JMP Academic Case Study 060

Nonlinear Regression Modeling for Cell Growth Optimization

Nonlinear Modeling, Curve DOE

Produced by

Benjamin Ingham, The University of Manchester

benjamin.ingham@manchester.ac.uk

jmp STATISTICAL DISCOVERY

Nonlinear Regression Modeling for Cell Growth Optimization

Nonlinear Modeling, Curve DOE

Key Ideas

The case study focuses on understanding the impact of categorical (cell line) and continuous (sugar concentration) factors on cell growth. Cell growth data is inherently nonlinear, with distinct growth phases and varying curve shapes that depend on experimental conditions. We need to apply nonlinear regression models with parameters that represent the key characteristics of a given curve. The individual parameter estimate values can then be captured and modeled against the experimental inputs using generalized regression. This approach allows you to profile the change of the growth curve over the experimental conditions and find conditions that provide a desired curve shape. With this direct understanding of how the growth curve is affected by experimental conditions, researchers can make more informed decisions, with a deeper understanding of the biological system, and achieve more optimal outcomes.

(This study uses Fit Curve and Curve DOE with JMP SE or JMP Pro.)

Background

You are attempting to grow a “wild type” bacteria and three genetically altered variants. You want to alter the percent of maltose in the growth media to see how the growth of the cell lines change. You plan to grow your cells in a 48-well plate and measure growth with absorbance at 600nm (optical density). Optical density is used to measure the amount of light scattered by microbial cells, since the more cells there are, the higher the optical density value. Your goals are to:

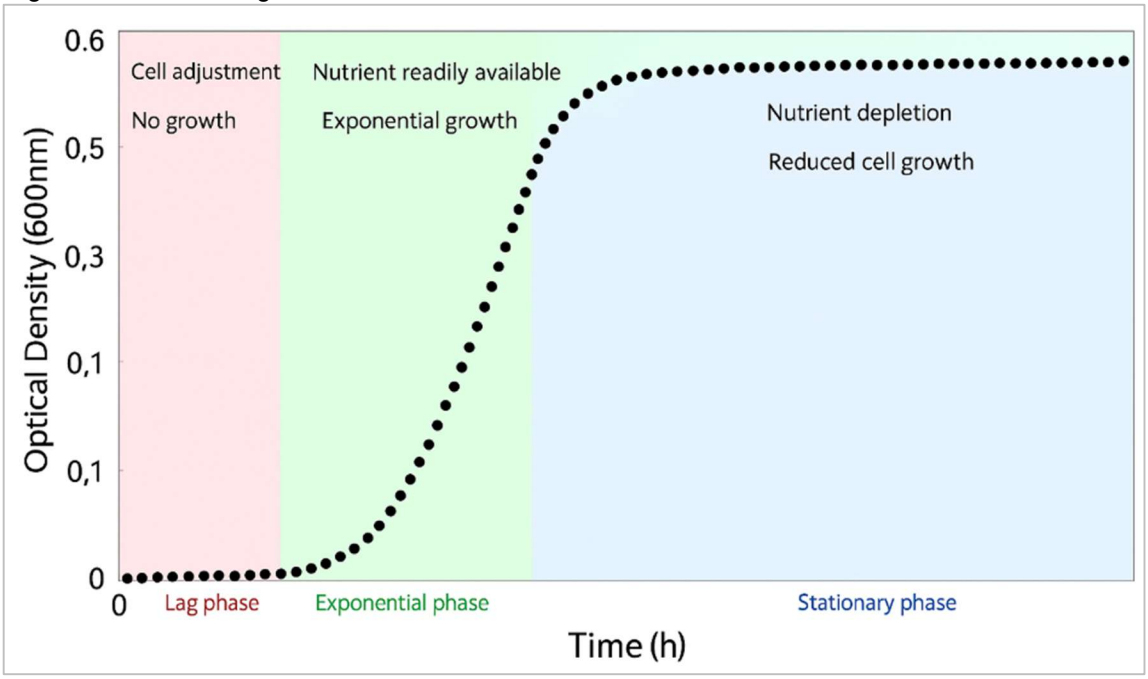
- Find the percent of maltose that achieves the “best growth” for each cell line (highest rate and highest final optical density).
- Find the best overall combination of maltose and cell line to achieve the highest number of cells.

The growth of bacteria generally follows a set growth profile that represents different phases of growth. Scientists are interested in capturing the occurrence of these stages to learn more about their microbe of study and to improve processes. The primary stages of growth (illustrated above) are:

- **Lag phase:** The initial period where cells are inoculated and acclimate to their conditions (environmental and media). In production processes, scientists want to ensure that the lag phase is as short as possible to reduce the total production time and reduce costs.
- **Exponential phase:** The period of exponential growth of the cells as they consume a readily available supply of nutrients. The growth rate during this stage is important to understand, as desirable rates are a balance between quick growth (reducing production time) and avoiding overwhelming a fermenter (i.e., consuming too much oxygen and nutrients leading to faster cell death). The midpoint of this phase (the midexponential) is a key characteristic that is considered when transferring seed/starting cultures into production processes.
- **Stationary phase:** The period when nutrients begin to deplete or toxic by-products begin to accumulate, leading to an equal portion of cell death and cell growth. The time that this occurs and the maximum value of biomass that is achieved are important as we often need processes that create high biomass and, in some cases, desired products are only expressed during the stationary phase. In this example, the stationary phase is not completely flat; even when a primary nutrient is depleted or toxins accumulate, cells can still grow (slowly) on alternative nutrients, if available.

- **Death phase (not shown):** The period when cells exhaust all nutrients or toxins build up to a level above tolerance. Cell death typically marks the end of a productive process, so knowing when it occurs can help inform when a production needs to stop.

When trying to analyze a growth curve, scientists want to profile the whole curve shape and find parameters that best represent each of these stages, which can be achieved with nonlinear regression/curve fitting.



The Task

You are entrusted with the following tasks:

- Test a range of maltose concentrations on the four different cell lines and measure the growth via optical density measurements.
- Use nonlinear curve analysis to understand how the curve shape is affected by changing factors.
- Identify optimized conditions for the growth of the cells.

You decide to grow each cell line using a 48-well plate, growing them under 0.25, 0.5, 1, 1.5, and 2% maltose in individual wells and incubating for 24.2 hours on a shaking plate reader. You run each cell line/maltose combination in triplicate.

The Data

The data set (**cell-growth.jmp**) contains:

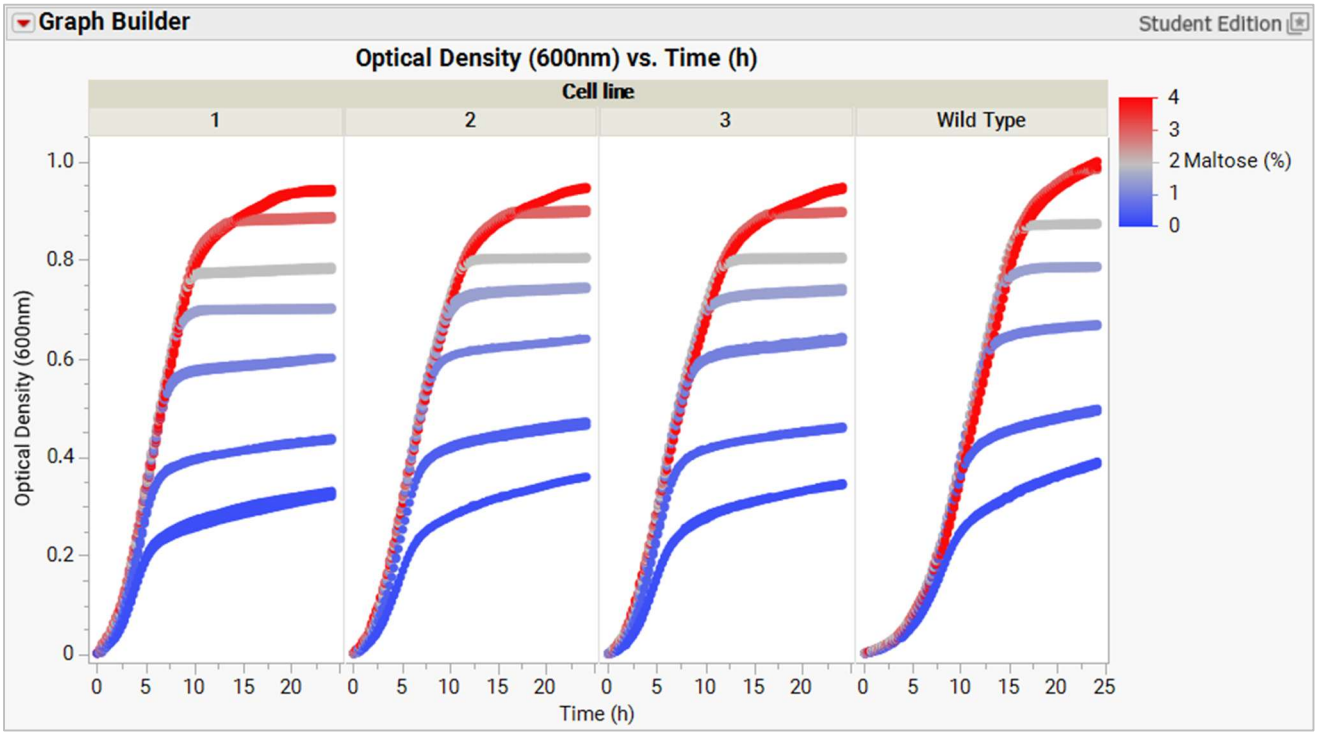
Well: A unique identifier for each growth curve under each setting; it is the actual location on the plate.
Run ID: A group identifier that defines the unique treatments of settings that were run in triplicate.
Cell line: The specific bacterial cell line that has been tested.
Maltose (%): The percentage of maltose added to the media at the beginning.
Time (h): The time reading from the plate reader.
Optical Density (600nm): The absorbance values from the reader at 600nm as a proxy of biomass.

Analysis

Exploring the raw data

Let's explore the data using Graph Builder in JMP. In Exhibit 1, we can see that changing the cell line and the percent of maltose causes a change to the shape of the growth curve. We can see that increasing the percent of maltose from 0.25% to 4% causes a higher final optical density value in all cell lines. Similarly, we can see that the genetically modified Cell Lines 1-3 start growing much faster than the wild type (shorter lag phase) regardless of maltose percent. But it is very difficult visually to determine a numeric value for the key characteristics that we're interested in and can take forward to perform statistical analysis on. To gain insight into characteristics such as growth rate, lag phase length, and mid-exponential time, we would have to manually measure each individual curve, which can quickly become overwhelming with large sample sizes. Similarly, we can't predict the curve shape between the maltose conditions that have been tested here – what if a better growth rate or set stationary phase value was found at 3.4% or 2.8%? It would be impossible to tell without running another experiment under those conditions. To solve this, we can apply nonlinear modeling to find these parameters and to model how the curve shape changes with the factors.

Exhibit 1 Raw data analysis of cell growth



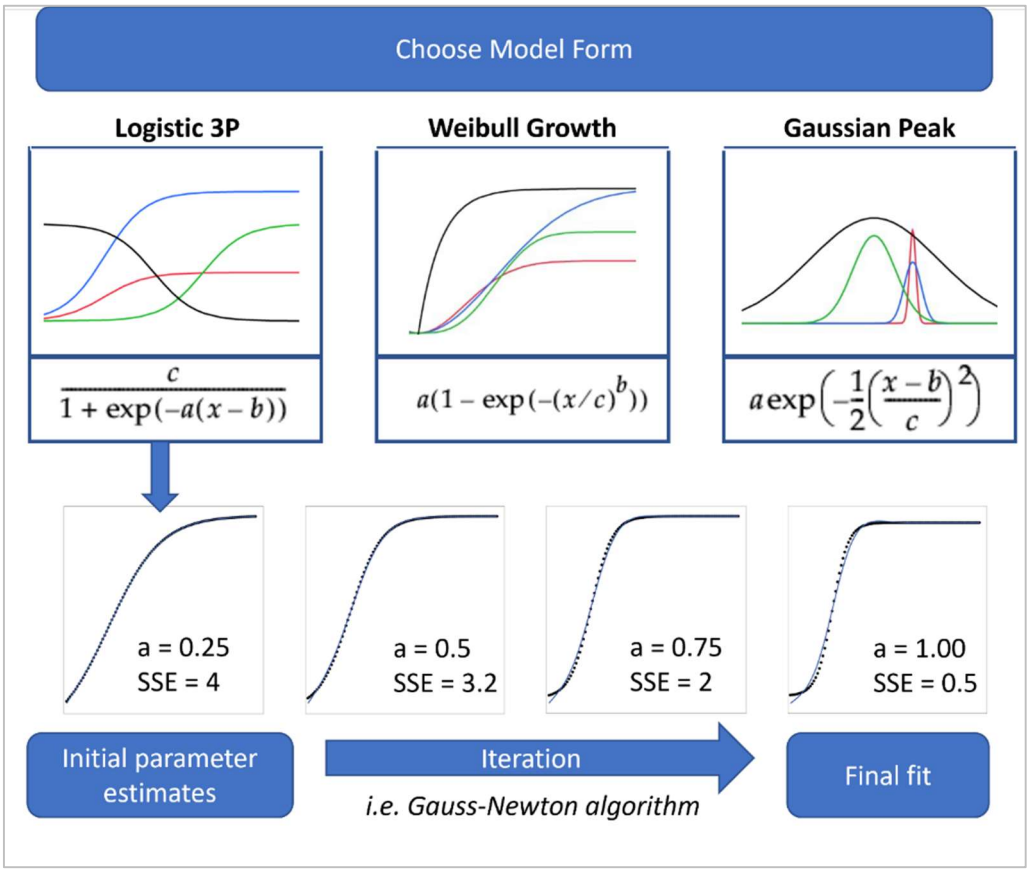
(To create this graph, go to Graph>Graph Builder. From the Control Panel, drag the Time (h) into the X axis of the graph and Optical Density (600nm) into the Y axis. To separate each data set into their respective experimental conditions, drag Cell Line into Group X and Maltose (%) into color.)

Linear vs. nonlinear regression modeling

Simple linear regression has a wide variety of uses for understanding a linear relationship between dependent and independent variables with a model, $y = mx+b$. However, if we were to apply it to the cell growth data, the model would fit poorly. Simple linear regression can be improved using higher order terms (e.g., quadratic and cubic) to capture curvilinear relationships. However, these may not be able to capture more complex relationships, such as those found in cell growth. In addition, the parameters in a polynomial model don't reflect components/phases of cell growth (e.g., growth rate, maximum optical density value).

Nonlinear regression is a more flexible technique that can be used to fit complex relationships between variables. The process of nonlinear regression is summarized in Exhibit 2. First, an appropriate model is selected, guided by domain knowledge, that best fits the curves of interest. This model is composed of parameters that define the shape of the curve (for example, Growth Rate A, Maximum Value B). These parameters are initially fit to the model with estimated parameter values and the difference between the fitted and actual Y values are compared with the sum of squares. From this, the model iteratively changes the parameter estimates until the sum of squares is minimized. It can be done through trial and error or determined through algorithms including Gauss-Newton or Levenberg-Marquardt. More modern nonlinear regression models will also use estimators, such as the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC), to select appropriate parameter values.

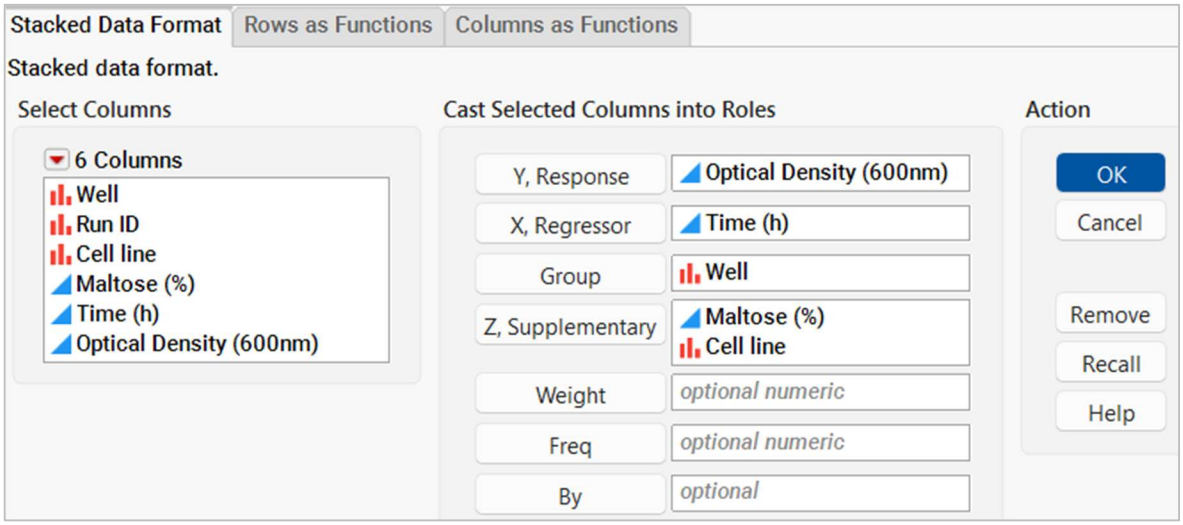
Exhibit 2 Nonlinear regression process summary



Applying a nonlinear model

To fit a suitable model to our growth curve data, we can use the Fit Curve platform in JMP to apply different nonlinear regression models as shown in exhibit 3.

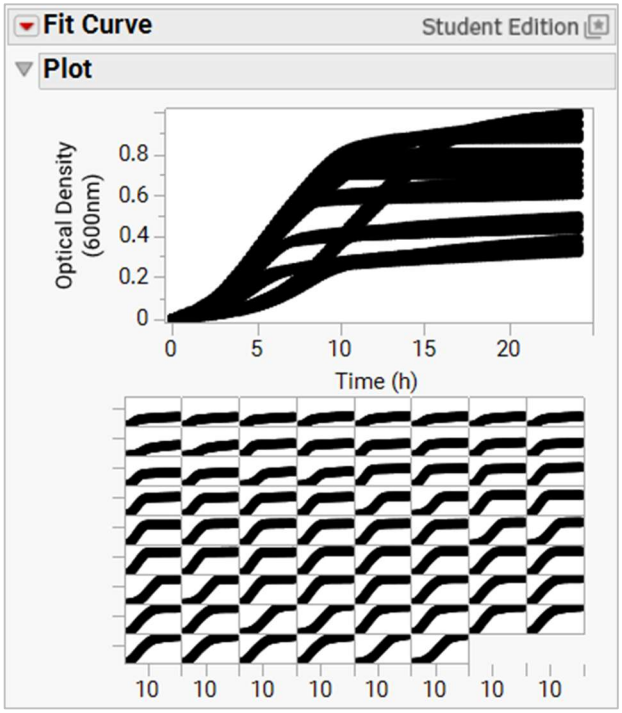
Exhibit 3 Fit Curve selection window



(To launch Fit Curve, select Analyze > Specialized Modelling > Fit Curve and input your data as shown above.)

The initial Fit Curve output is shown in Exhibit 4.

Exhibit 4 Initial Fit Curve output

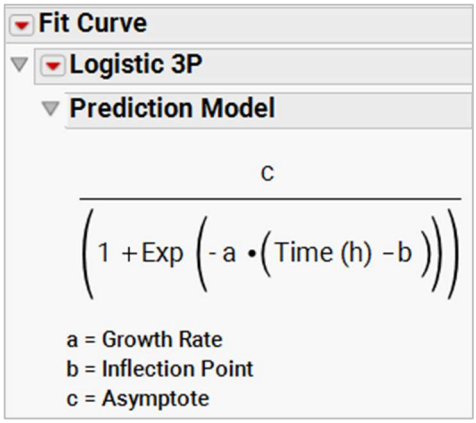


The platform displays the experimental run data overlaid (top) and individually (lower). From this point, we need to select a suitable model to fit the data. In this example, a sigmoid or “logistic” model has been fit. There are several types of sigmoid models, all of which differ slightly in how they capture the curves. Typically, the parameters include:

- **Growth rate:** The rate at which the dependent variable (number of cells) increases or decreases over time. Initially, the growth rate is high and then gradually slows down.
- **Inflection point:** The point at which the growth rate of the logistic function changes from increasing to decreasing. It marks the midpoint of the curve’s S-shaped growth pattern.
- **Asymptote (upper and lower):** A straight line or curve that a function approaches but never touches. As the independent variable approaches a certain value, the dependent values get closer and closer to the asymptote, but never actually reach it. This parameter can be used as an indicator for the maximum (upper) and minimum (lower) limit. In cases where a value should not reach a value below 0, a single asymptote (upper by default) is selected.

In this example we have selected a Logistic 3P model to best represent our curves. This model contains three parameters: growth rate, inflection point, and maximum asymptote. As the data involves cell growth, we would not expect to see any optical density values less than 0 (you can’t have a negative population), so there is no need to include a lower asymptote parameter (this can be found in 4- and 5-parameter logistic models). Logistic models find the inflection point exactly between the two asymptotes, producing a symmetrical curve at that point. Other options, such as the Gompertz 3P model, add additional parameters that allow for the inflection point to be asymmetrical (not centered between the two asymptotes).

Exhibit 5 Sigmoid Logistic 3P model equation



(To select this model, select the red triangle on Fit Curve > Sigmoid Curves > Logistic Curves > Fit Logistic 3P.)

We can see how a Gompertz 3P and Logistic 3P differ using model comparisons (Exhibit 6), with values such as the SSE and AICc. We can see that a Logistic 3P model has a lower AICc and SSE value, indicating a better fit to the data. While it appears that the Logistic 3P model is the better choice for this data set, you should also use your domain understanding to inform the choice of model.

Exhibit 6 Nonlinear model comparison

Model Comparison

Model	AICc	AICc Weight	.2 .4 .6 .8	BIC	SSE	MSE	RMSE	R-Square
Logistic 3P	-43446.74	1		-42002.85	1.098931	0.0001539	0.0124061	0.9982427
Gompertz 3P	-41649.89	0		-40206.01	1.4032717	0.0001965	0.0140192	0.997756

(To compare against the Logistic 3P model, select the red triangle on Fit Curve > Sigmoid Curves > Fit Gompertz 3P.)

Exhibit 7 Logistic 3P model output: parameter values

Logistic 3P

Prediction Model

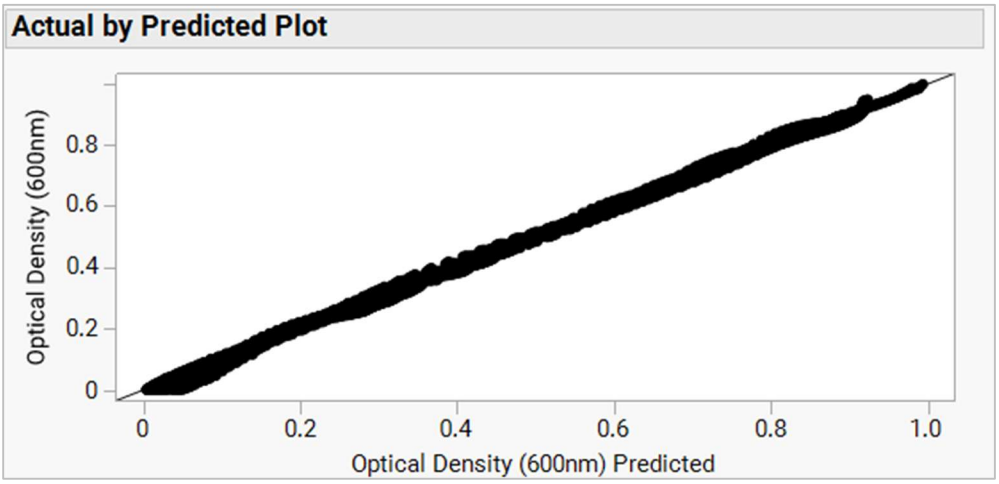
Summary of Fit

Group Summary

Group	Growth Rate	Inflection Point	Asymptote
B10	0.4178285	8.6113697	0.3641161
C1	0.7902571	4.1775371	0.4171163
C2	0.8104748	4.3011713	0.4172074
C3	0.8472807	4.2303269	0.4151712
C4	0.7244545	4.7856835	0.4445249
C5	0.7204813	4.8473498	0.4510777
C6	0.7440607	4.8556361	0.4419537
C7	0.7519593	4.8716634	0.4410772
C8	0.7434436	4.9182619	0.4427704
C9	0.5612101	8.1510878	0.4746681
C10	0.5612385	8.2310527	0.4796236
D1	0.7974184	4.681692	0.5879515
D2	0.8031786	4.727463	0.5924882
D3	0.8204335	4.6819885	0.5918444
D4	0.7026591	5.2067451	0.6277923

As shown in Exhibit 7, each Group (Well ID) has a unique value for each of the parameters of the Logistic 3P model. We can use these values and place them in the model equation in Exhibit 5 to recreate the curve shape of the original data. We can see how well this fit compares to the real data by looking at the summary fit and actual vs. predicted plots (Exhibit 8).

Exhibit 8 Logistic 3P actual by predicted plot



(To create, select the red triangle on Logistic 3P > Plot Actual by Predicted.)

The application of nonlinear modeling has allowed us to characterize the variety of growth curves in our experiment with parameter estimates that can be used to recreate any given curve. With these parameter values, we can now model the effect of experimental factors on the growth curve.

Analyzing the parameter estimates

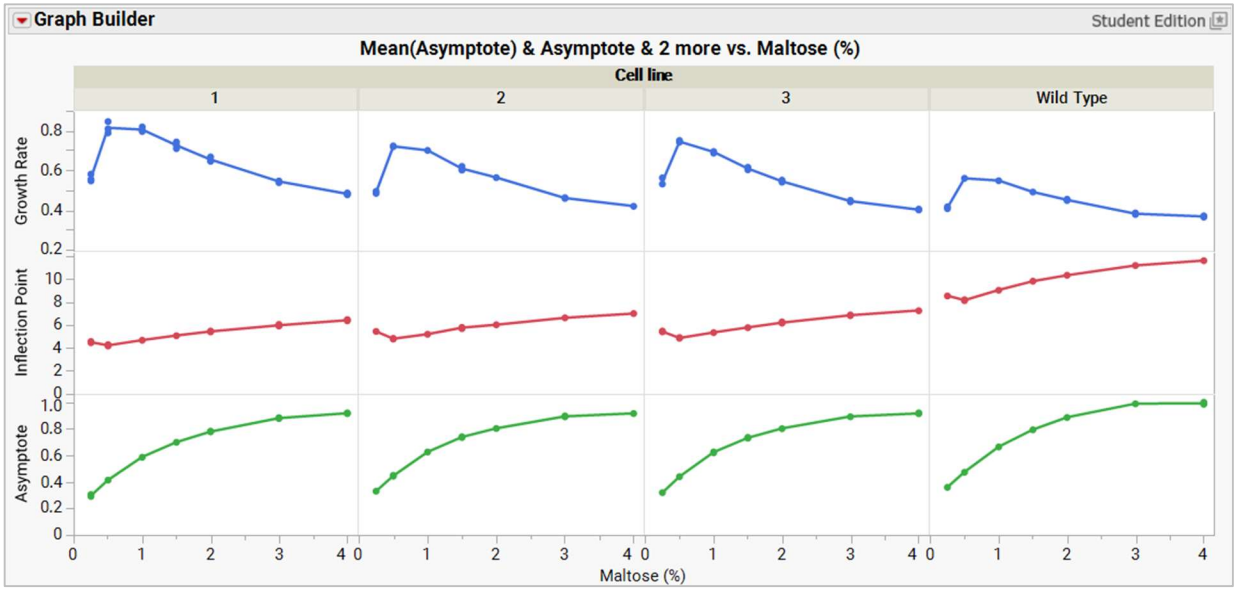
The nonlinear model has provided unique parameter estimates that represent a characteristic of the sigmoidal curve shape. As mentioned, biologists aim to simplify a cell growth curve down to set stages of growth (lag, log/exponential, stationary). Using this preference, we can link the parameters of the logistics model and represent these stages of growth, as summarized in Exhibit 9.

Exhibit 9 Equivalent parameters between cell growth and sigmoidal models

Cell growth parameter	Logistic 3P Parameter
Cell growth rate	Growth rate
Mid-exponential phase	Inflection point
Peak growth (stationary phase)	Asymptote

We can take the parameter values from the Logistic 3P model and analyze them as output variables against the input variables of cell line and maltose percent. We can look at this graphically (Exhibit 8) or by using regression analysis (standard least squares regression) to see the effect of cell growth/maltose. At a quick glance, we can see that Cell line 1 at 0.5% maltose has the highest growth rate and inflection point, while the Wild Type has the highest asymptote (indicating the highest final optical density value).

Exhibit 10 Graph Builder fit of parameter values



(To create this graph, click the red triangle on Logistic 3P and select Make Parameter Table. Select Graph Builder and place Growth Rate, Asymptote and Inflection Point in the Y axis (dragging on to the top section of the Y axis, Maltose (%) in the X axis and Cell line in the Group X. Replace the Smoother by the Line element (use shift-key to combine elements)).

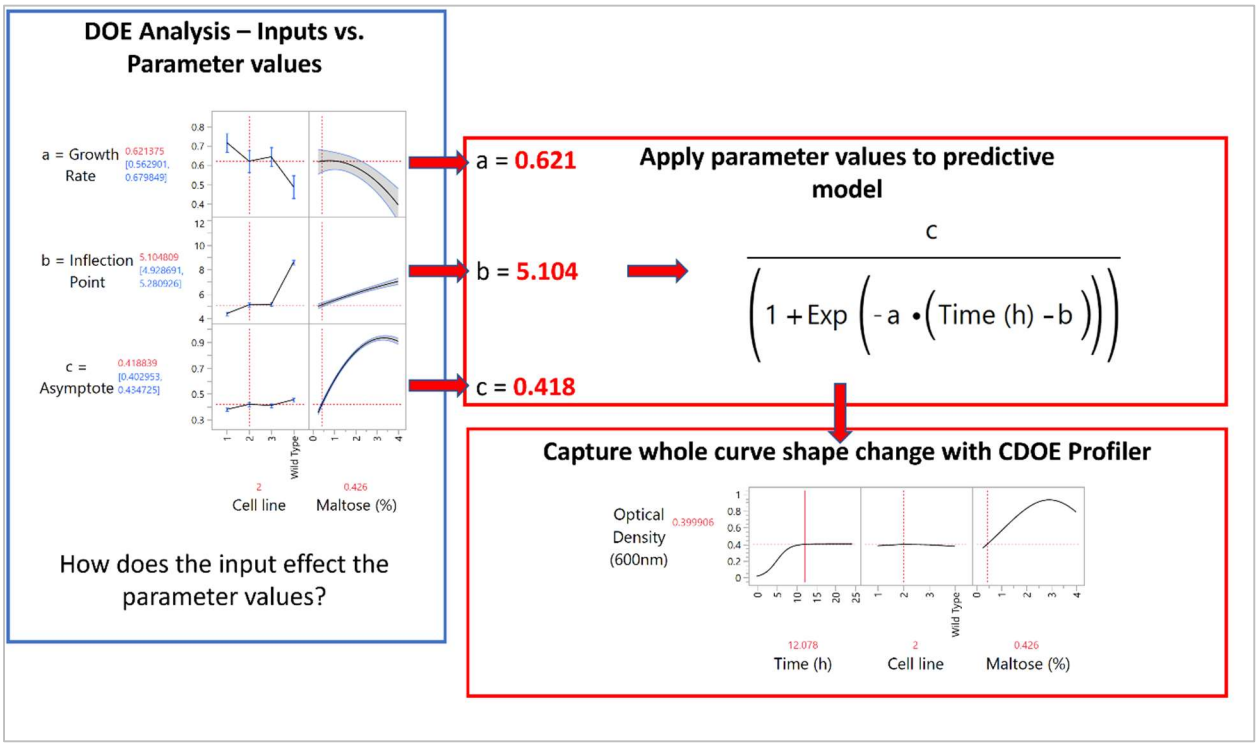
While analysis of the individual parameters can be useful for assessing the effect of a condition on the growth of cells, simplifying a curve down to these components for analysis is limited. Analysis can quickly become cumbersome when trying to compare across models and draw meaningful conclusions – especially with more complex nonlinear models (more parameters) and experiments (more factors). In addition, focusing on individual parameter estimates does not enable you to understand the change in the whole curve. The effect of changing a single parameter may not be linear, and the overall impact on growth may be more significant than the individual estimates suggest. By only considering individual parameters, researchers may miss important interactions between different factors that affect the curve's shape and behavior, which can lead to incomplete or inaccurate conclusions about the effect of changing

experimental conditions on growth. Instead, we can use all of the parameter values together with curve design of experiments analysis (Curve DOE) to understand the change in the shape of the entire curve.

Curve DoE analysis

From the nonlinear Logistics 3P model, we have generated a set of parameter values that represent key characteristics of the curve. We may have also generated individual regression models for each of these terms, but how do we relate these values/models to the effect of the input on the overall shape? Curve DOE analysis works by using the parameter values as a response (output) in a DOE model against the cell line and maltose percent (inputs or factors), estimating its effect on the parameter value and determining any interactions or curvature (cubic or quadratic). This method produces a model where parameter scores can be predicted for any cell line or maltose percent, including conditions that haven't been tested. Once we have a model that can predict the parameter values under a range of simulated conditions, we can take these values and substitute them into the original Logistic 3P prediction model and see how the shape of the curve changes, producing a profiler that links the change of the inputs to the change in the curve shape. This whole process is contained within the Curve DOE tool, which automatically performs the steps outlined in Exhibit 11.

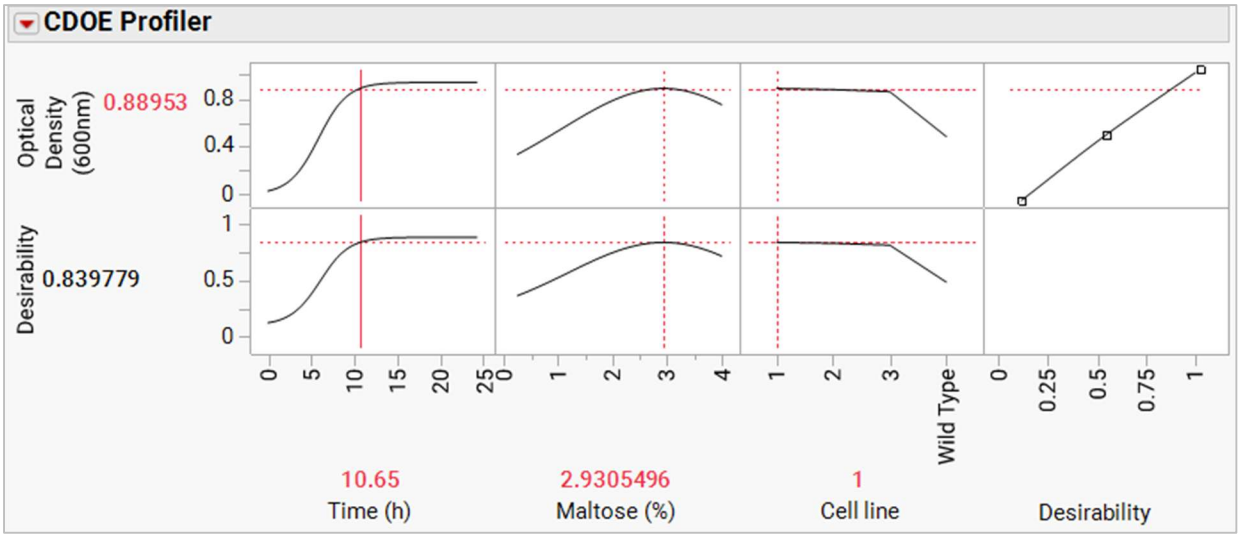
Exhibit 11 Curve DOE process



From the CDOE profiler (exhibit 12), we can explore how the changes in cell line and maltose percent alter the whole curve shape, allowing us to consider multiple desirable characteristics (quick growth, short lag phase, short exponential phase) in one simple, easy-to-read profiler. The cell line and maltose percent can be altered manually to determine optimum conditions by eye, however the inbuilt optimization tools in the Curve DOE tool can be used for more accurate predictions. Initially, the profiler "locks" the time profiler, allowing the user to alter the cell line and maltose values to find the optimum values at the specific time selected in the model (e.g., 10.65h in Exhibit 10). This can be useful when wanting to maximize growth within a short time frame, such as determining which conditions would provide the highest number of cells at a set time. This tool can be used to guide experimental work by suggesting conditions that have not been physically tested, for example, 2.93% maltose and cell line 1 in Exhibit 1 for

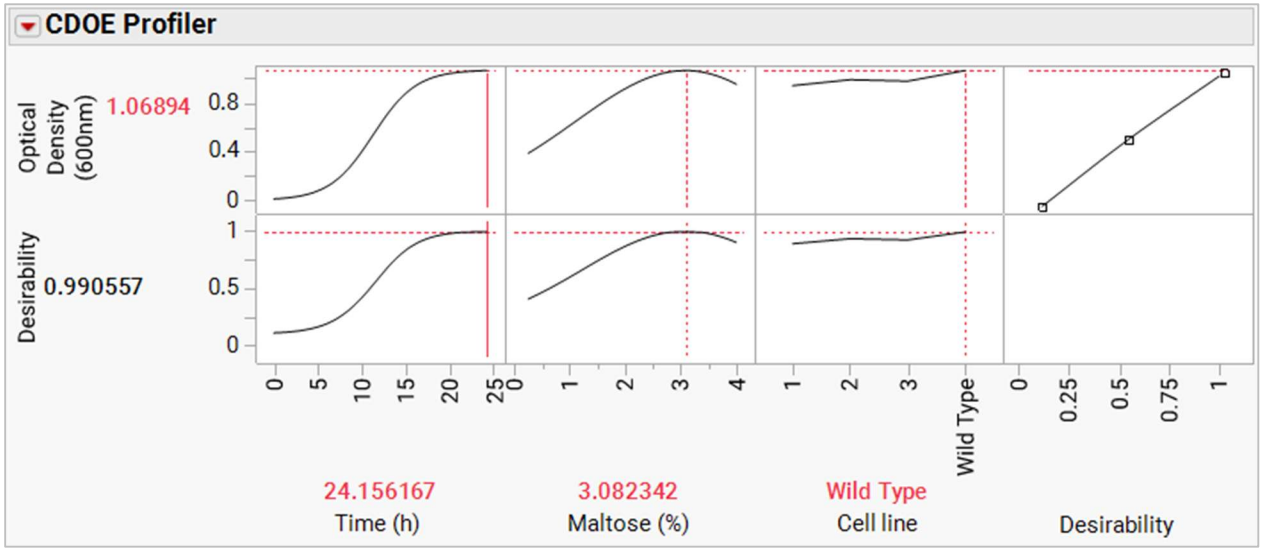
maximum growth at 10.65h. The profiler can also be used to find the best conditions for maximum growth, finding the best cell line, maltose percent, and time where cell growth is highest. This can be done by “unlocking” the time parameter to allow it to be changed (see Exhibit 13).

Exhibit 12 Curve DOE Profiler: optimized settings at specified times



(To create this profiler, select the red triangle on the Logistics 3P model section and click Curve DOE Analysis. Optimization settings can be found under the CDOE Profiler red triangle>Optimization and Desirability>Desirability Functions. Select Maximize Desirability to optimize your settings; desirability is set to Maximize as default.)

Exhibit 13 Curve DOE Profiler: overall optimum settings with unlocked profilers



(To add this Profiler, select the red triangle on the Logistics 3P model section and click Curve DOE Analysis. Optimization settings can be found under the CDOE Profiler red triangle>Optimization and Desirability>Desirability Functions. Select Maximize Desirability to optimize your settings; desirability is set to Maximize as default. To unlock the Time (h) profiler, hold Ctrl and click the time profile, then deselect Lock Factor settings.)

Summary

Statistical insights

Understanding cell growth profiles is essential in a wide variety of industries, and useful insights are needed to understand key stages in cell growth. In this example we have used nonlinear models to find parameters that can represent key traits in cell growth (e.g., asymptotes for maximum growth) that can be used as outputs to model against growth conditions (cell line and maltose percent), with single parameter analysis with graphing and standard least squares regression. We have applied Curve DOE analysis and profiled the effect of the growth conditions on the overall curve shape. From this method, we can conclude:

- The maximum cell growth is obtained with the Wild Type Cell line and 3.08% Maltose at 24.15h.
- The highest measured growth rate and lowest inflection point is in Cell line 1 at 0.5% Maltose.

Managerial/business implications

Traditional cell line or fermentation process development is laborious when analyzing cell growth data due to the complexity and size, typically requiring analysis by eye for each experimental run one by one. Fit Curve allows for rapid assessment of cell growth experiments by extracting key parameters that can be related to key biological growth stages. Curve DOE can then take these parameter values and summarize all the experimental runs into an easy to interpret CDOE profiler – simplifying data analysis, reducing turnaround time for results and improving the quality of analysis.

JMP features and hints

JMP was used in this case study to:

- Visually represent the cell growth data using Graph Builder.
- Fit curves to the cell growth data with a Logistic 3P nonlinear model.
- Visually represent the nonlinear model parameters using Graph Builder.
- Model the effect of experimental conditions (cell line and maltose percent) on the shape of the growth curves with Curve DOE and the CDOE Profiler.

Exercise

1. Try to apply different sigmoidal fits to the data. Are any of them better for fitting this curve data? Look at how the CDOE fit changes with each model type.
2. What are the potential pitfalls of using a nonlinear model with predetermined parameters? Look at the fit of the logistic models on Wells B1-B8, for example, focusing on the initial stages of growth.
3. What would you suggest as a follow-up experiment?