 18 main effects, 2 strong heredity interactions, 5 weak heredity interactions and 1 interaction without heredity (hierarchy and heredity principles), with a mean of 5.2 active effects per

response (sparsity principle).

Those results agree with what was theoretically expected and make the models more interpretable too. Although it is not presented in this document, it was verified that if no heredity was imposed in the stepwise regression procedure, the selected models would be larger and composed mostly by non-hereditary interactions, and so the data-driven modeling results would diverge from what happens in reality.

Other analyses were also performed, including the study of uncontrolled factors whose results led the experimenters to find new useful paths to meet more consistently the defined objectives, and the creation of predictive models that can be used to guide future planned experiments. The investigation of external validation trials supports the idea that the models proposed in this work are better and more precise than the ones obtained using traditional methods.

Finally, the multi-objective optimization using desirability functions showed that with the suggested models, it is possible to obtain a combination of factor values that in theory ensures that all the targets are simultaneously achieved, something that was not possible with the standard approach models, being a *proof of concept* of the proposed methodology.

Some limitations must be highlighted as well, like the small sample size and the lack of replicates in the design of experiments trials. Although the absence of such constraints would probably lead to better models, it is important to note that those are intrinsic characteristics of screening experiments, and the goal is precisely to get the maximum knowledge possible with such small number of trials, due to resources constraints. Taking this into consideration, the approaches used in this work provide arguably great progress and the results give confidence that the final objectives of the project can be achieved and that it can be done more efficiently.

7.1 Future Work

The problem being studied in this work is related to the identification of important effects in screening experiments and the indication of guidelines for the next steps of the process development. Therefore, it is expected that these analyses can be used to better advise the follow-up experiments where more data will be collected to be then employed in optimization procedures.

The increase of experimental data can also be considered for the improvement of some performed approaches, such as the predictive models, and for the execution

of new procedures that are not possible to accomplish with the amount of data available at this stage, such as the creation of an inline process control.

Regarding the identification of the important effects, the selection procedure can be enhanced to include physical and chemical first principles, i.e. the known theoretically relations between the factors. The incorporation of effect heredity is already a way to introduce prior and physical knowledge to the data-driven modeling but it states only the structure relations and not the semantics ones.

Appendices

A

Selection Paths

A.1 Effects Selection

A.1.1 Response CQA-D.1

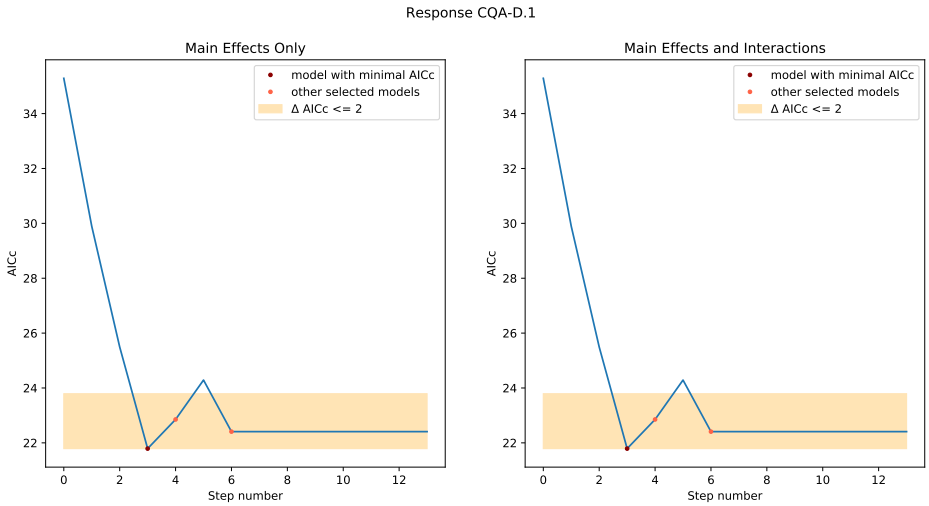


Figure A.1: Stepwise regression selection path, for response CQA-D.1.

Table A.1: All the selected models for response CQA-D.1.

Effects type	Selected effects	AICc	Generalized R^2
only main effects	FP-1.1, PP-2.2, FP-3.5	21.79	0.789
	FP-1.1, PP-2.2, FP-3.5, FP-2.1, FP-5.1	22.41	0.914
	FP-1.1, PP-2.2, FP-3.5, FP-5.1	22.85	0.805
main effects and interactions	FP-1.1, PP-2.2, FP-3.5	21.79	0.789
	FP-1.1, PP-2.2, FP-3.5, FP-2.1, FP-5.1	22.41	0.914
	FP-1.1, PP-2.2, FP-3.5, FP-2.1, FP-5.1	22.85	0.805

A.1.2 Response CQA-D.2

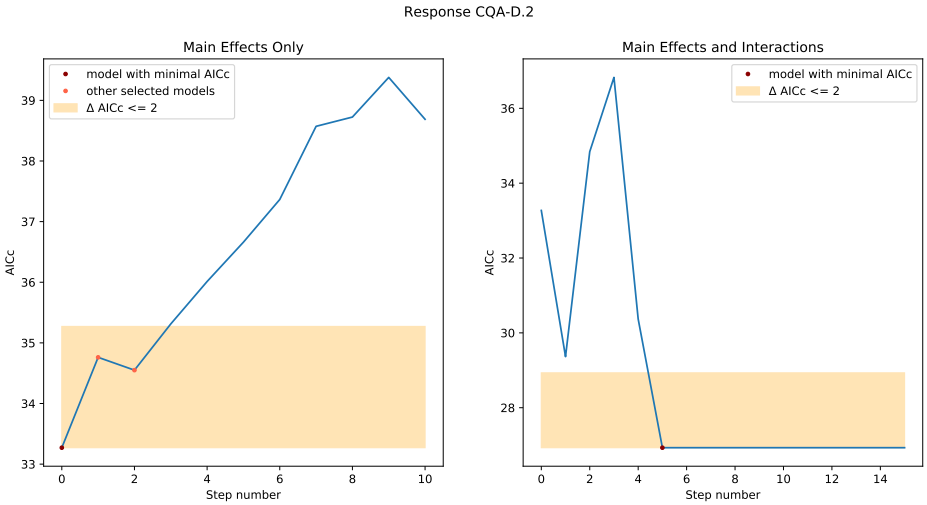


Figure A.2: Stepwise regression selection path, for response CQA-D.2.

Table A.2: All the selected models for response CQA-D.2.

Effects type	Selected effects	AICc	Generalized R^2
only main effects	-	33.27	-
	FP-3.4, FP-4.1	34.55	0.292
	FP-4.1	34.76	0.161
main effects and interactions	PP-2.2, PP-3.1, FP-4.1, PP-2.2*PP-3.1, PP-2.2*FP-4,1	26,93	0.828

A.1.3 Response CQA-C.1

A.1.3.1 Generalized Linear Model - Lognormal

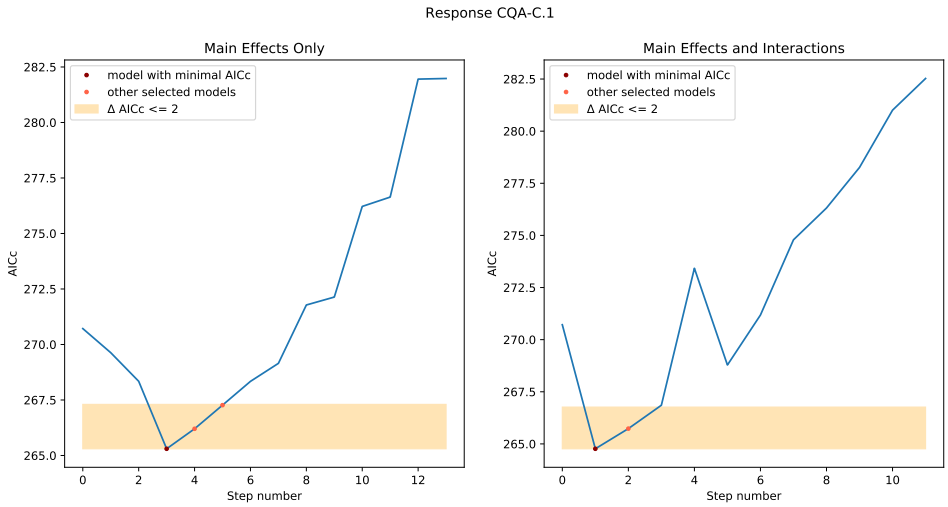


Figure A.3: Stepwise regression selection path, for response CQA-C.1, using lognormal-GLM.

Table A.3: All the selected models for response CQA-C.1, obtained using lognormal-GLM.

Effects type	Selected effects	AICc	Generalized R^2
only main effects	FP-1.3, PP-3.1, FP-4.1	265.30	0.446
	FP-1.3, PP-3.1, FP-4.1, FP-1.2	266.20	0.505
	FP-1.3, PP-3.1, FP-4.1, FP-3.4	267.27	0.483
main effects and interactions	FP-1.3, PP-5.1, FP-1.3*PP-5.1	264.77	0.458
	FP-1.3, PP-2.2, FP-3.4, PP-5.1, FP-1.3*PP-5.1, PP-2.2*FP-3.4	265.73	0.662

A.1.3.2 Ordinary Least Squares

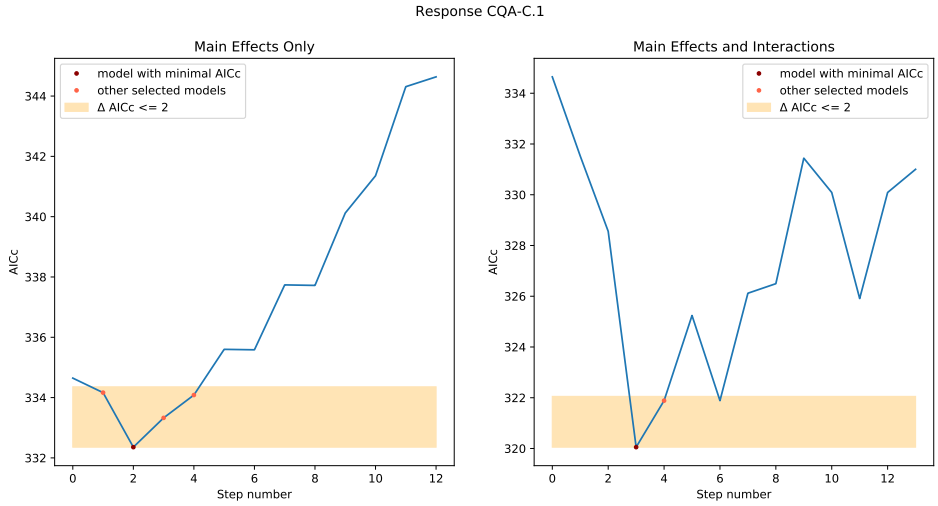


Figure A.4: Stepwise regression selection path, for response CQA-C.1, using OLS.

Table A.4: All the selected models for response CQA-C.1, obtained using OLS.

Effects type	Selected effects	AICc	R^2
only main effects	FP-1.3, FP-4.1	332.36	0.278
	FP-1.3, FP-4.1, FP-3.3	333.33	0.343
	FP-1.3, FP-4.1, PP-3.1	334.08	0.322
	FP-1.3	334.16	0.122
main effects and interactions	FP-1.3, PP-2.2, PP-5.1, FP-1.3*PP-5.1, PP-1.3*PP-2.1, PP-2.1*PP-5.1	320.05	0.773
	FP-1.3, PP-2.2, FP-3.4, PP-5.1, FP-1.3*PP-5.1, PP-1.3*PP-2.1, PP-2.1*PP-5.1, FP-3.4*PP-5.1	321.89	0.847

A.1.4 Response CQA-C.2

A.1.4.1 Generalized Linear Model - Lognormal

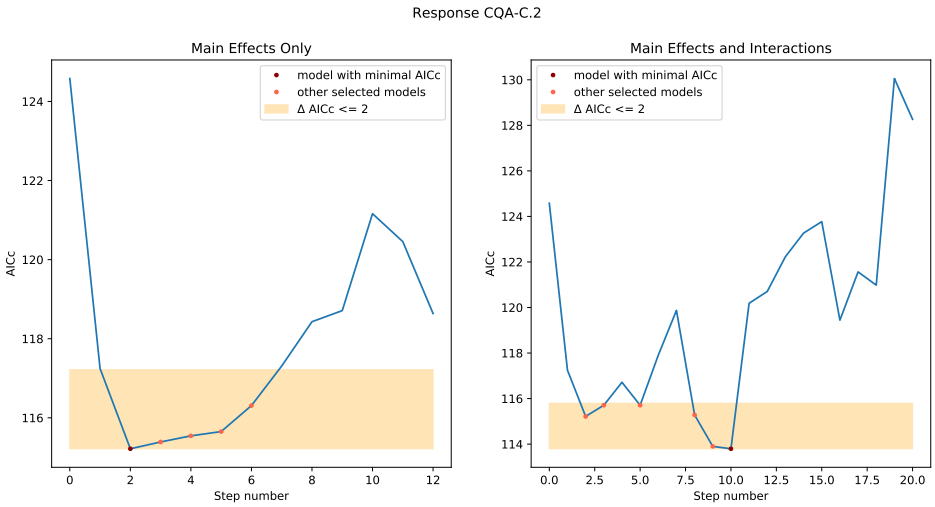


Figure A.5: Stepwise regression selection path, for response CQA-C.2, using lognormal-GLM.

Table A.5: All the selected models for response CQA-C.2, obtained using lognormal-GLM.

Effects type	Selected effects	AICc	Generalized R^2
only main effects	FP-3.3, FP-4.1	115.22	0.462
	FP-3.3, FP-4.1, PP-2.1	115.39	0.527
	PP-2.1, FP-3.3, PP-3.1	115.54	0.524
	PP-2.1, FP-3.3, PP-2.1, PP-3.1	115.65	0.588
	PP-2.1, FP-3.3, PP-2.1, PP-3.1, FP-1.3	116.31	0.643
main effects and interactions	FP-1.3, PP-2.1, FP-3.3, FP-3.5, FP-4.1, PP-5.2, FP-3.3*PP-5.2	113.80	0.787
	FP-1.3, PP-2.1, FP-3.3, FP-4.1, PP-5.2, FP-3.3*PP-5.2	113.90	0.733
	FP-3.3, FP-4.1	115.22	0.462
	FP-1.3, FP-3.3, FP-4.1, PP-5.2, FP-3.3*PP-5.2	115.28	0.658
	FP-3.3, FP-4.1, PP-5.2, FP-3.3*PP-5.2	115.71	0.587

A.1.4.2 Ordinary Least Squares

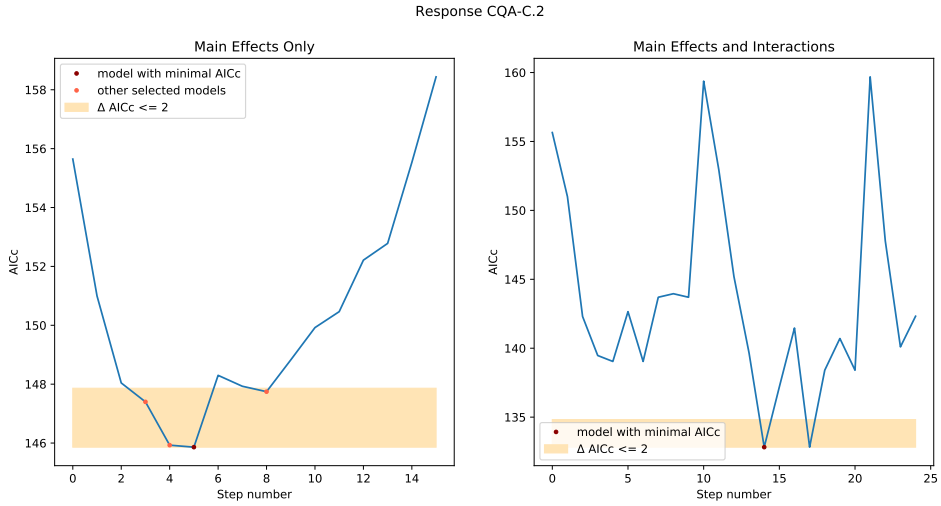


Figure A.6: Stepwise regression selection path, for response CQA-C.2, using OLS.

Table A.6: All the selected models for response CQA-C.2, obtained using OLS.

Effects type	Selected effects	AICc	R^2
only main effects	FP-1.1, PP-2.1, FP-3.3, FP-3.5, FP-4.1	145.86	0.665
	PP-2.1, FP-3.3, FP-3.5, FP-4.1	145.92	0.602
	PP-2.1, FP-3.3, FP-4.1	147.40	0.508
	FP-1.1, PP-2.1, FP-3.3, FP-3.5, FP-4.1, FP-5.1	147.74	0.700
main effects and interactions	FP-1.1, PP-2.1, FP-3.5, FP-4.1, FP-1.1*FP-3.5, FP-1.1*FP-3.5, FP-2.1*FP-3.5, FP-2.1*FP-4.1, FP-3.5*FP-4.1	132.82	0.925

A.1.5 Response CQA-C.3

A.1.5.1 Generalized Linear Model - Lognormal

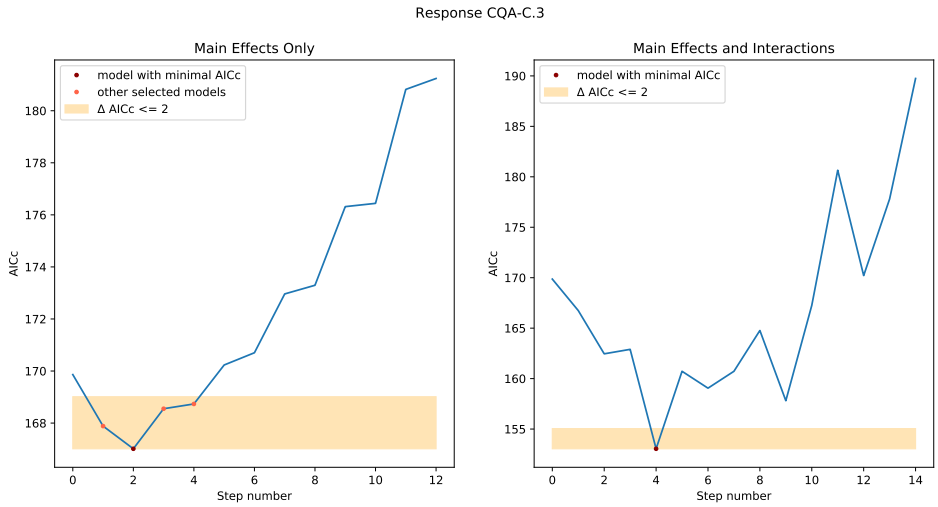


Figure A.7: Stepwise regression selection path, for response CQA-C.3, using lognormal-GLM.

Table A.7: All the selected models for response CQA-C.3, obtained using lognormal-GLM.

Effects type	Selected effects	AICc	Generalized R^2
only main effects	PP-3.1, FP-4.1	167.01	0.295
	PP-3.1	167.88	0.175
	PP-3.1, FP-4.1, PP-5.2	168.55	0.343
	PP-3.1, FP-4.1, FP-1.1	168.73	0.338
main effects and interactions	PP-2.1, PP-2.2, PP-3.2, FP-3.4, PP-5.1, PP-2.2*PP-3.2, PP-2.2*FP-3.4, PP-2.1*PP-2.2	153.05	0.871

A.1.5.2 Ordinary Least Squares

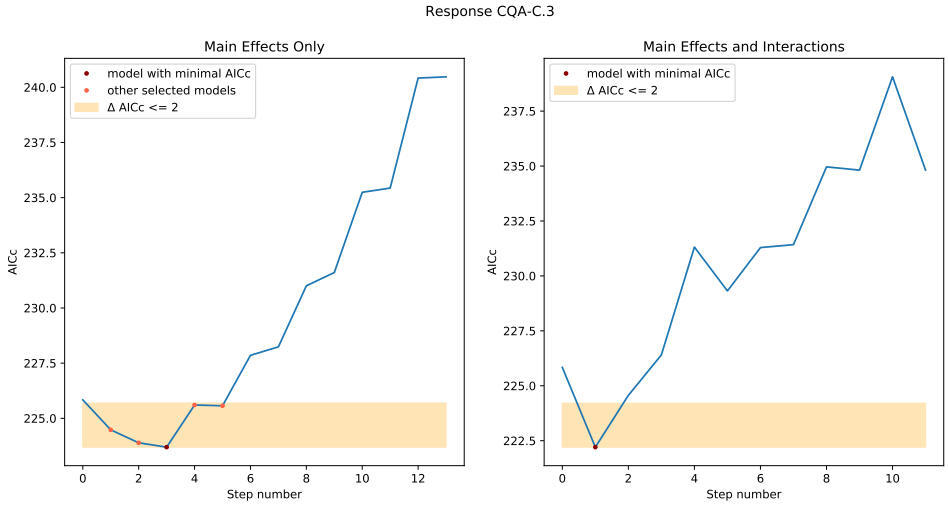


Figure A.8: Stepwise regression selection path, for response CQA-C.3, using OLS.

Table A.8: All the selected models for response CQA-C.3, obtained using OLS.

Effects type	Selected effects	AICc	R^2
only main effects	PP-3.1, FP-4.1	167.01	0.295
	PP-3.1	167.88	0.175
	PP-3.1, FP-4.1, PP-5.2	168.55	0.343
	PP-3.1, FP-4.1, FP-1.1	168.73	0.338
main effects and interactions	PP-2.1, PP-2.2, PP-3.2, FP-3.4, PP-5.1, PP-2.2*PP-3.2, PP-2.2*FP-3.4, PP-2.1*PP-2.2	153.05	0.871

A.2 Classifiers Grid Search

A.2.1 Response CQA-D.1

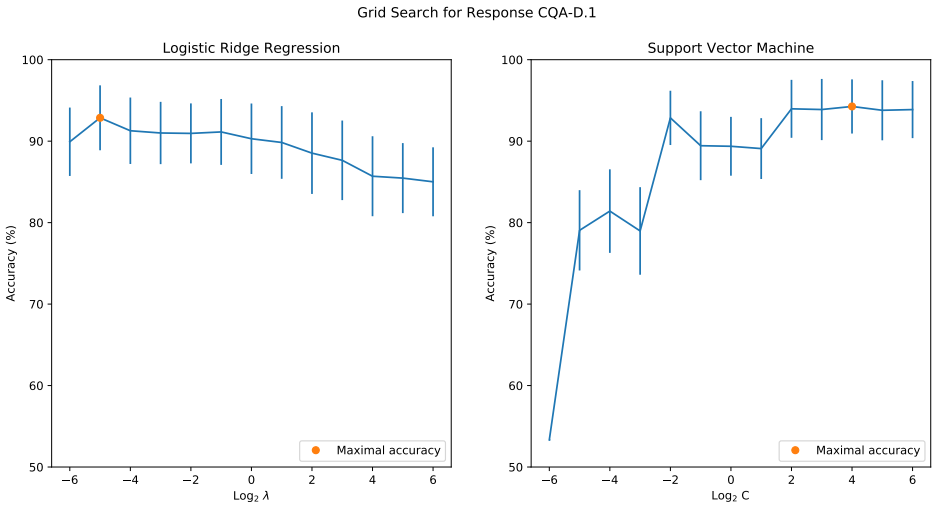


Figure A.9: Grid search optimization for penalization parameter of classifiers, for response CQA-D.1.

A.2.2 Response CQA-D.2

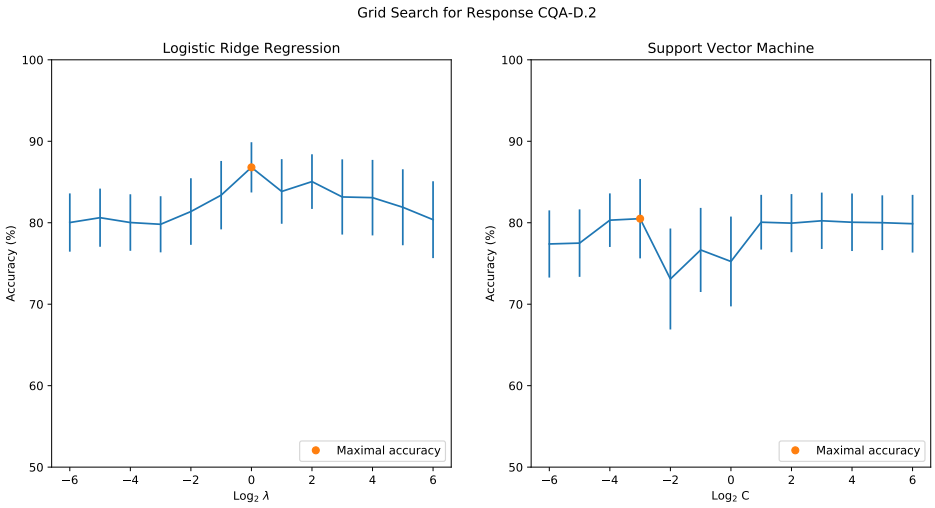


Figure A.10: Grid search optimization for penalization parameter of classifiers, for response CQA-D.2.

References

[1] U.S. Food & Drug Administration, "Pharmaceutical CGMPs for the 21s Century - A risk-based approach", 2004.

[2] U.S. Food & Drug Administration, "Guidance for Industry PAT - A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance", 2004.

[3] U.S. Food & Drug Administration, "Guidance for Industry - Process Validation: General Principles and Practices", 2011.

[4] International Council for Harmonisation, "Pharmaceutical Quality System Q10", 2008.

[5] International Council for Harmonisation, "Quality Risk Management Q9", 2005.

[6] International Council for Harmonisation, "Pharmaceutical Development Q8(R2)", 2009.

[7] M. J. Juran, *Juran on quality by design: the new steps for planning quality into goods and services.* Free Press, 1992.

[8] S. Roy, "Quality by design : A holistic concept of building quality in pharmaceuticals", *International Journal of Pharmaceutical and Biomedical Research*, vol. 3, no. 2, pp. 100–108, 2012.

[9] U.S. Food & Drug Administration, "Guidance for Industry and Review Staff - Target Product Profile - A Strategic Development Process Tool", tech. rep., 2007.

[10] L. X. Yu, G. Amidon, M. A. Khan, S. W. Hoag, J. Polli, G. K. Raju, and J. Woodcock, "Understanding Pharmaceutical Quality by Design", *The AAPS Journal*, vol. 16, pp. 771–783, jul 2014.

[11] B. Cooney, S. D. Jones, and H. L. Levine, "Quality by design for monoclonal antibodies, Part 1: Establishing the foundations for process development", *BioProcess International*, vol. 14, no. 6, 2016.

[12] U.S. Food & Drug Administration, "Approved Drug Products with Terapeutic Equivalence Evaluations", 2019.

[13] U.S. Food & Drug Administration, "GDUFA II Commitment Letter", 2016.

[14] C.-F. Wu and M. Hamada, *Experiments : planning, analysis, and optimization*. Wiley, 2009.

[15] X. Li, N. Sudarsanam, and D. D. Frey, "Regularities in data from factorial experiments", *Complexity*, vol. 11, pp. 32–45, may 2006.

[16] Douglas C. Montgomery, *Montgomery: Design and Analysis of Experiments*. 2000.

[17] J. Gabrielsson, N.-O. Lindberg, and T. Lundstedt, "Multivariate methods in pharmaceutical applications", *Journal of Chemometrics*, vol. 16, pp. 141–160, mar 2002.

[18] L. M. Collins, J. J. Dziak, and R. Li, *Design of Experiments With Multiple Independent Variables: A Resource Management Perspective on Complete and Reduced Factorial Designs*, vol. 14. NIH Public Access, sep 2009.

[19] J. Wass, "First Steps in Experimental Design–The Screening Experiment", *Journal of Validation Technology*, pp. 49–57, 2010.

[20] M. S. Elazazy, "Factorial Design and Machine Learning Strategies: Impacts on Pharmaceutical Analysis", in *Spectroscopic Analyses - Developments and Applications*, InTech, 2017.

[21] M. Uy and J. K. Telford, "Optimization by Design of Experiment Techniques", in *2009 IEEE Aerospace conference*, pp. 1–10, IEEE, mar 2009.

[22] P. Goos and B. Jones, *Optimal Design of Experiments: A Case Study Approach*. Wiley, 2011.

[23] J. Antoy, *Design of Experiments for Engineers and Scientists*. Elsevier, 2014.

[24] B. Bergquist, E. Vanhatalo, and M. L. Nordenvaad, "A Bayesian Analysis of Unreplicated Two-Level Factorials Using Effects Sparsity, Hierarchy, and Heredity", *Quality Engineering*, vol. 23, pp. 152–166, mar 2011.

[25] F. Zimmer, A. Souza, A. Silveira, M. Santos, M. Matsushita, N. Souza, and A. Rodrigues, "Application of Factorial Design for Optimization of the Synthesis of Lactulose Obtained from Whey Permeate", *Journal of the Brazilian Chemical Society*, vol. 28, no. 12, pp. 2326–2333, 2017.

[26] W. Satianrangsarith, "Design of Experiments Approach for Improving Wire Bonding Quality", *International Journal of Innovation, Management and Technology*, 2012.

[27] M. W. Hester and J. M. Usher, "Factor screening experiments using fractional factorial split plot designs and regression analysis in developing a top-down nanomanufacturing system for recycling of welding rod residuals", *Production and Manufacturing Research*, vol. 5, no. 1, pp. 118–139, 2017.

[28] N. Tiwari, U. Bellur, S. Sarkar, and M. Indrawan, "Identification of critical parameters for MapReduce energy efficiency using statistical Design of Experiments", in *2016 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pp. 1170–1179, IEEE, may 2016.

[29] M. Hamada and N. Balakrishnan, "Analyzing unreplicated factorial experiments: A review with some new proposals", *Statistica Sinica*, vol. 8, no. 1, pp. 1–41, 1998.

[30] C. Daniel, "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments", *Technometrics*, vol. 1, pp. 311–341, nov 1959.

[31] A. G. Holms and J. N. Berrettoni, "Chain-Pooling ANOVA for Two-Level Factorial Replication-Free Experiments", *Technometrics*, vol. 11, pp. 725–746, nov 1969.

[32] D. A. Zahn, "Modifications of and revised critical values for the half-normal plot", *Technometrics*, vol. 17, no. 2, pp. 189–200, 1975.

[33] A. Seheult and J. W. Tukey, "Some resistant procedures for analyzing factorial experiments", *Utilitas Mathematica*, vol. 21B, 1982.

[34] G. E. Box and R. D. Meyer, "An Analysis for Unreplicated Fractional Factorials", *Technometrics*, vol. 28, pp. 11–18, feb 1986.

[35] E. G. Johnson and J. W. Tukey, "Graphic Exploratory Analysis of Variance Illustrated on a Splitting of the Johnson and Tsao Data", p. 73, 1987.

[36] D. T. Voss, "Generalized modulus-ratio tests for analysis of factorial designs with zero degrees of freedom for error", *Communications in Statistics - Theory and Methods*, vol. 17, pp. 3345–3359, jan 1988.

[37] H. C. Benski, "Use of a Normality Test to Identify Significant Effects in Factorial Designs", *Journal of Quality Technology*, vol. 21, pp. 174–178, jul 1989.

[38] R. V. Lenth, "Quick and Easy Analysis of Unreplicated Factorials", *Technometrics*, vol. 31, p. 469, nov 1989.

[39] A. F. Bissell, "Interpreting mean squares in saturated fractional designs", *Journal of Applied Statistics*, vol. 16, pp. 7–18, jan 1989.

[40] K. N. Berk and R. R. Picard, "Significance Tests for Saturated Orthogonal Arrays", *Journal of Quality Technology*, vol. 23, pp. 79–89, apr 1991.

[41] A. F. Bissell, "Mean squares in saturated fractional designs revisited", *Journal of Applied Statistics*, vol. 19, pp. 351–366, jan 1992.

[42] N. D. Le and R. H. Zamar, "A global test for effects in 2k factorial design without replicates", *Journal of Statistical Computation and Simulation*, vol. 41, pp. 41–54, may 1992.

[43] J. Juan and D. Pefña, "A simple method to identify significant effects in unreplicated two-level factorial designs", *Communications in Statistics - Theory and Methods*, vol. 21, pp. 1383–1403, jan 1992.

[44] W.-Y. Loh, "Identification of active contrasts in unreplicated factorial experiments", *Computational Statistics & Data Analysis*, vol. 14, pp. 135–148, aug 1992.

[45] F. Dong, "On the Identification of Active Contrasts in Unreplicated Fractional Factorials", 1993.

[46] H. Schneider, W. J. Kasperski, and L. Weissfeld, "Finding Significant Effects for Unreplicated Fractional Factorials Using the n Smallest Contrasts", *Journal of Quality Technology*, vol. 25, pp. 18–27, jan 1993.

[47] J. H. Venter and S. J. Steel, "A Hypothesis-Testing Approach Toward Identifying Active Contrasts", *Technometrics*, vol. 38, pp. 161–169, may 1996.

[48] M. J. Anderson and P. J. Whitcomb, *DOE Simplified*. Productivity Press, aug 2017.

[49] D. Gunasegaram, D. Farnsworth, and T. Nguyen, "Identification of critical factors affecting shrinkage porosity in permanent mold casting using numerical simulations based on design of experiments", *Journal of Materials Processing Technology*, vol. 209, pp. 1209–1219, feb 2009.

[50] H. Chipman, "Bayesian variable selection with related predictors", *Canadian Journal of Statistics*, vol. 24, pp. 17–36, mar 1996.

[51] E. I. George and R. E. McCulloch, "Variable Selection via Gibbs Sampling", *Journal of the American Statistical Association*, vol. 88, pp. 881–889, sep 1993.

[52] H. Chipman, M. Hamada, and C. F. Wu, "A Bayesian Variable-Selection Approach for Analyzing Designed Experiments With Complex Aliasing", *Technometrics*, vol. 39, pp. 372–381, nov 1997.

[53] S. D. Beattie, D. K. H. Fong, and D. K. J. Lin, "A Two-Stage Bayesian Model Selection Strategy for Supersaturated Designs", *Technometrics*, vol. 44, pp. 55–63, feb 2002.

[54] J. O. Berger and L. R. Pericchi, "The Intrinsic Bayes Factor for Model Selection and Prediction", *Journal of the American Statistical Association*, vol. 91, pp. 109–122, mar 1996.

[55] M. Hamada and C. F. J. Wu, "Analysis of Designed Experiments with Complex Aliasing", *Journal of Quality Technology*, vol. 24, pp. 130–137, jul 1992.

[56] D. K. J. Lin, "A New Class of Supersaturated Designs", *Technometrics*, vol. 35, pp. 28–31, feb 1993.

[57] P. H. Westfall, S. S. Young, and D. K. J. Lin, "Forward selection error control in the analysis of supersaturated designs", *Statistica Sinica*, vol. 8, no. 1, pp. 101–117, 1998.

[58] B. Abraham, H. Chipman, and K. Vijayan, "Some Risks in the Construction and Analysis of Supersaturated Designs", *Technometrics*, vol. 41, pp. 135–141, may 1999.

[59] S. C. Pinault, "An analysis of subset regression for orthogonal designs", *American Statistician*, vol. 42, no. 4, pp. 275–277, 1988.

[60] X. Lu and X. Wu, "A Strategy of Searching Active Factors in Supersaturated Screening Experiments", *Journal of Quality Technology*, vol. 36, pp. 392–399, oct 2004.

[61] D. Xing, H. Wan, M. Y. Zhu, S. M. Sanchez, and T. Kaymal, "Simulation screening experiments using Lasso-optimal supersaturated design and analysis: A maritime operations application", in *2013 Winter Simulations Conference (WSC)*, pp. 497–508, IEEE, dec 2013.

[62] Bahr Kadhim Mohammed, "Robust Lasso Variable Selection for Factorial Experiments Analysis with Application", *International Journal of Statistics and Applications*, vol. 8, no. 2, pp. 79–87, 2018.

[63] M. Yuan, V. R. Joseph, and Y. Lin, "An Efficient Variable Selection Approach for Analyzing Designed Experiments", *Technometrics*, vol. 49, pp. 430–439, nov 2007.

[64] R. Tibshirani, I. Johnstone, T. Hastie, and B. Efron, "Least angle regression", *The Annals of Statistics*, vol. 32, pp. 407–499, apr 2004.

[65] N. H. Choi, W. Li, and J. Zhu, "Variable Selection With the Strong Heredity Constraint and Its Oracle Property", *Journal of the American Statistical Association*, vol. 105, pp. 354–364, mar 2010.

[66] H. Noguchi, Y. Ojima, and S. Yasui, "A practical variable selection for linear models", in *Frontiers in Statistical Quality Control 10*, (Heidelberg), pp. 349–360, Physica-Verlag HD, 2012.

[67] D.-H. Jang and C. M. Anderson-Cook, "Examining robustness of model selection with half-normal and LASSO plots for unreplicated factorial designs", *Quality and Reliability Engineering International*, vol. 33, pp. 1921–1928, dec 2017.

[68] R. Li and D. K. Lin, "Data analysis in supersaturated designs", *Statistics & Probability Letters*, vol. 59, pp. 135–144, sep 2002.

[69] J. Fan and R. Li, "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties", *Journal of the American Statistical Association*, vol. 96, pp. 1348–1360, dec 2001.

[70] R. Li and D. K. J. Lin, "Variable Selection for Screening Experiments.", *Quality technology & quantitative management*, vol. 6, no. 3, pp. 271–280, 2009.

[71] F. K. Phoa, Y.-H. Pan, and H. Xu, "Analysis of supersaturated designs via the Dantzig selector", *Journal of Statistical Planning and Inference*, vol. 139, pp. 2362–2372, jul 2009.

[72] E. Candes and T. Tao, "The Dantzig selector: Statistical estimation when p is much larger than n", *The Annals of Statistics*, vol. 35, pp. 2313–2351, dec 2007.

[73] K. Drosou and C. Koukouvinos, "A new variable selection method based on SVM for analyzing supersaturated designs", *Journal of Quality Technology*, vol. 51, pp. 21–36, jan 2019.

[74] Q.-Z. Zhang, R.-C. Zhang, and M.-Q. Liu, "A method for screening active effects in supersaturated designs", *Journal of Statistical Planning and Inference*, vol. 137, pp. 2068–2079, jun 2007.

[75] M. A. Wolters and D. Bingham, "Simulated Annealing Model Search for Subset Selection in Screening Experiments", *Technometrics*, vol. 53, pp. 225–237, aug 2011.

[76] N. Balakrishnan, C. Koukouvinos, and C. Parpoula, "An information theoretical algorithm for analyzing supersaturated designs for a binary response", *Metrika*, vol. 76, pp. 1–18, jan 2013.

[77] K. Drosou, C. Koukouvinos, and A. Lappa, "Screening Active Effects in Supersaturated Designs with Binary Response via Control Charts", *Quality and Reliability Engineering International*, vol. 33, pp. 1475–1483, nov 2017.

[78] F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information", *Journal of Machine Learning Research*, vol. 5, pp. 1531–1555, 2004.

[79] Hanchuan Peng, Fuhui Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1226–1238, aug 2005.

[80] B. D. Rege, J. Gawel, and J. H. Kou, "Identification of critical process variables for coating actives onto tablets via statistically designed experiments", *International Journal of Pharmaceutics*, vol. 237, pp. 87–94, apr 2002.

[81] T. Zahel, L. Marschall, S. Abad, E. Vasilieva, D. Maurer, E. Mueller, P. Murphy, T. Natschläger, C. Brocard, D. Reinisch, P. Sagmeister, and C. Herwig, "Workflow for Criticality Assessment Applied in Biopharmaceutical Process Validation Stage 1", *Bioengineering*, vol. 4, no. 4, p. 85, 2017.

[82] SAS Institute Inc., "JMP® 14 Fitting Linear Models", 2018.

[83] K. P. Anderson and D. A. Burnham, *Model Selection and Multi-Model Inference : A Practical Information-Theoretic Approach*, vol. 172. Springer, 2002.

[84] A. Pickles, *An introduction to likelihood analysis*. Geo Books, 1985.

[85] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection - A review and recommendations for the practicing statistician", *Biometrical Journal*, vol. 60, pp. 431–449, may 2018.

[86] H. Akaike, "Information Theory and an Extension of the Maximum Likelihood Principle", pp. 199–213, Springer, New York, NY, 1998.

[87] N. Sugiura, "Further analysts of the data by akaike' s information criterion and the finite corrections", *Communications in Statistics - Theory and Methods*, vol. 7, pp. 13–26, jan 1978.

[88] P. V. Bertrand, Y. Sakamoto, M. Ishiguro, and G. Kitagawa, "Akaike Information Criterion Statistics.", *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, vol. 151, no. 3, p. 567, 1988.

[89] C. M. Hurvisch and C.-L. Tsai, "Regression and time series model selection in small samples", *Biometrika*, vol. 76, pp. 297–307, jun 1989.

[90] Y. Dodge, *The concise encyclopedia of statistics*. Springer, 2008.

[91] H. C. Thode, *Testing for normality*, vol. 34. Marcel Dekker, 2002.

[92] S. S. Shapiro and M. B. Wilk, "An Analysis of Variance Test for Normality (Complete Samples)", *Biometrika*, vol. 52, p. 591, dec 1965.

[93] R. D'Agostino, "An omnibus test of normality for moderate and large size samples", *Biometrika*, vol. 58, pp. 341–348, aug 1971.

[94] R. D'Agostino and E. S. Pearson, "Tests for Departure from Normality. Empirical Results for the Distributions of b2 and b1", *Biometrika*, vol. 60, p. 613, dec 1973.

[95] A. Agresti, "Categorical data analysis", 1990.

[96] D. F. Polit, *Data analysis & statistics for nursing research*. Appleton & Lange, 1996.

[97] H. Kyngäs and M. Rissanen, "Support as a crucial predictor of good compliance of adolescents with a chronic disease", *Journal of Clinical Nursing*, vol. 10, pp. 767–774, jul 2008.

[98] J. A. Nelder and R. W. M. Wedderburn, "Generalized Linear Models", *Journal of the Royal Statistical Society. Series A (General)*, vol. 135, p. 370, may 1972.

[99] P. P. McCullagh and J. A. Nelder, *Generalized linear models*. Chapman and Hall, 1989.

[100] T. Hastie, R. Tibshirani, and J. Friedman, "Linear Methods for Regression", pp. 1–57, 2009.

[101] S. Raschka and V. Mirjalili, *Python Machine Learning Second Edition*. Packt Publishing, 2017.

[102] L. Breiman and P. Spector, "Submodel Selection and Evaluation in Regression. The X-Random Case", *International Statistical Review / Revue Internationale de Statistique*, vol. 60, p. 291, dec 1992.

[103] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection", *International Joint Conference of Artificial Intelligence*, 1995.

[104] N. J. D. Nagelkerke, "A note on a general definition of the coefficient of determination", *Biometrika*, vol. 78, pp. 691–692, sep 1991.

[105] E. C. Harrington, "The desirability function", *Industrial Quality Control*, vol. 21, no. 10, pp. 494–498, 1965.

[106] G. Derringer and R. Suich, "Simultaneous Optimization of Several Response Variables", *Journal of Quality Technology*, vol. 12, pp. 214–219, oct 1980.

[107] M. A. Udokang Anietie, E. Raji Surajudeen, and T. Bello Latifat Kemi, "An Empirical Study of Generalized Linear Model for Count Data", *Journal of Applied & Computational Mathematics*, vol. 04, no. 05, 2015.

[108] S. R. Naffees Gowsar, M. Radha, and N. Devi, "A Comparison of Generalized Linear Models for Insect Count Data", *International Journal of Statistics and Analysis*, vol. 9, no. 1, pp. 1–9, 2019.

[109] F.-W. Wellmer, "The Use of the Lognormal Distribution", in *Statistical Evaluations in Exploration for Mineral Deposits*, pp. 67–94, Berlin, Heidelberg: Springer Berlin Heidelberg, 1998.

[110] P. Armitage, G. G. Berry, and J. N. S. Matthews, *Statistical methods in medical research*. Blackwell Scientific Publications, 2002.