Functional Data Analysis and Nonlinear Regression Models: An Information Quality Perspective

Ron S. Kenett[1] and Chris Gotwalt[2]

[1] The KPA Group and the Samuel Neaman Institute, Technion, Israel
[2] JMP Division, SAS, Research Triangle, NC, USA

Abstract

Data from measurements over time can be analyzed in different ways. In this paper we compare functional data analysis and nonlinear regression models. As an example, we focus on dissolution profiles of drug tablets where tablets under test are compared to reference tablets. An initial simple example is used to introduce functional data analysis and non liner regression models. A more complex example is then provided where statistical designed mixture experiments are used to optimization tablet formulation in order to match the target dissolution profile. To evaluate these approaches we refer to the Information Quality (InfoQ) framework. InfoQ provides a checklist for evaluating the information quality derived from applying an analytic method such as functional data analysis or nonlinear regression. The JMP platform (www.jmp.com) is used to demonstrate the points made in the paper. A JMP add in enables to score the information quality of a specific study.

1. Introduction

When you collect data from measurements over time it can be analyzed, among other methods, with functional data analysis (FDA) or non-linear regression (NLR) methods. Examples include chromatograms from high-performance liquid chromatography (HPLC) systems, dissolution profiles of drug tablets, and sensor measurements in production systems. Both FDA and NLR can be used in analyzing such repeated measurements. FDA is non-parametric. In this paper we apply B-splines which one fits with a finite number of knots. In contrast NLR is parametric. Here we use, three parameter Gomperz and Weibull functions referring to the asymptote, growth rate and inflection point parameters. A comprehensive coverage of functional data analysis is provided in Ramsey and Silverman (2002). See also Woodall et al (2004) and Thakur et al (2021). For nonlinear regression applications, see Bates and Watts (2007) and Kenett and Zacks (2021).

The paper presents simple and complex applications of FDA and NLR to tablet dissolution profiles. In conclusion, FDA and NLR are discussed using the information quality (InfoQ) framework introduced in Kenett and Shmueli (2014). Moreover, the information quality perspective enables a comparison of analytic methods like FDA and NLR. An ensemble of models can enhance information quality. This was proposed in the context of customer satisfaction surveys data analysis models (Kenett and Salini, 2011). The next section is an introduction to information quality.

2. Introduction to Information Quality

Information quality (InfoQ), was introduced in Kenett and Shmueli (2014). InfoQ is a framework for planning, tracking and assessing information derived from data and data analysis. It is formally defined as the utility, U, of applying a particular statistical analysis, f, to a particular data set, X, conditioned on a given goal. Given the 4 components: g, U, f and X, the formal definition is InfoQ = U(f(X|g)).

Besides these 4 components, InfoQ involves 8 dimensions: Data Resolution, Data Structure, Data Integration, Temporal Relevance, Chronology of Data and Goal, Generalizability, Operationalization, and Communication. Further developments of InfoQ and application examples are provided in Kenett and Shmueli (2016), Reis and Kenett (2018) and Kenett and Gotwalt (2020).

The following checklist of questions is used in evaluating the InfoQ dimensions, with reference to g, U, f and X:
(1) *Data Resolution*: Is the data granularity adequate for the intended goal? Has measurement uncertainty been evaluated and found appropriate?
(2) *Data Structure*: Have you considered using data from different sources reflecting on the problem at hand?
(3) *Data Integration*: If relevant, how is the integration from different data sources done? Are there linkage issues that affect data privacy?
(4) *Temporal Relevance*: Does the time gap between data collection and analysis cause any concern?
(5) *Chronology of Data & Goal*: Are the analytic findings communicated to the right persons in a timely manner?
(6) *Generalizability*: Can you derive general conclusions based on the study, beyond what was explicitly studied, for example to other products or processes?
(7) *Operationalization*: Are the measured variables themselves of relevance to the study goal? Are stated action items derived from the study?
(8) *Communication*: Are findings properly communicated to the intended audience?

With this checklist, one can compare methods of analysis by assessing the generated level of information quality they provide. The last section in this paper provides such a comparison. The next two sections are introductions to FDA and NLR using tablet dissolution profiles as an example. Section 5 is a complex example where mixture experiments are conducted to match tablet dissolution profiles with a target profile. Section 6 concludes the paper with a discussion of FDA and NLR implementations in terms of information quality.

3. Introduction to Functional Data Analysis

Functional data analysis (FDA) is about modeling data profiles with functions. FDA often uses splines. A spline is a continuous function which coincides with a polynomial on every subinterval of the whole interval on which it is defined. In other words, splines are functions which are piecewise polynomial. The coefficients of the polynomials differ from interval to interval, but the order of the polynomial is fixed. Splines originated in description of soil properties in agricultural plots and in the design of ship hulls.

A basis spline (or B-spline) is a spline function that has minimal support with respect to a given degree, smoothness, and domain partition. A spline function of given degree can be expressed as a linear combination of B-splines of that degree. The term was coined by Shoenberg (1958) who has been recognized as the “father of splines”. B-splines of order n, are basis functions for spline functions of the same order defined over the same knots. All possible spline functions can be built from a linear combination of B-splines, and there is only one unique combination for each spline function (Karlin and Pinkus, 1976).

A spline of order n is defined as a piecewise polynomial function of degree n-1, in a variable x. The values of x, where the pieces of polynomial meet, are known as knots. These are denoted as t_0 , $t_{1,}$ $t_{2\ldots}$ t_n and sorted into nondecreasing order. When the knots are distinct, the first n-2 derivatives of the polynomial pieces are continuous across each knot. When r knots are coincident, then only the first n-r-1 derivatives of the spline are continuous across that knot.

For a given sequence of knots, there is, up to a scaling factor, a unique spline B{i,n}(x) satisfying

$$B_{i,n}(x) = \begin{cases} 0 & \text{if} \quad x < t_i \quad \text{or} \quad x \geq t_{i+n} \\ \text{nonzero} & \text{otherwise} \end{cases}$$

If we add the additional constraint $\sum_i B_{i,n}(x) = 1$ for all x between the first and last knot, then the scaling factor of $B_{i,n}(x)$ becomes fixed. The resulting $B_{i,n}(x)$ spline functions are called B-splines. The higher order B-splines are defined by recursion. The usefulness of B-splines lies in the fact that any spline function of order n, on a given set of knots, can be expressed as a linear combination of B-splines.

The term P-spline stands for "penalized B-spline". It refers to using the B-spline representation where the coefficients are determined partly by the data to be fitted, and partly by an additional penalty function that imposes smoothness to avoid overfitting.

We present next an example of a quadratic B-spline application. The data used in the case study consists of 12 test and reference tablets measured under dissolution conditions at 5, 10, 15, 20, 30 and 45 minutes. The level of dissolution recorded at these time instances is the basis for the dissolution functions to be analysed. The test tablets behaviour is compared to the target reference paths. Ideally the tested generic product is identical to the target brand reference. Figure 1 shows the reference and tested paths for the 12 tablets tested in each group, with a superimposed smoother.

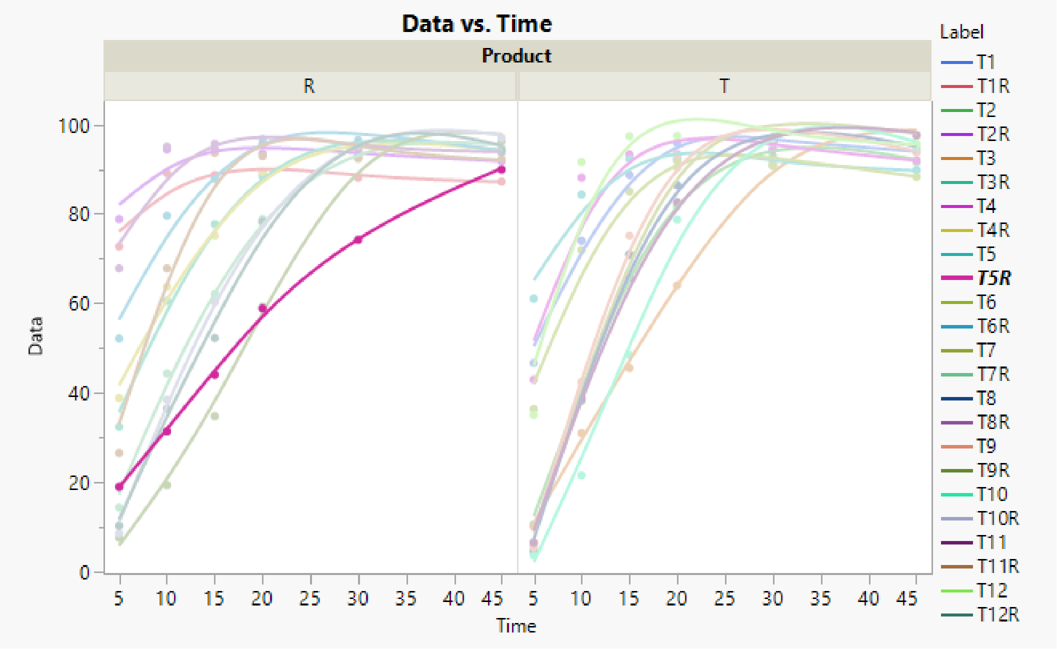


Figure 1: Dissolution paths of reference and tested tablets, with smoother (T5R is highlighted)

Functional data analysis extends the capabilities of traditional statistical techniques by considering functions tracking change over time, or space, or some other dimension. Because we are observing curves rather than individual values, the vector-valued observations $X_1, \ldots, X_n$ are replaced by the univariate functions $X_1(t), \ldots, X_n(t)$, where t is a continuous index varying within a closed interval [0, T]. In functional principal component analysis (FPCA), each sample curve is considered to be an independent realization of a univariate stochastic process X(t) with smooth mean function E{X(t)} = μ(t) and covariance function cov{X(s),X(t)} = σ(s, t). A spectral decomposition of the covariance function expresses σ as an orthogonal expansion (in the L_2 sense) in terms of its eigenvalues λj and associated eigenfunctions $V_j(t)$, so that

$$\sigma(s,t) = \sum_{j=1}^{\infty} \lambda_j V_j(s) V_j(t),$$

where the eigenvalues quickly tend to zero and the first few eigenfunctions are slowly varying. The covariance function, σ, is positive-definite and hence, the eigenvalues are nonnegative and can be ordered: λ1 ≥ λ2 ≥· · · ≥ 0. The goal is to determine the primary components of functional variation in σ(s, t), where the eigenvalues indicate the amount of total variance attributed to each component. A random curve can then be expressed as

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j V_j(t),$$

where the coefficient

$$\xi_j = \int [X(t) - \mu(t)] V_j(t) dt$$

is a scalar random variable called the j^th^ FPC score with $E\{\xi_j\} = 0$, $\mathrm{var}\{\xi_j\} = \lambda_j$, $\Sigma \lambda_j < \infty$, and $\mathrm{cov}\{\xi_j, \xi_k\} = 0$, j≠k.

The eigenfunctions $\{V_j(t)\}$, called FPC functions, satisfy:

$$\int [V_j(t)]^2 dt = 1, \quad \int V_j(t)V_k(t)dt = 0, j \neq k,$$

where the integrals are taken over [0, T], which may be periodic. This expansion is known as the Karhunen–Loeve expansion of X(t). Thus, X(t) − μ(t) may be thought of as a finite sum of orthogonal curves each having uncorrelated random amplitudes.

In analysing functional data, on first smooths each individual sample curve (e.g., using spline methods or local-linear smoothers), and then apply functional PCA assuming that the smooth curves are the completely observed curves. This gives a set of eigenvalues $\{\lambda_j\}$ and (smooth) eigenfunctions $\{V_j(t)\}$ extracted from the sample covariance matrix of the smoothed data. Typically, the first and second estimated eigenfunctions exhibit location of individual curve variation. Figure 2 presents outputs from JMP functional data explorer (FDE) applied to the dissolution data displayed in Figure 1. (https://www.jmp.com/support/help/en/16.1/#page/jmp/model-reports-2.shtml?os=win&source=application&utm_source=helpmenu&utm_medium=application#ww328146)

Figure 2 is showing the fit of a quadratic B-splines to the dissolution data to the set of reference tablets. One observes the unusual straight-line path of T5R which was highlighted in Figure 1. In Figure 3, we display a scatterplot of the top two functional principal components of the reference paths. Here T5R clearly stands out. Could be that the dissolution at 30 minutes was misreported as too low. A double check of the record indicated this was not the case.

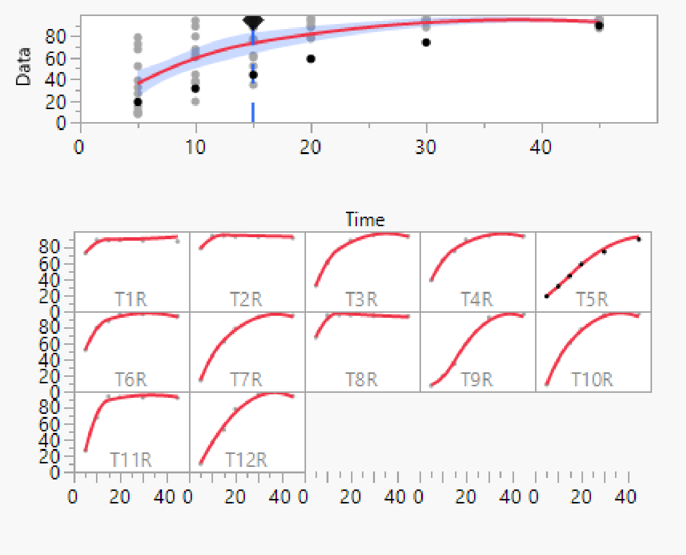


Figure 2: Functional form of reference tablets dissolution paths

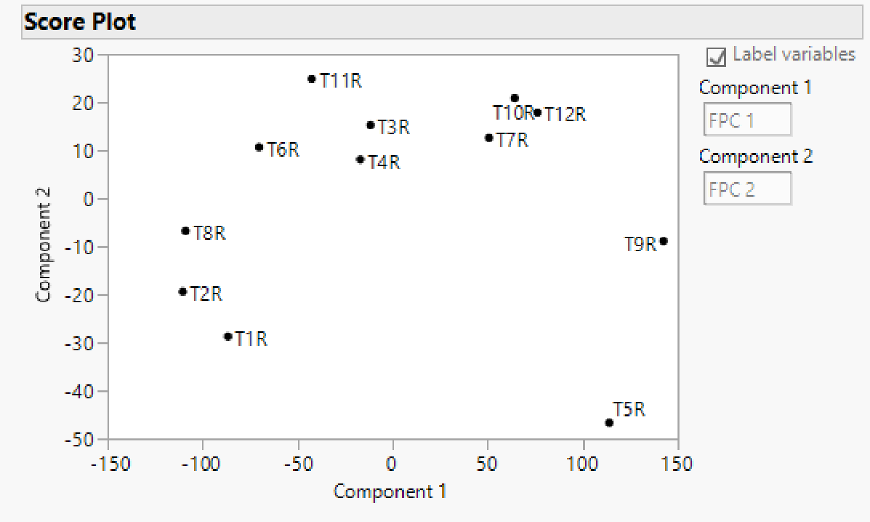


Figure 3: Scatter plot of top two functional principal components of reference tablet dissolution paths

FDA can be used to identify anomalies in profiles. A graphical comparison of the mean functional data of the reference and test tablets dissolution paths is shown in Figure 4. Except for the initial dissolution phase, they almost fully overlap indicating that, on average, the tablets under test are compatible with the refence product, in terms of dissolution.

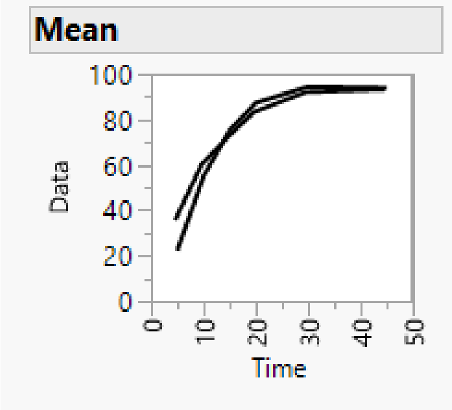


Figure 4: Mean functional data of reference tablets and tested tablets dissolution paths.

4. Non Linear Regression Models

The general linear regression model is formulated as:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdot \cdot \cdot + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

Linear regression models, include first-order models in p – 1 predictor variables and also more complex models. For example, a polynomial regression model in one or more predictor variables is linear in the parameters. Such as a model includes two predictor variables with linear, quadratic, and interaction terms:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X^2_{i1} + \beta_3 X_{i2} + \beta_4 X^2_{i2} + \beta_5 X_{i1} X_{i2} + \varepsilon_i$$

Models with transformed variables that are linear in the parameters, also belong to the class of linear regression models. An example is the following model:

$$\log_{10} Y_i = \beta_0 + \beta_1 \sqrt{X_{i1}} + \beta_2 \exp(X_{i2}) + \varepsilon_i$$

In general, a linear regression model has the form:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\beta}) + \varepsilon_i$$

where X_i is the vector of the observations on the predictor variables for the i^{th} case:

$$\mathbf{X}_i = \begin{bmatrix} 1 \\ X_{i1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$$

β is the vector of the regression coefficients. f(X_i , β) represents the expected value E{Y_i }, which for linear regression models is equal to:

$$f(\mathbf{X}_i, \boldsymbol{\beta}) = \mathbf{X}_i' \boldsymbol{\beta}$$

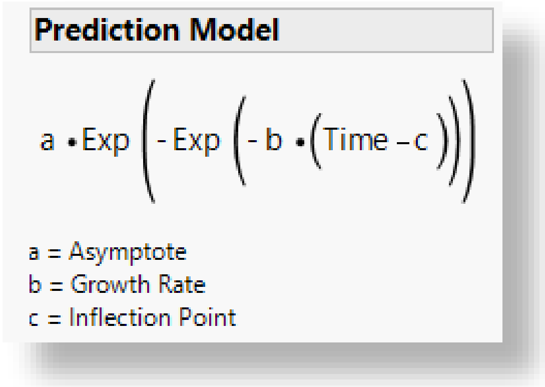
Nonlinear regression models (NLR) are of the same form as linear regression models: Y_i = f (X_i, γ) + ε_i. An observation Y_i is the sum of a mean response f (X_i , γ), given by the nonlinear response function f (X, γ) and the error term ε_i. The error terms are typically assumed to have expectation zero, constant variance, and to be uncorrelated, just as for linear regression models. Often, a normal error model is invoked assuming that the error terms are independent normal random variables with constant variance. The parameter vector in the response function f (X, γ) is denoted by γ rather than β in the linear model. This emphasizes that the response function is nonlinear in the parameters. A difference between linear and nonlinear regression models is that the number of regression parameters is not necessarily directly related to the number of X variables in the model. In linear regression models, if there are p – 1, X variables in the model, then there are p regression coefficients in the model. If the number of X variables in the nonlinear regression model is denoted by q, and we continue to denote the number of regression parameters in the response function by p. The general form of a nonlinear regression model is expressed as:

$$Y_i = f(\mathbf{X}_i, \boldsymbol{\gamma}) + \varepsilon_i$$

where

$$\underset{q\times 1}{\mathbf{X}_i} = \begin{bmatrix} X_{i1} \\ X_{i2} \\ \vdots \\ X_{iq} \end{bmatrix} \qquad \underset{p\times 1}{\boldsymbol{\gamma}} = \begin{bmatrix} \gamma_0 \\ \gamma_1 \\ \vdots \\ \gamma_{p-1} \end{bmatrix}$$

Like in linear regression models, estimation of parameters of a nonlinear regression model can be carried out by the method of least squares or the method of maximum likelihood. Like in linear regression, both methods of estimation yield the same parameter estimates when the error terms are independent normal with constant variance. Unlike linear regression, it is usually not possible to find analytical expressions for the least squares and maximum likelihood estimators for nonlinear regression models. Instead, numerical search procedures are used with both estimation procedures. For example, The Gauss-Newton method, a.k.a. as the linearization method, uses a Taylor series expansion to approximate the nonlinear regression model with linear terms and then employs ordinary least squares to estimate the parameters. Iteration of these steps generally leads to a solution to the nonlinear regression problem. We fit a Gompertz 3 parameter non linear model to the data analysed in section 4. The functional form of the model is shown in Figure 5.



Prediction Model

$$a \cdot \text{Exp}\left(-\text{Exp}\left(-b \cdot \left(\text{Time} - c\right)\right)\right)$$

a = Asymptote
b = Growth Rate
c = Inflection Point

Figure 5: Gompertz 3 parameter non linear model

Figure 6 presents the fitted model to the individual dissolution curves. Comparing Figure 6 to Figure 2, derived with quadratic B-splines, provides similar qualitative information. The three Gompertz parameters for the reference tablets are presented in Figure 7. Tablet T5R stands out because of low growth rate (0.07) and high inflection point (11.5).

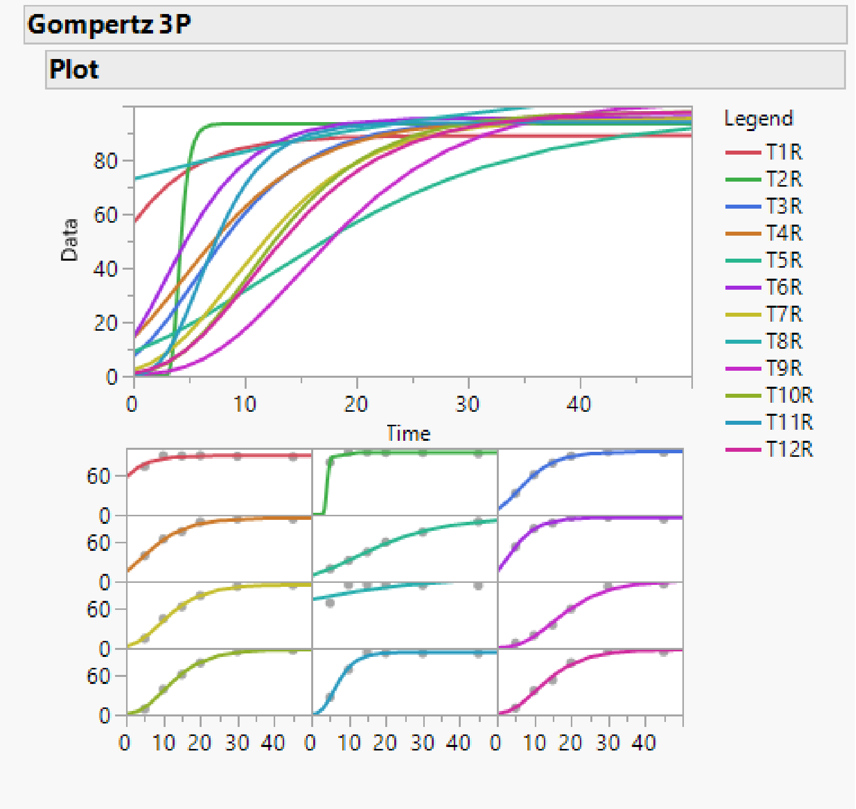
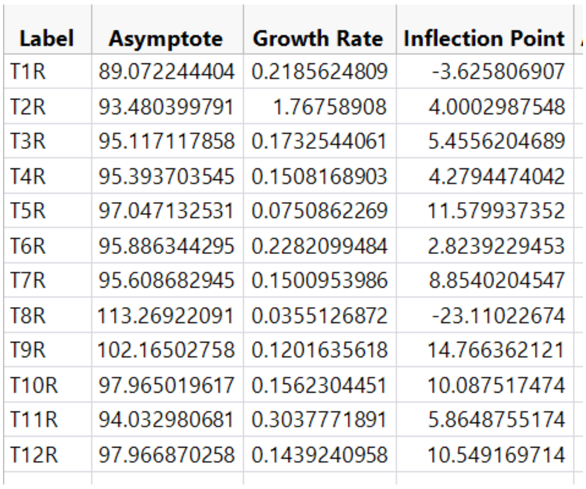


Figure 6: Gompertz 3 parameter non-linear model fit to reference tablets dissolution paths



Label	Asymptote	Growth Rate	Inflection Point
T1R	89.072244404	0.2185624809	-3.625806907
T2R	93.480399791	1.76758908	4.0002987548
T3R	95.117117858	0.1732544061	5.4556204689
T4R	95.393703545	0.1508168903	4.2794474042
T5R	97.047132531	0.0750862269	11.579937352
T6R	95.886344295	0.2282099484	2.8239229453
T7R	95.608682945	0.1500953986	8.8540204547
T8R	113.26922091	0.0355126872	-23.11022674
T9R	102.16502758	0.1201635618	14.766362121
T10R	97.965019617	0.1562304451	10.087517474
T11R	94.032980681	0.3037771891	5.8648755174
T12R	97.966870258	0.1439240958	10.549169714

Figure 7: The 3 Gompertz parameters from fit to reference tablets dissolution paths

Sections 3 and 4 were designed to introduce FDA and NLR in an application to tablet dissolution curves. In the next section we present a complex case study derived from tablet formulations designed with mixture experiments to match a target dissolution profile.

5. Optimizing tablet dissolution profiles with mixture experiments

The goal, *g*, of this case study is to find polymer amounts and compression values leading matching a target reference dissolution curve. The matching gap is the utility function, *U*, used in this study. The data, *X*, consists of 102 tablets tested over 4 dissolution times. The tablets were produced in 16 formulations involving two polymers (A and B) and a tableting compression force. Six tablets were produced for each formulation. Dissolution profiles of six tablets of a reference product were used to set a target profile. The analysis, *f*, involves FDA and NRL. To conduct an information quality assessment we first describe the FDA and NLR analysis. The InfoQ assessment is presented in Section 6.

Fitting a quadratic B-spline to the 16 formulation experiments with 6 replicates each produces the fit shown in Figure 8.

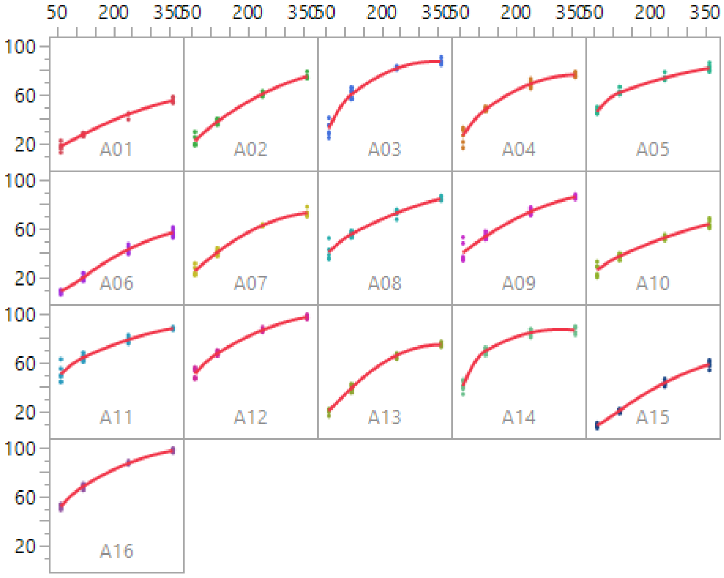
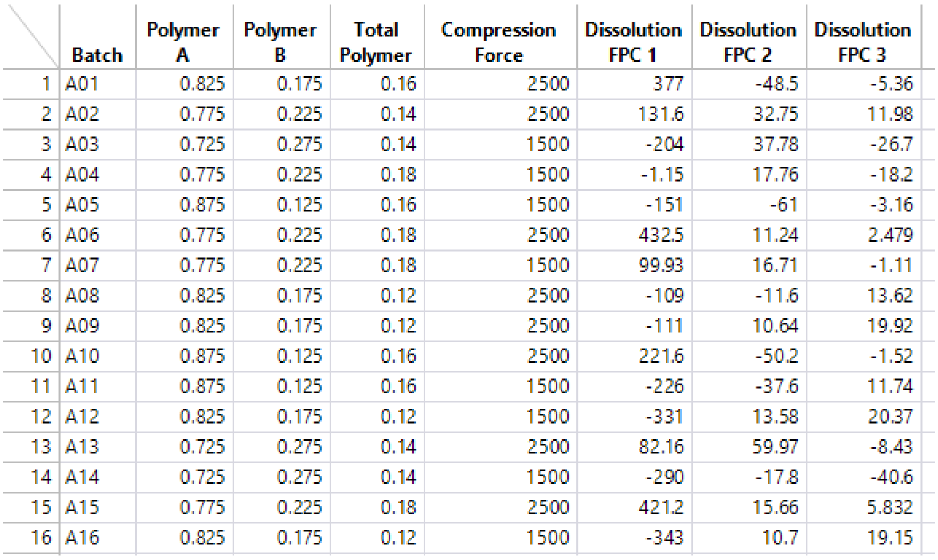


Figure 8: Quadratic B-spline fit to the 16 formulation tablet dissolution paths

These 16 paths correspond to 16 mixture experiments representing mixtures of Polymer A and Polymer B, the total amount of polymer and the compression force. The experimental array with the first three functional principal components (FPC) is shown in Figure 9.



	Batch	Polymer A	Polymer B	Total Polymer	Compression Force	Dissolution FPC 1	Dissolution FPC 2	Dissolution FPC 3
1	A01	0.825	0.175	0.16	2500	377	-48.5	-5.36
2	A02	0.775	0.225	0.14	2500	131.6	32.75	11.98
3	A03	0.725	0.275	0.14	1500	-204	37.78	-26.7
4	A04	0.775	0.225	0.18	1500	-1.15	17.76	-18.2
5	A05	0.875	0.125	0.16	1500	-151	-61	-3.16
6	A06	0.775	0.225	0.18	2500	432.5	11.24	2.479
7	A07	0.775	0.225	0.18	1500	99.93	16.71	-1.11
8	A08	0.825	0.175	0.12	2500	-109	-11.6	13.62
9	A09	0.825	0.175	0.12	2500	-111	10.64	19.92
10	A10	0.875	0.125	0.16	2500	221.6	-50.2	-1.52
11	A11	0.875	0.125	0.16	1500	-226	-37.6	11.74
12	A12	0.825	0.175	0.12	1500	-331	13.58	20.37
13	A13	0.725	0.275	0.14	2500	82.16	59.97	-8.43
14	A14	0.725	0.275	0.14	1500	-290	-17.8	-40.6
15	A15	0.775	0.225	0.18	2500	421.2	15.66	5.832
16	A16	0.825	0.175	0.12	1500	-343	10.7	19.15

Figure 9: Experimental array and functional principal components of Quadratic B-spline fit to the 16 formulation tablet dissolution paths

The mean and three eigenvalues corresponding to FPC1, FPC2 and FPC3 are shown in Figure 10.

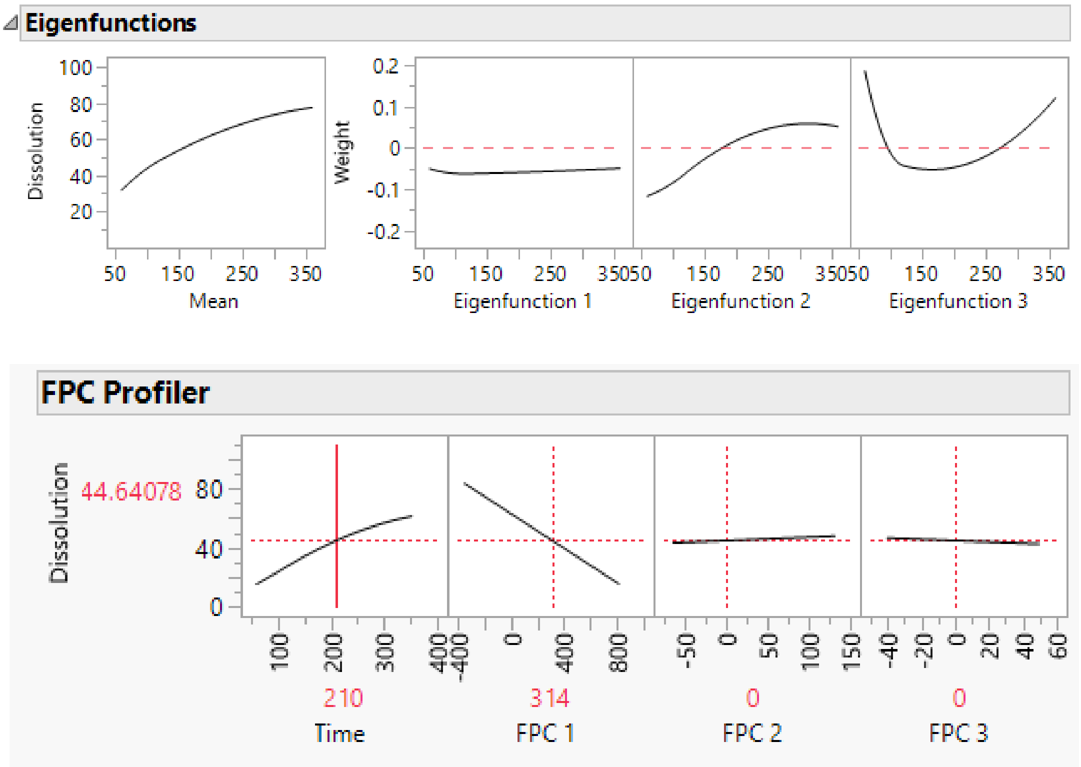
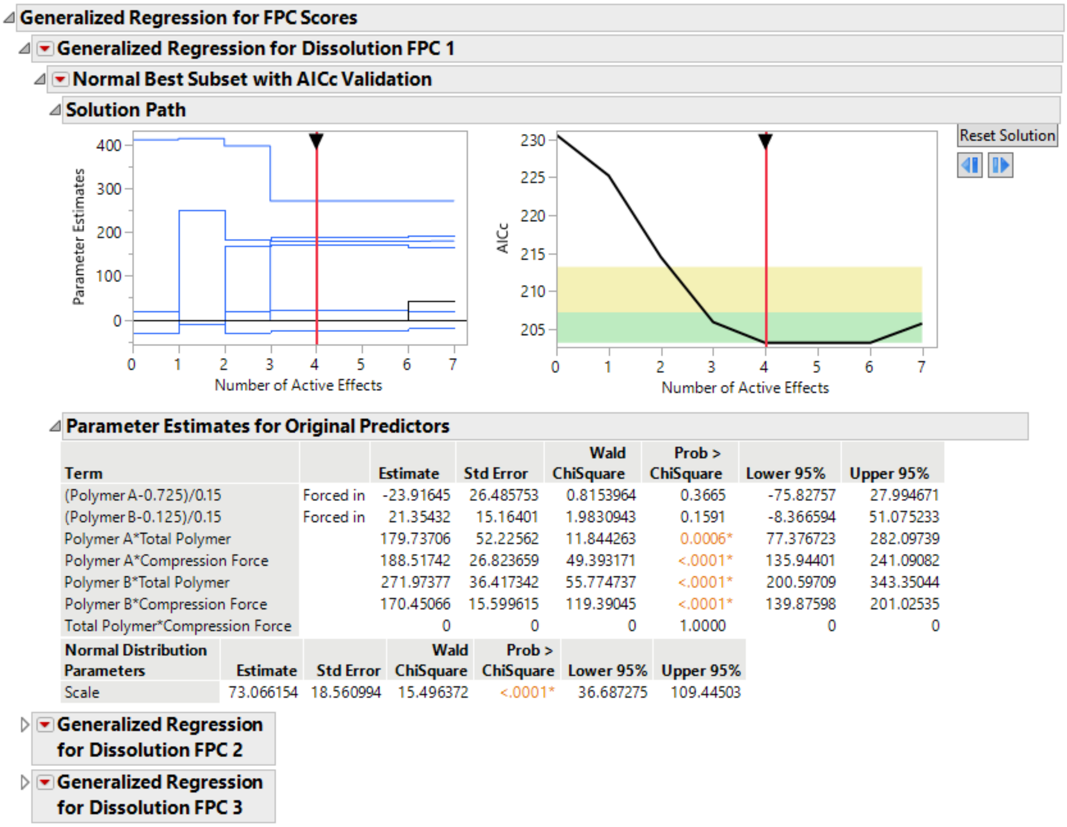


Figure 10: Eigenvalues (top) and profiler (bottom) of first three functional principal components of Quadratic B-spline

The first eigenfunction appears as affecting dissolution level and the second eigenfunction being related to dissolution quadratic shapes. See Section 3 for more details on FPC and eigenfunctions.

Generalized regression models the impact of the four experimental factors on the three functional principal components: FPC1, FPC2 and FPC3. Figure 11 presents the fit of a best subset regression to FPC1. The interactions of Polymer A and B with Total Polymer and Compression Force are significant on FPC1 that affects the dissolution level.



Generalized Regression for FPC Scores

Generalized Regression for Dissolution FPC 1

Normal Best Subset with AICc Validation

Solution Path

Parameter Estimates for Original Predictors

Term		Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
(Polymer A-0.725)/0.15	Forced in	-23.91645	26.485753	0.8153964	0.3665	-75.82757	27.994671
(Polymer B-0.125)/0.15	Forced in	21.35432	15.16401	1.9830943	0.1591	-8.366594	51.075233
Polymer A*Total Polymer		179.73706	52.22562	11.844263	0.0006*	77.376723	282.09739
Polymer A*Compression Force		188.51742	26.823659	49.393171	<.0001*	135.94401	241.09082
Polymer B*Total Polymer		271.97377	36.417342	55.774737	<.0001*	200.59709	343.35044
Polymer B*Compression Force		170.45066	15.599615	119.39045	<.0001*	139.87598	201.02535
Total Polymer*Compression Force		0	0	0	1.0000	0	0

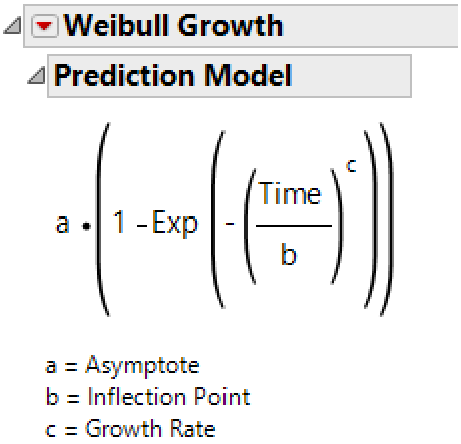
Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	73.066154	18.560994	15.496372	<.0001*	36.687275	109.44503

Generalized Regression for Dissolution FPC 2

Generalized Regression for Dissolution FPC 3

Figure 11: Best subset regression of first functional principal component versus main effect and interaction features.

Following the application of FDA we apply NLR to the same data. The Weibull growth non linear model used is based on domain specific knowledge of dissolution curves, see Langenbucher, F. (1972). Figure 12 presents the Weibull function. Figure 13 is an output is fitting the Weibull model to the 16 formulation experiments and the reference data.



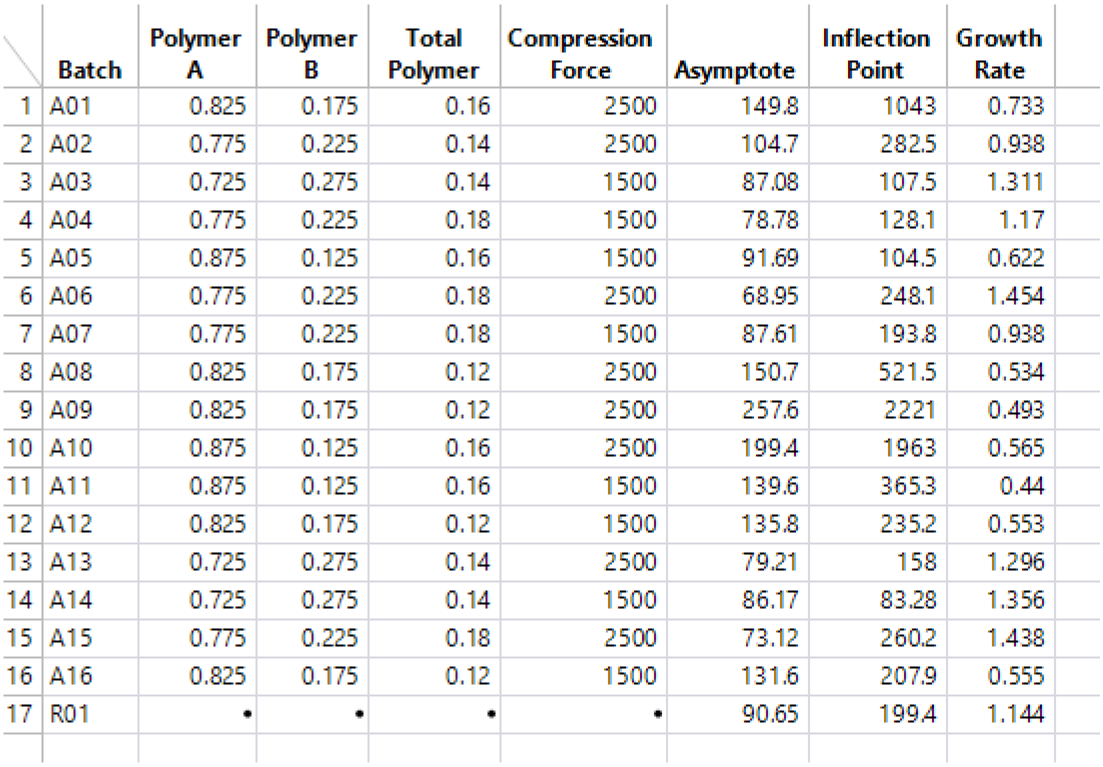
Weibull Growth

Prediction Model

$$a \cdot \left(1 - \mathrm{Exp}\left(-\left(\frac{\mathrm{Time}}{b}\right)^{c}\right)\right)$$

a = Asymptote
b = Inflection Point
c = Growth Rate

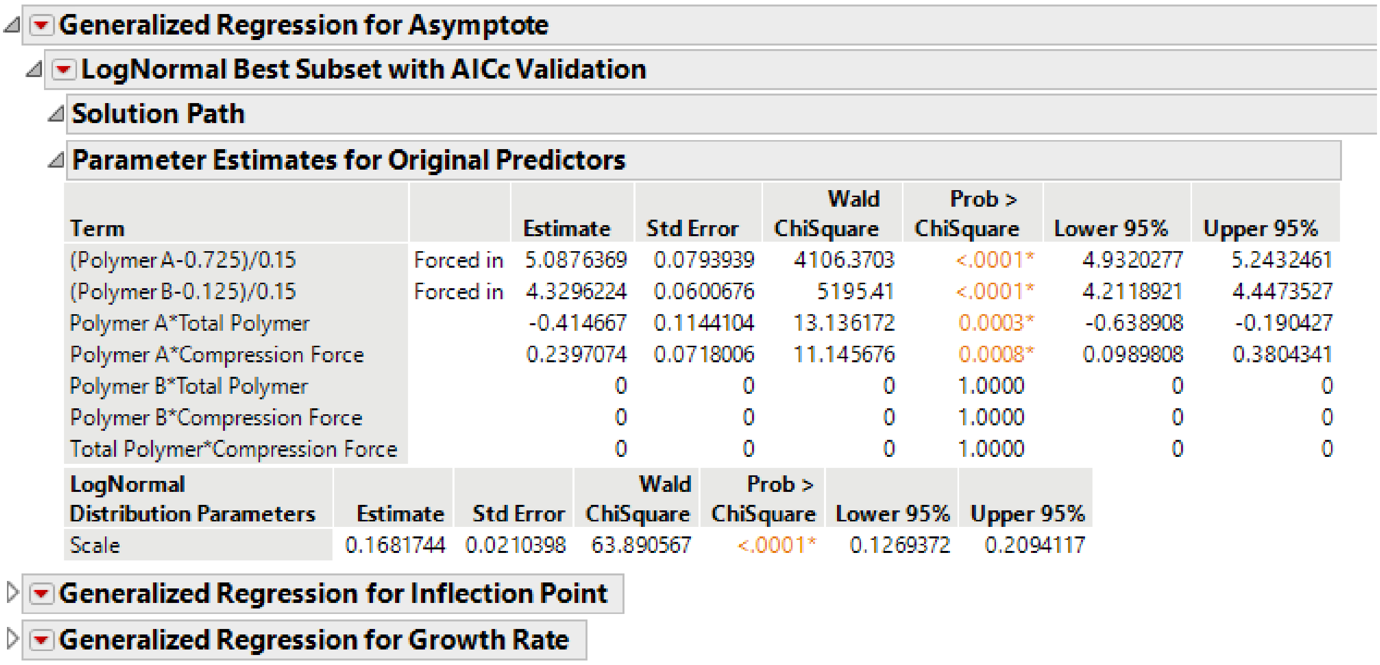
Figure 12: Weibull growth model



	Batch	Polymer A	Polymer B	Total Polymer	Compression Force	Asymptote	Inflection Point	Growth Rate
1	A01	0.825	0.175	0.16	2500	149.8	1043	0.733
2	A02	0.775	0.225	0.14	2500	104.7	282.5	0.938
3	A03	0.725	0.275	0.14	1500	87.08	107.5	1.311
4	A04	0.775	0.225	0.18	1500	78.78	128.1	1.17
5	A05	0.875	0.125	0.16	1500	91.69	104.5	0.622
6	A06	0.775	0.225	0.18	2500	68.95	248.1	1.454
7	A07	0.775	0.225	0.18	1500	87.61	193.8	0.938
8	A08	0.825	0.175	0.12	2500	150.7	521.5	0.534
9	A09	0.825	0.175	0.12	2500	257.6	2221	0.493
10	A10	0.875	0.125	0.16	2500	199.4	1963	0.565
11	A11	0.875	0.125	0.16	1500	139.6	365.3	0.44
12	A12	0.825	0.175	0.12	1500	135.8	235.2	0.553
13	A13	0.725	0.275	0.14	2500	79.21	158	1.296
14	A14	0.725	0.275	0.14	1500	86.17	83.28	1.356
15	A15	0.775	0.225	0.18	2500	73.12	260.2	1.438
16	A16	0.825	0.175	0.12	1500	131.6	207.9	0.555
17	R01	•	•	•	•	90.65	199.4	1.144

Figure 13: Fit of formulation experiments data to Weibull growth model

The generalized regression analysis of the 16 formulation experiments for the asymptote, inflection point and asymptotes is presented in Figure 14.



Generalized Regression for Asymptote

LogNormal Best Subset with AICc Validation

Solution Path

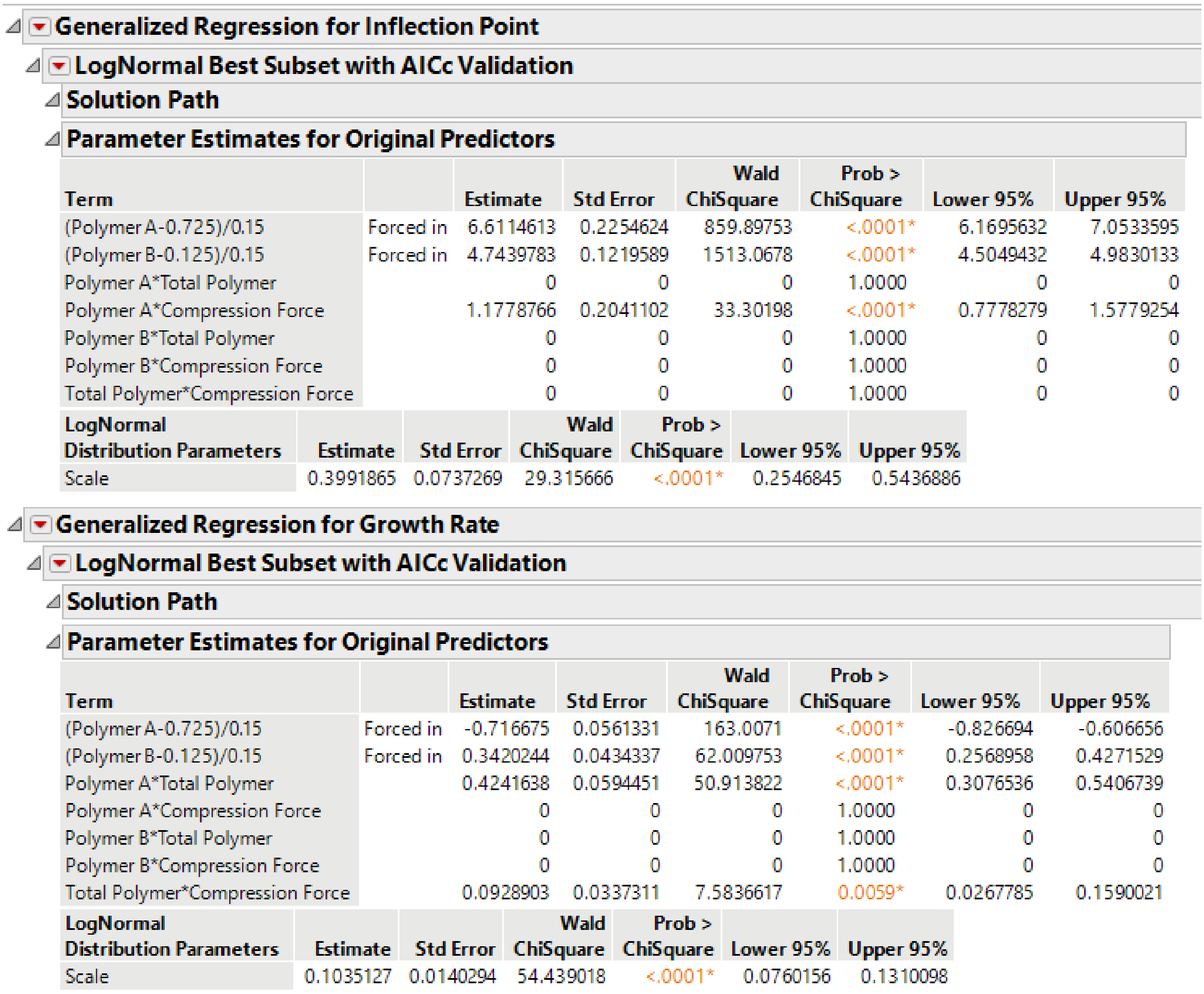
Parameter Estimates for Original Predictors

Term		Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
(Polymer A-0.725)/0.15	Forced in	5.0876369	0.0793939	4106.3703	<.0001*	4.9320277	5.2432461
(Polymer B-0.125)/0.15	Forced in	4.3296224	0.0600676	5195.41	<.0001*	4.2118921	4.4473527
Polymer A*Total Polymer		-0.414667	0.1144104	13.136172	0.0003*	-0.638908	-0.190427
Polymer A*Compression Force		0.2397074	0.0718006	11.145676	0.0008*	0.0989808	0.3804341
Polymer B*Total Polymer		0	0	0	1.0000	0	0
Polymer B*Compression Force		0	0	0	1.0000	0	0
Total Polymer*Compression Force		0	0	0	1.0000	0	0

LogNormal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	0.1681744	0.0210398	63.890567	<.0001*	0.1269372	0.2094117

Generalized Regression for Inflection Point

Generalized Regression for Growth Rate



Generalized Regression for Inflection Point

LogNormal Best Subset with AICc Validation

Solution Path

Parameter Estimates for Original Predictors

Term		Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
(Polymer A-0.725)/0.15	Forced in	6.6114613	0.2254624	859.89753	<.0001*	6.1695632	7.0533595
(Polymer B-0.125)/0.15	Forced in	4.7439783	0.1219589	1513.0678	<.0001*	4.5049432	4.9830133
Polymer A*Total Polymer		0	0	0	1.0000	0	0
Polymer A*Compression Force		1.1778766	0.2041102	33.30198	<.0001*	0.7778279	1.5779254
Polymer B*Total Polymer		0	0	0	1.0000	0	0
Polymer B*Compression Force		0	0	0	1.0000	0	0
Total Polymer*Compression Force		0	0	0	1.0000	0	0

LogNormal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	0.3991865	0.0737269	29.315666	<.0001*	0.2546845	0.5436886

Generalized Regression for Growth Rate

LogNormal Best Subset with AICc Validation

Solution Path

Parameter Estimates for Original Predictors

Term		Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
(Polymer A-0.725)/0.15	Forced in	-0.716675	0.0561331	163.0071	<.0001*	-0.826694	-0.606656
(Polymer B-0.125)/0.15	Forced in	0.3420244	0.0434337	62.009753	<.0001*	0.2568958	0.4271529
Polymer A*Total Polymer		0.4241638	0.0594451	50.913822	<.0001*	0.3076536	0.5406739
Polymer A*Compression Force		0	0	0	1.0000	0	0
Polymer B*Total Polymer		0	0	0	1.0000	0	0
Polymer B*Compression Force		0	0	0	1.0000	0	0
Total Polymer*Compression Force		0.0928903	0.0337311	7.5836617	0.0059*	0.0267785	0.1590021

LogNormal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	0.1035127	0.0140294	54.439018	<.0001*	0.0760156	0.1310098

Figure 14: Fit of generalized regression the three to Weibull growth model parameters

The asymptote is affected by Polymer A and B and the interaction of Polymer A with the total amount of polymer and the Compression Force. The Inflection point is determined by Polymer A and B and the interaction of Polymer A and the Compression Force. The growth rate is set up by Polymer A and B, the interaction of Polymer A with the total amount of polymer and of Total Polymer with the Compression Force.

To design the ideal formulation matching the reference tablets s we minimize the distance between the optimized model and the reference model. The optimized set up using FDA and NRL are listed in Table 1.

Table 1: Optimized formulation with FDA and NLR model

Model	Polymer A	Polymer B	Total Polymer	Compression Force
FDA	0.725	0.275	0.17	1700
NLR	0.758	0.242	0.16	2100

Under these optimized formulations we compare the designed formulation dissolution curve and the reference dissolution curve. Figure 15 presents these curves. NLR clearly matches the Reference path better than the FDA path.

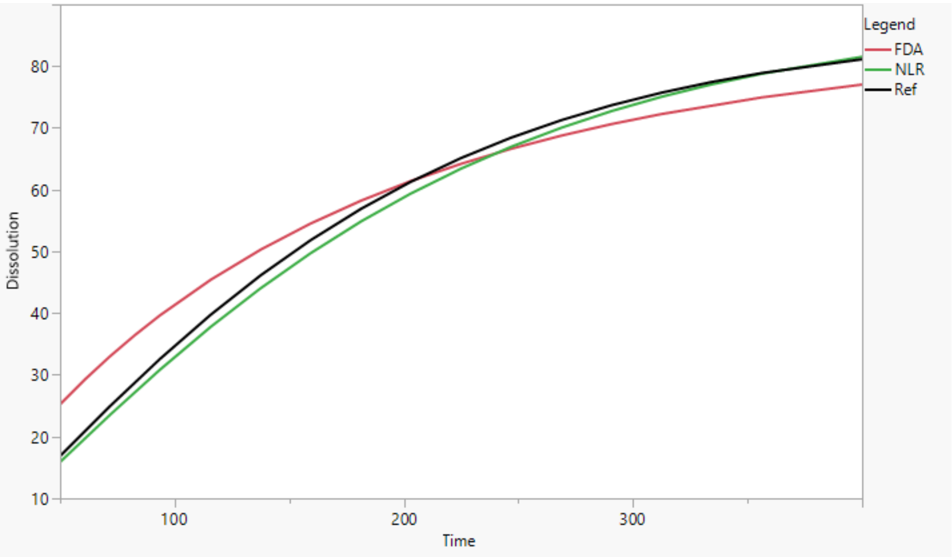


Figure 15: Fit of optimized formulations under FDA and NLR and Reference path

The next section is comparing the models using the information quality framework.

6. Information Quality Assessment and Discussion

In assessing information quality derived from applying the FDA and NLR models we find similar positive responses to the checklist questions in Section 2 for: Data Resolution, Data Structure, Data Integration, Temporal Relevance and Chronology of Data ad Goal. These dimensions were determined to be similar for FDA and NLR. The remaining three dimensions turned out to produce difference responses for FDA ad NLR. The questions, for these dimensions, also listed in Section 2, with evaluation responses are:

(6) *Generalizability*: Can you derive general conclusions based on the study, beyond what was explicitly studied, for example to other products or processes?
FDA: The non-parametric FDA model is focused on functional form.
NLR: The Weibull growth function is based on the work of Langenbucher (1972). The implication being that the three function parameters are interpretable on mechanistic knowledge.
NLR had better generalizability properties than FDA.

(7) *Operationalization*: Are the measured variables themselves of relevance to the study goal? Are stated action items derived from the study?
Computationally NLR is also affected by starting point and convergence properties. FDA also requires numerical tweaking such as setting the number and position of knots and the type of spline function used. Overall FDA is easier to operationalize than NLR.

(8) *Communication*: Are findings properly communicated to the intended audience?
The dissolution curves parametrized version from NLR analysis arrive better communicated than the functional FDA form. In addition the NLR parameters meet the model based regulatory approach described in the guidance for dissolution (Food and Drug Administration, 1997).NLR has better communication properties than FDA.

This high-level assessment of information quality indicates that NLR generates higher information quality than FDA. Combining NLR with FDA can handle better operationalization.

References

- Bates, D. and Watts D. (2007) Nonlinear Regression Analysis and Its Applications, Wiley
- Food and Drug Administration (1997) Guidance for Industry Dissolution Testing of Immediate Release Solid Oral Dosage Forms, Center for Drug Evaluation and Research (CDER), https://www.fda.gov/media/70936/download
- Karlin, S. and Pinkus, A. (1976) Interpolation by Splines with Mixed Boundary Conditions, in Studies in Spline Functions and Approximation Theory, S. Karlin, C. A. Micchelli, A. Pinkus, I. J. Schoenberg, 305–325, Academic Press, N. Y.
- Kenett, RS and Gotwalt, C. (2021) Maximizing Data Science Success with Information Quality (InfoQ) and JMP® (2021-EU-45MP-750) JMP Discovery Summit Europe, https://community.jmp.com/t5/Discovery-Summit-Europe-2021/Maximizing-Data-Science-Success-with-Information-Quality-InfoQ/ta-p/349217
- Kenett, R.S. and Salini S. (2011). Modern Analysis of Customer Surveys: comparison of models and integrated analysis, with discussion, Applied Stochastic Models in Business and Industry, 27, pp. 465–475
- Kenett, R.S. and Shmueli G, (2014) On Information Quality, Journal of the Royal Statistical Society, Series A (with discussion), Vol. 177, No. 1, pp. 3-38,
- Kenett, R.S. and Shmueli G, (2016) Information Quality: The Potential of Data and Analytics to Generate Knowledge. John Wiley and Sons.
- Kenett, R.S. and Zacks, S. (2021) Modern Industrial Statistics: With Applications in R, MINITAB, and JMP, 3rd Edition, ISBN: 978-1-119-71490-3
- Langenbucher, F. (1972) Letters to the Editor: Linearization of dissolution rate curves by the Weibull distribution. J. Pharm. Pharmacol, 24 (12), 979–981.
- Ramsay, J. O. and Silverman, B. W. (2002). Applied Functional Data Analysis: Methods and Case Studies. Springer, New York, NY.
- Reis, M. and Kenett, R.S. (2018) Assessing the Value of Information of Data-Centric Activities in the Chemical Processing Industry 4.0, AIcHe, Process Systems Engineering, 64(11), pp. 3868-3881
- Schoenberg, J. (1958), "Spline functions, convex curves and mechanical quadratures", Bull. Amer. Math. Soc. 64, 352–357.
- Thakur, G, Ghumade, P. and Anurag S. Rathore, A. (2021) Process analytical technology in continuous processing: Model-based real time control of pH between capture chromatography and viral inactivation for monoclonal antibody production, Journal of Chromatography A, 68, https://doi.org/10.1016/j.chroma.2021.46261
- Woodall, W., Spitzner, S., Montgomery D. Shilpa Gupta S. (2004) Using Control Charts to Monitor Process and Product Quality Profiles, Journal of Quality Technology, 36(3), pp. 309-320.