



JMP[®] for Mixed Models

Ruth M. Hummel
Elizabeth A. Claassen
Russell D. Wolfinger

The correct bibliographic citation for this manual is as follows: Hummel, Ruth M., Elizabeth A. Claassen, and Russell D. Wolfinger. 2021. *JMP® for Mixed Models*. Cary, NC: SAS Institute Inc.

JMP® for Mixed Models

Copyright © 2021, SAS Institute Inc., Cary, NC, USA

ISBN 978-1-952365-21-8 (Hardcover)

ISBN 978-1-951684-02-0 (Paperback)

ISBN 978-1-951684-03-7 (Web PDF)

ISBN 978-1-952363-85-6 (EPUB)

ISBN 978-1-952363-86-3 (Kindle)

All Rights Reserved. Produced in the United States of America.

For a hard copy book: No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

For a web download or e-book: Your use of this publication shall be governed by the terms established by the vendor at the time you acquire this publication.

The scanning, uploading, and distribution of this book via the Internet or any other means without the permission of the publisher is illegal and punishable by law. Please purchase only authorized electronic editions and do not participate in or encourage electronic piracy of copyrighted materials. Your support of others' rights is appreciated.

U.S. Government License Rights; Restricted Rights: The Software and its documentation is commercial computer software developed at private expense and is provided with RESTRICTED RIGHTS to the United States Government. Use, duplication, or disclosure of the Software by the United States Government is subject to the license terms of this Agreement pursuant to, as applicable, FAR 12.212, DFAR 227.7202-1(a), DFAR 227.7202-3(a), and DFAR 227.7202-4, and, to the extent required under U.S. federal law, the minimum restricted rights as set out in FAR 52.227-19 (DEC 2007). If FAR 52.227-19 is applicable, this provision serves as notice under clause (c) thereof and no other notice is required to be affixed to the Software or documentation. The Government's rights in Software and documentation shall be only those set forth in this Agreement.

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

June 2021

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to <http://support.sas.com/thirdpartylicenses>.

Contents

About This Book	vii
About the Authors	xi
Acknowledgements	xiii
1 Introduction	1
1.1 What is a Mixed Model?	1
1.2 Cell Viability Example	4
1.3 Mixed Model Assumptions	8
1.4 Nominal and Continuous Variables	12
1.5 Experimental Units and Blocking, Cell Growth Example	12
1.6 Confounding	15
1.7 JMP and JMP Pro	17
1.8 Exercises	19
2 ANOVA with a Single Blocking Effect	21
2.1 Motivating Examples	21
2.2 Blocking Designs and Skeleton ANOVA	21
2.3 Metal Bond Breaking Example	27
2.4 Balanced Incomplete Blocks Example	38
2.5 Exercises	44
3 Models with Factorial Treatment Designs	49
3.1 Motivating Examples	49
3.2 Conceptual Background	49
3.3 RCBD with Factorial Treatments, Tensile Strength Example	50
3.4 Split-Plot Design, Greenhouse Example	60
3.5 What About Interactions Between Fixed and Random Effects?	71
3.6 Nested Design, Semiconductor Example	73
3.7 Exercises	83

4	Multiple Random Effects	85
4.1	Motivating Examples	85
4.2	Conceptual Background	85
4.3	Latin Square - Blocking in Two Orthogonal Directions	86
4.4	Mouse Condition: Negative Block Variance Example	92
4.5	Exercises	102
5	Regression, Random Coefficients, and Multilevel Models	105
5.1	Motivating Examples	105
5.2	Conceptual Background	105
5.3	Stability Trial	106
5.4	Student Achievement Example	114
5.5	Exercises	119
6	Repeated Measures and Longitudinal Data	121
6.1	Motivating Example	121
6.2	Conceptual Background	121
6.3	Repeated Measures Skeleton ANOVA and Statistical Model	128
6.4	Respiratory Ability	129
6.5	When Might We Choose Other Models?	141
6.6	Exercises	144
7	Spatial Models	147
7.1	Motivating Examples	147
7.2	Conceptual Background	147
7.3	Hazardous Waste Example	149
7.4	Alliance Wheat Trial	154
7.5	Further Statistical Details	164
7.6	Exercises	164
8	Simulation and Power Analysis	167
8.1	Motivating Examples	167
8.2	Simulation for Precision and Power	168
8.3	Type I Error Control Using the Semiconductor Design	176
8.4	Power Using the Semiconductor Design	179
8.5	Confidence Interval Coverage Using the Winter Wheat Example	187
8.6	Simulating Mixed Model Data Directly with JSL	193
8.7	Exercises	203
9	Generalized Linear Mixed Models	205
9.1	Motivating Examples	205
9.2	Conceptual Background	205

9.3	Binomial Response: Shrub Coverage	207
9.4	Binary Response: Salamander Mating	212
9.5	Count Response: Manufacturing Imperfections	215
9.6	Exercises	221
10	Mixed Models Amidst Modern Debates	223
10.1	Statistical Pragmatism	223
10.2	Fixed versus Random Effects	224
10.3	Frequentist versus Bayesian, p -values	226
10.4	Causality versus Association	228
10.5	Explanation versus Prediction	231
10.6	Randomized Experiments versus Observational Studies	231
10.7	Exercises	233
A	List of Examples Used in This Book	235

About This Book

What Does This Book Cover?

Mixed models, which are an extension of classic statistical linear models (including analysis-of-variance and regression), are one of the most powerful and useful collection of methods for analyzing data from designed experiments. Variations of mixed models have been one of the strongest capabilities of SAS software since its beginnings in the mid 1970s. In parallel, JMP (a SAS product launched in 1989) has evolved into an incredibly powerful and popular tool for scientists and engineers. This book brings together these two legacies, and in example-driven fashion, walks through the core concepts of mixed models and how to best apply them in practice.

Mixed models are largely about how to handle experimental observations that are correlated. After introducing foundational concepts and terminology of mixed models with examples, the book covers increasing levels of complexity, revealing the richness and wide applicability of mixed models in most any discipline that collects data in well-formed experiments.

The first four chapters focus on mixed models in the context of analysis-of-variance (ANOVA). We find that ANOVA is a good place to start as it helps organize thinking around factors with a discrete number of levels, as are nearly always found in designed experiments. We proceed further and utilize a helpful construct known as a Skeleton ANOVA to help clearly break down and understand degrees of freedom as well as how information from experimental units in the design is being allocated to effects in the model.

Chapters 5 and 6 shift focus to continuous effects as commonly found in linear regression and repeated measures contexts. They fit quite naturally into the mixed model framework and can be effectively combined with ANOVA-style effects to handle a wide variety of common experimental setups. Chapter 7 covers spatial models, which extend mixed models further to handle covariance over two or more dimensions.

Chapter 8 shows how you can use simulation to rigorously explore deeper statistical properties of mixed models such as power and sampling distributions of outputs. Chapter 9 provides an introduction to generalized linear mixed models, which are used when the response is no longer normally distributed, such as when it is a discrete number of

successes or a count. Chapter 10 concludes the book with discussions on how mixed models relate to current controversies in the statistical and broader scientific and engineering communities.

Is This Book For You?

JMP for Mixed Models builds on the success of the *SAS for Mixed Models* book series as well as several other related books and articles. In contrast to the SAS procedures and code forming the basis of previous books, JMP and JMP Pro offer the ability to fit mixed models from a dynamically interactive, mouse-driven interface. This enables you to use mixed models without having to write code and get to important results faster and with less effort. This book is designed as an instructional guide along these lines and is the very first of its kind in this regard.

If you fit one of the following two characterizations, this book is likely for you:

- You are a scientist or engineer running experiments in which subsets of the observations are correlated due to the design or the nature of the experimental units themselves. This includes designs such as a randomized block or split-plot as generated by JMP's rich design-of-experiments (DOE) routines.
- You are familiar with running mixed models, hierarchical linear models, or multi-level models in SAS, R, or other languages and want to learn an easy, point-and-click interface to fit them and obtain dynamically integrated statistics and graphics to aid in their interpretation and presentation.

If you take the time to learn mixed models in JMP, they will likely become one of the most useful tools that you have for analyzing designed experiments.

What Are the Prerequisites for This Book?

We assume you have knowledge of introductory statistical concepts such as those taught in an advanced high school or first-year college curriculum. This includes topics such as the following:

- Statistical Testing (of the mean, of the difference between two means), standard errors (of the mean, difference between two means), and t tests
- Distributions (normal, binomial, uniform, t, chi-square, F)
- One-Way ANOVA
- Factorial ANOVA
- Regression
- ANCOVA (regression with groups)

In JMP or JMP Pro, the Fit Model platform is the central one we will use, and some basic familiarity with it will be very helpful.

What Should You Know about the Examples?

Each topical chapter in this book begins with a description of several motivating examples that utilize the topic, and then we present the necessary conceptual background. With the background in place, we analyze the examples using both JMP and JMP Pro, including full interpretations of the output.

If you already have a decent understanding of mixed models and/or JMP, you may want to skip straight to examples that best match the problem that you want to analyze. Although the book roughly proceeds from simpler to more complex topics in a somewhat logical fashion, it is also designed to be a reference book in which you can find an example that most closely matches your current problem and skip directly to it.

Software Used to Develop the Book's Content

We use both JMP and JMP Pro throughout, highlighting key differences as they arise.

Example Code and Data

JMP tables for all of the examples are available in the books supplemental information web page. JMP Scripting Language (JSL) programs are either included with the tables themselves or provided as stand-alone programs that you can open in JMP and run.

Output and Graphics

All of the books output and graphics are generated on an Apple MacIntosh. If you are running on Microsoft Windows, the aesthetics of the output will be somewhat different but content should be the same.

We use several typeface conventions throughout the book to help demarcate between data set names, variable names, commands, etc. **Data sets**, **variables**, and **functions** are monospace. *Platforms*, *menus*, *options*, *variable roles*, *buttons*, and *function groups*, basically anything you click on, are italicized. Bold font is mostly used only for section headings, but it is also used to call out **table names** in JMP reports.

We Want to Hear from You

SAS Press books are written by SAS Users for SAS Users. We welcome your participation in their development and your feedback on SAS Press books that you are using. Please visit sas.com/books to do the following:

- Sign up to review a book
- Recommend a topic
- Request information on how to become a SAS Press author
- Provide feedback on a book

Do you have questions about a SAS Press book that you are reading? Contact the author through saspress@sas.com or support.sas.com/author_feedback. SAS has many

resources to help you find answers and expand your knowledge. If you need additional help, see our list of resources: sas.com/books.

Chapter 1

Introduction

1.1 What is a Mixed Model?

Imagine you are lab scientist studying the effect of two chemicals, A and B, on cell viability. You prepare nine plates of media with healthy cells growing on each, and then apply A and B to randomly assigned halves of each plate. After a suitable incubation period, you collect treated cells from the halves of each plate and perform an assay on each sample to compute a measurement Y of interest. Four of the samples are accidentally contaminated during processing and produce no assay results. Your data table in JMP looks like Figure 1.1.

How should you analyze these data? A primary goal is to estimate the causal effect of Chemical on Y , while taking appropriate account of the experiment design based on Plate. A standard way to begin is to formulate a statistical model of Y as a function of Chemical and Plate. A *statistical model* is a mathematical equation formed using parameters and probability distributions to approximate a data-generating process. We refer to Y as the *response* in the model, or alternatively as the *dependent variable* or *target*. We refer to Chemical and Plate as *factors* or *independent variables*.

Note the different natures of Chemical and Plate. Chemical has two specifically chosen levels, A and B, whereas the levels of Plate are effectively a random set of such plates you routinely make in your lab. This is a most basic example of a case in which you would want to use a *mixed model*, which is a statistical model that includes both *fixed effects* and *random effects*. Here Chemical would be considered a fixed effect and Plate a random effect.

Figure 1.1: Cell Viability Data

	Plate	Chemical	Y
1	1	A	3.7
2	1	B	3.8
3	2	A	•
4	2	B	1.4
5	3	A	5.1
6	3	B	6.1
7	4	A	5.1
8	4	B	7.9
9	5	A	3.9
10	5	B	•
11	6	A	•
12	6	B	8.1
13	7	A	4.3
14	7	B	8.3
15	8	A	2.7
16	8	B	•
17	9	A	5
18	9	B	6.8

Key Terminology

Fixed Effect A statistical modeling factor whose specific levels and associated parameters are assumed to be constant in the experiment and across a population of interest. Scientific interest focuses on these specific levels. For example, when modeling results from three possible treatments, your focus is on which of the three is best and how they differ from each other.

Random Effect A statistical modeling factor whose observed values are assumed to arise from a probability distribution, typically assumed to be normal (Gaussian). Random effects can be viewed as a random sample from a population that forms part of the process that generates the data you observe. You want to learn about characteristics of the population and how it drives variability and correlations in your data. You want inferences about fixed effects in the same model to apply to the population corresponding to this random effect. You may also want to estimate or predict the realized values of the random effects.

Mixed Model A statistical model that includes both fixed effects and random effects.

Why is the distinction between fixed and random effects important? Many, if not most, real-life data sets do not satisfy the standard statistical assumption of independent observations. In the example above, we naturally expect observations from the same plate to be correlated as opposed to those from different plates. Random effects provide an easy and effective way to directly model this correlation and thereby enable more accurate inferences about other effects in the model. In the example, specifying Plate as a random effect enables us to draw better inferences about Chemical. Failure to appropriately model design structure such as this can easily result in biased inferences. With an appropriate mixed model, we can estimate primary effects of interest as well as compare sources of variability using common forms of dependence among sets of observations.

The use of fixed and random effects have a rich history, with countless successful applications in most every major scientific discipline over the past century. They often go by several other names, including *blocking models*, *variance component models*, *nested and split-plot designs*, *hierarchical linear models*, *multilevel models*, *empirical Bayes*, *repeated measures*, *covariance structure models*, and *random coefficient models*. They also overlap with *longitudinal*, *time series*, and *spatial smoothing* models. Mixed models are one of the most powerful and practical ways to analyze experimental data, and if you are a scientist or engineer, investing time to become skilled with them is well worth the effort. They can readily become the most handy method in your analytical toolbox and provide a foundational framework for understanding statistical modeling in general.

This book builds on the strong tradition of mixed model software offered by SAS Institute, beginning with PROC VARCOMP and PROC TSCSREG in the 1970s, to PROC MIXED, PROC PHREG, PROC NLMIXED, and PROC PANEL in the 1990s, PROC GLIMMIX in the 2000s, and more recently PROC HP MIXED, PROC LMIXED, PROC MCMC, PROC BGLIMM, and related Cloud Analytic Service actions in SAS Viya. We borrow extensively from *SAS for Mixed Models* by [Littell et al. \(2006\)](#) and [Stroup et al. \(2018\)](#). Mixed model software in various forms has evolved extensively and somewhat independently over the past several decades in other packages including R (lme4, lmer, nlme), SPSS Mixed, Stata xtmixed, HLM, MLwiN, GenStat, ASREML, MIXOR, WinBUGS/OpenBUGS, Stan, Edward, Tensorflow Probability, PyMC, and Pyro (web search each for details). The existence and popularity of all of these also speaks to the power and usefulness of mixed model methodology. Some differences in syntax, terminology, and philosophy naturally occur between the various implementations, and we hope the explanations and coverage in this book are clear enough to enable translation to other software should the need arise.

Mixed model functionality has been available in JMP since 2000 (JMP 4), and a dedicated mixed model personality in Fit Model was released in 2013 (JMP Pro 11). It continues to be an area of active development. The unique and powerful point-and-click interface of JMP, designed intrinsically around dynamic interaction between graphics

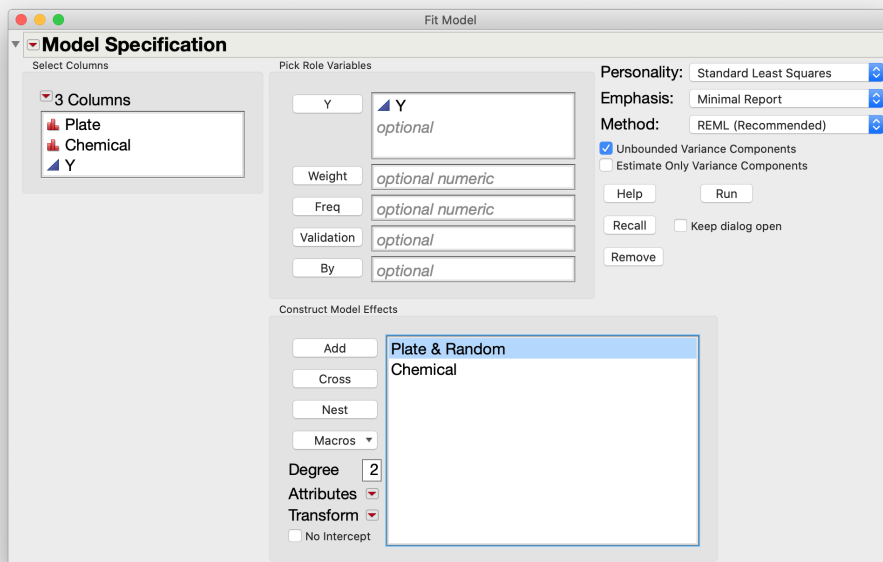
and statistics, makes it an ideal environment within which to fit and explore mixed models. Analyzing mixed models in JMP offers some natural conveniences over any approach that requires you to write code, especially with regards to the engaging interplay between numerical and pictorial results of statistical modeling. To get an initial idea of how it works, let's dive right into our first mixed model analysis in JMP.

1.2 Cell Viability Example

Consider the cell viability data shown the previous section and contained in `Cell Viability.jmp`.

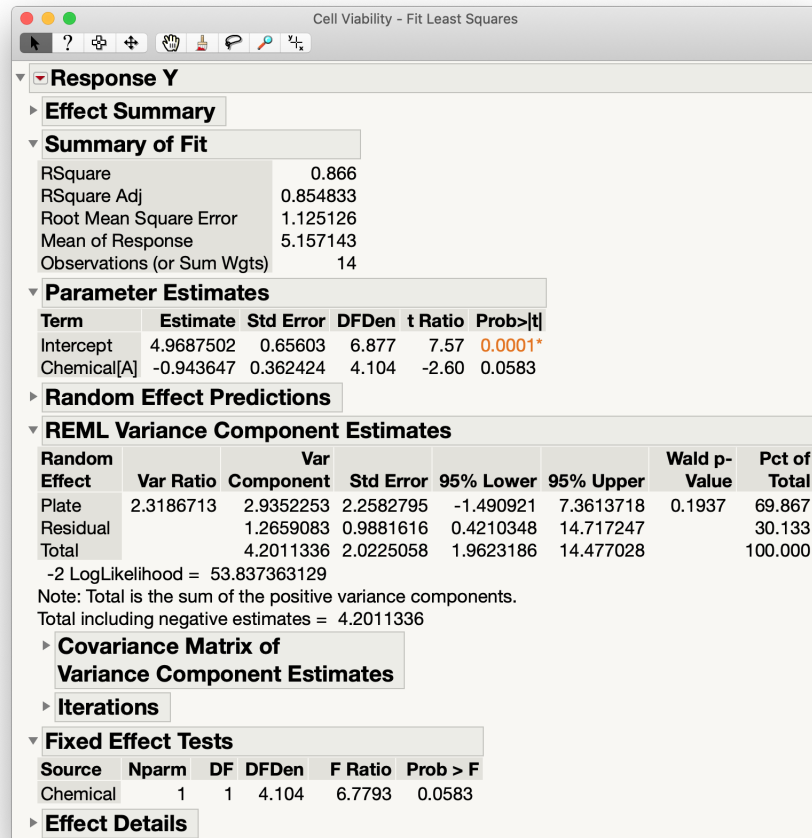
Using JMP

With the `Cell Viability` table open and active, from the top menu bar click *Analyze > Fit Model* to bring up a dialog box. On the left side, choose `Y`, then assign it to the `Y` role. Make sure the *Standard Least Squares* personality is selected in the upper right corner. Then select `Chemical` and `Plate`, and click *Add* to assign them to the *Construct Model Effects* box. In that box, select `Plate`, then click the red triangle  beside *Attributes* and select *Random Effect*. You will see "& Random" added beside `Plate` in the box, confirming designation as a random effect. Click *Run* to fit the model.



The model fitting results in Figure 1.2 are comprehensive with numerous statistics and details. We only focus on a few of the most important ones here and explore others in more depth in later chapters.

Figure 1.2: Mixed Model Results for Cell Viability Data



In the **Parameter Estimates** box in Figure 1.2, the row beginning with "Chemical[A]" contains the estimate of the effect of Chemical A (-0.94) along with its estimated standard error (0.36). Note this standard error is computed accounting for the random effect Plate in the model. Taking the ratio of these two numbers produces a t -statistic (signal-to-noise ratio) of -2.60. The associated p -value is 0.058, just above the classical 0.05 rule of thumb for statistical significance. As emphasized in recent commentary (see *The American Statistician* (2019)), such a borderline "non-significant" result should be inter-

preted in conjunction with the effect estimate itself and how it relates to estimated levels of variability in the context of the experiment.

Not shown is the estimate of Chemical B, which is automatically set equal to the negative of Chemical A in order to identify the model using the traditional sum-to-zero parameterization for linear models. The statistics for Chemical B are therefore identical to those from Chemical A, but the main effect and *t*-statistic have opposite signs. Our main conclusion is that Chemical B is estimated to have an overall effect around 1.9 units higher than Chemical A.

The **REML Variance Component Estimates** box provides estimates of the variance components along with associated statistics. Here we see that the estimate of plate-to-plate variability is 2.3 times larger than within-plate (residual error) variability. Such a result speaks to the two primary sources of random variability in this experiment and prompts questions as to why plates are varying to this degree.

Key Terminology

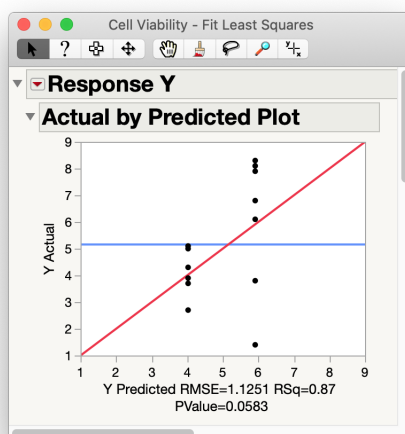
The acronym REML refers to restricted (or residual) maximum likelihood, the best-known method for fitting mixed models assuming that any missing data are missing at random, and equivalent to full information maximum likelihood from econometrics. Refer to [Stroup et al. \(2018\)](#) for details and theory behind REML in mixed models.

All of the mixed model results help to answer various aspects of research questions involving these chemicals and the assay used to assess them. Note you can also obtain confidence intervals by clicking the small red triangle  near the upper left corner of the report (just to the left of **Response Y**) then selecting *Regression Reports > Show All Confidence Intervals*. The 95% interval for the estimate of the Chemical A effect in this case is (-1.94, 0.05), just barely containing zero.

The red triangle  menu is loaded with several additional analyses, including many graphical displays. A key philosophy behind the design of JMP is to utilize relevant interactive graphics directly alongside statistics. As one good example, click  > *Row Diagnostics > Plot Actual by Predicted* to produce Figure 1.3.

Here the predicted values are based only on the fixed chemical effect from the model and not the random plate effect, explaining why there are only two distinct values on the X axis. A key aspect revealed by this plot is the increased variability in predictions corresponding to Chemical B on the right, as compared to those from Chemical A on the left. This is driven by the two lowest predicted values on the right. Selecting these two points in the graph and looking back at the table reveals they come from the first two plates. This is a reason to recheck that nothing unusual occurred on these two plates.

Figure 1.3: Actual by Predicted Values for Cell Viability Data



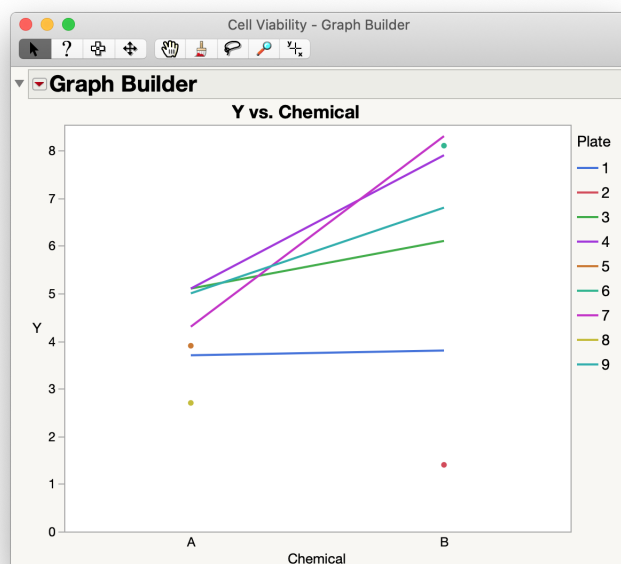
To track this further, let's plot the raw data. With the Cell Viability table in focus, click *Graph > Graph Builder*, assign Y to the Y axis drop zone, Chemical to the X axis drop zone, and Plate in the *Overlay* role. Then select the Line plot element and click *Done* to produce Figure 1.4.

The points in Figure 1.4 have similar orientation to the model-based ones in Figure 1.3, but now points occurring on the same plate are connected with a line. The four singleton points correspond to the four plates that contain only one observation, with the other one missing. The point in the far bottom right might be considered an outlier as well as influential in the estimate of plate-to-plate variability. A decision to potentially remove it depends critically on the quality of experiment protocol, and care must be taken to maintain data and research integrity.

The preceding analysis flow illustrates how to perform a basic mixed model analysis within JMP and the ease with which you can effectively utilize graphical displays to reveal potentially hidden or unusual patterns in your data suggested by mixed modeling. The combination of advanced statistical models and targeted graphics is a powerful one. In many cases you might want to do some graphical explorations in JMP before mixed modeling, and that is a great way to proceed as well.

Note it is also possible to analyze these data using *Analyze > Specialized Modeling > Matched Pairs*, which performs a classic paired *t*-test along with a rotated graph. However the four rows with missing values of Y are dropped and results are less efficient than those shown here. As one illustration of the difference, the degrees of freedom used to com-

Figure 1.4: Cell Viability Raw Data Plot



pute the p -values in the mixed model analysis are fractional (for example, 4.1 in the F test near the bottom of the output above). These are obtained using an advanced algorithm (Kenward and Roger, 1997) to more accurately approximate the small sample distributions of the statistics given the imbalance in the data due to the four missing values. A mixed model is able to handle missing data like this and deliver better results than a classical paired- t analysis.

In this example, the observed data that we analyze are the three columns in the Cell Viability table: the assigned levels of Chemical and Plate, and the response Y . All unknown parameters from the mixed model are estimated with REML using these quantities as inputs. Under key assumptions the estimated parameters enable us to make direct quantitative assessments of the causal effect of Chemical on Y amongst plate-to-plate and residual variability. Let's explore these assumptions in detail.

1.3 Mixed Model Assumptions

Several key assumptions are behind the validity of the preceding modeling results for the cell viability data. Continuing with this example as a prototype, we now describe the key statistical and structural form of these assumptions. We begin with a statistical description of a basic mixed model.

Statistical Mixed Model

$$y_{ij} = \mu + \chi_i + p_j + e_{ij}$$

response
intercept
chemical effect
plate effect $\sim N(0, \sigma_p^2)$
residual $\sim N(0, \sigma_e^2)$

$$\chi_i \perp\!\!\!\perp p_j \perp\!\!\!\perp e_{ij}$$

This is the simplest possible mixed model, with one fixed effect (Chemical) and one random effect (Plate). It is a linear mixed model, because it is an additive function of all primary components. The subscripts i and j index the individual observations; here $i = 1, 2$ and $j = 1, \dots, 9$, and y_{ij} is the response for the i th chemical on the j th plate.

Each term on the right hand side of the model contains unknown parameters that we estimate from the data. We adopt the convention here and throughout the book that Greek letters denote fixed effects and Roman letters denote random effects.

The first fixed effect is μ , which models the central tendency of the data, also known as an *intercept*. We expect its estimated value to be near the simple mean of Y . For the fixed effect Chemical, we specify two parameters, χ_1 and χ_2 , to model the effects of Chemical A and B, respectively. These are our primary parameters of interest for this experiment.

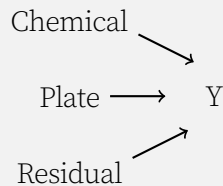
The notation $p_j \sim N(0, \sigma_p^2)$ is a shorthand way of stating that the random plate effect consists of independent and identical realizations from a normal (Gaussian) probability distribution with mean 0 and variance σ_p^2 . The errors e_{ij} have the same form of probability assumption and serve as a catch-all for the numerous, small, unobserved effects driving variability of Y within each plate, also known as residuals. The notation $\chi_i \perp\!\!\!\perp p_j \perp\!\!\!\perp e_{ij}$ denotes statistical independence (Dawid, 1979) among its three components. Even though χ_i are considered fixed unknown parameters, the independence here refers to the treatment assignment mechanism of the levels of χ_i to the half-plates.

This completes the formal set of assumptions that we make when viewing a mixed model as a *statistical* model, suitable for assessing associational relationships and for making predictions. We have defined the full conditional probability distribution of y_{ij} given all elements on the right hand side of the model.

Given our randomized experiment setting, we can readily move from association to causality and infer the causal effects of Chemical and Plate on Y . This entails viewing

the model as *structural* and assuming each term on the right hand side is *exogenous*, that is, wholly and causally independent of other variables in the system. We can depict this with a directed acyclic graph (DAG) as follows.

Structural Mixed Model



Note the direction of the causal arrows from the three causes to Y . Importantly, the absence of arrows into and between Chemical, Plate, and Residual indicates their exogeneity. Furthermore, the absence of any additional arrows into Y indicates there are no unmeasured causes or confounders besides those included in Residual. In addition, residual error is no longer just defined by algebraic subtraction, but consists of independent noise effects uniquely influencing each observed value of Y . Assumptions along these lines are required for causal inference. Refer to [Pearl \(2009\)](#), [Heckman \(2008\)](#), [Imbens and Rubin \(2015\)](#), [Hernán and Robins \(2020\)](#), and Chapter [10](#) for a comprehensive discussion.

It is critical that you fully understand the preceding modeling assumptions and their implications, keeping them in mind as you interpret modeling results. Strictly speaking, the assumptions may not be precisely true, but they do not need to be. As long as the assumptions provide a reasonably adequate approximation to the true data-generating mechanism, you can make sufficiently reliable associational and causal conclusions along with a statement of accompanying uncertainty.

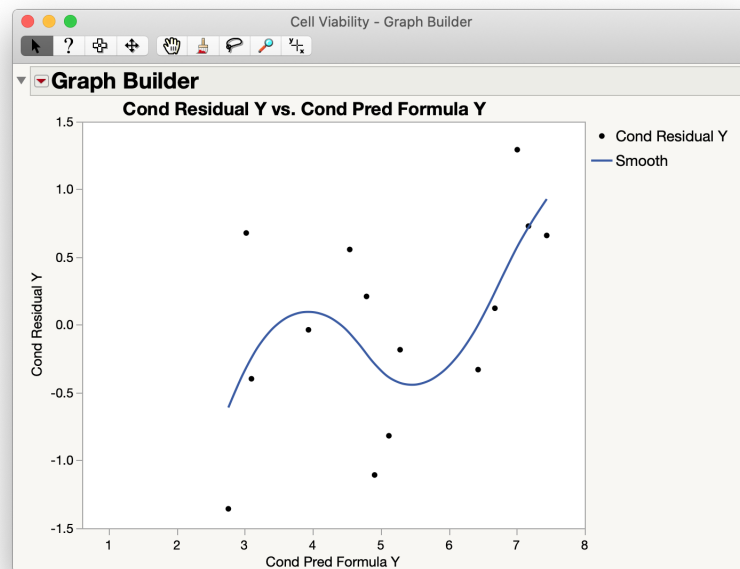
For the cell viability example, the assumptions on p_j and e_{ij} made above imply that y_{ij} is normally distributed with a well-defined mean and covariance structure, and the validity of printed t -statistics are made under this assumption. This model typically would not be appropriate for a response that is nonnormal (e.g. binary, count, or time-to-event), but you can handle such situations with extensions such as the generalized linear mixed model discussed in Chapter [9](#), or with transformations of the data that enable better alignment with the underlying assumptions. You are also free to only adopt standard statistical assumptions or go further and make causal ones, depending on the objectives of your analysis.

As we proceed with various examples throughout the book, we will indicate various ways of checking the aforementioned assumptions. The methods can be statistical or graphical, and often involve analyzing deviations from fitted model predictions. Some

assumptions, especially those for causal inference, are only indirectly or even fully untestable from observed data. For these you must rely on your scientific know-how and common sense, always maintaining a healthy degree of skepticism about how well your model approximates the true data-generating process of the system that you are studying.

As one example of the type of graph that is helpful for assumption checking, fit a mixed model on the **Cell Viability** data as in Section 1.2. Recall we left the previous analysis of these data wondering about a potential outlier. How well does our mixed model fit this value? Near the upper left corner of the analysis report, click the red triangle  > *Save Columns > Conditional Pred Formula*. From the same menu also save *Conditional Residuals*. Return to the Cell Viability JMP table and note two new columns have been added. Click *Graph > Graph Builder*, assign **Cond Residual Y** to the Y axis and **Cond Pred Formula Y** to the X axis, to obtain a graph like Figure 1.5.

Figure 1.5: Cell Viability Conditional Residuals



The residuals in Figure 1.5 are original Y values minus conditional predicted values (predictions including random effects). Graph Builder overlays a smooth curve by default. Note the range of the Y axis is from -1.5 to 1.5. This type of plot under usual mixed model assumptions should exhibit randomly scattered noise around a horizontal line at 0, with the fitted smooth curve also horizontal. Here the smooth curve is somewhat far

from that ideal. If you bring up the previous raw data plot side-by-side with this one, you can interactively select points in one plot and see them highlighted in the other. The outlier apparent in the raw data graph corresponds to the lowest left point in this plot. The three largest residuals in the upper right correspond to the three largest Y values. The mixed model predictions are shrunk somewhat towards the overall mean. This is a small data set with a fair amount of noise, and given the lack of fit, any final conclusions should be considered tentative.

1.4 Nominal and Continuous Variables

In the Cell Viability example from the previous sections, the Chemical and Plate variables in the JMP table are assigned modeling type *Nominal* (having named, discrete, unordered levels), as indicated by the red histogram icon , whereas Y is *Continuous* (having numerical values with an implied distance measure between them), as indicated by the blue triangle icon . The Nominal type of Chemical and Plate is crucial while modeling, as it notifies JMP to create distinct levels when constructing the parameters to be estimated. This is particularly important for variables like Plate, whose values in the table are numeric.

When first entering a numeric variable in JMP, it is assigned by default to be Continuous (blue triangle ). To change this attribute in any JMP table or dialog, click on the variable's icon and select the desired modeling type such as Nominal . A third possibility is Ordinal (green increasing histogram ) , but it is typically not needed for the common mixed models in this book.

Nominal or Continuous modeling type specification in JMP is in contrast to the use of a CLASS statement in various SAS/STAT procedures like PROC MIXED and PROC GLIMMIX. JMP has no CLASS statement; rather, you must prespecify effects to be Nominal or Continuous before specifying them in a modeling dialog. Note it is possible to create effectively identical models by converting a nominal variable to continuous ones with values 0 or 1, also known as indicator variables, dummy variables, or one-hot encoding.

Using independent variables that are continuous in a mixed model produces regression-style models. Such variables are often referred to as *covariates* or *regressors*. Standard regression analysis views such variables as fixed effects and the estimated parameters multiplying them in a linear model correspond to slopes. An effective mixed model extension of linear regression enables you to specify random slopes corresponding to meaningful clusters of the data, a type of model we refer to as *random coefficients*. These are a form of hierarchical linear models popular in social science and econometrics applications; see Chapter 5.

1.5 Experimental Units and Blocking, Cell Growth Example

When considering data from an experiment a fundamental question to ask is: On precisely what entities have treatment levels been applied or randomly assigned? In our

cell viability example, at first glance you might consider the entities to be the plates or the individual cells growing on them; however, neither of these is exactly right with respect to the Y responses. Such considerations involve the fundamental concept of an *experimental unit*.

Key Terminology

Experimental Unit The smallest entity to which a treatment is independently assigned. In the cell viability example, the experimental unit for Chemical is a half-plate.

Note there is no variable in the cell viability table for half plate even though it is the experimental unit for Chemical. This is because half plates correspond to the rows in the table itself and JMP and other common mixed modeling software is able to recognize this. Now suppose you subdivide samples into three replicates and triple the number of measurements and rows in the table. You would then want to add new columns like HalfPlate and Replicate to the table to designate the experimental units for Chemical and the replicate numbers, respectively.

While our cell viability example is relatively simple, the data and experiments that you commonly analyze are likely more complex. Consider the data in Figure 1.6, which represent an extension of the cell viability experiment and are available in `Cell Growth.jmp`.

The Plate, Chemical, and Y data values in `Cell Growth` are identical to `Cell Viability`, but there are now two additional factors: Incubation and Batch. Our experiment objectives are extended to investigate the effect of three different incubation periods (short, medium, and long). In addition, the incubation chamber has room for only three plates, and the Batch variable indicates which plates are incubated together.

Given what we know so far about mixed models, how would you designate the new factors Incubation and Batch with regard to being fixed or random effects? The three levels of Incubation are ordered and constant for this experiment, and we want to directly compare how they change Y. Incubation is thus considered to be a fixed effect. Furthermore, notice that levels of Incubation are applied to entire plates, so the experimental unit for Incubation is a plate. We therefore now have two different sizes of experimental units: plates and half-plates. This arrangement is known as a *split-plot design*, which we cover in detail in Chapter 3.

What about Batch? We can naturally consider the effect of a particular run in the incubation chamber to be transient and sampled from a theoretical population of such runs. Batch is therefore a random effect. The fact that Batch groups sets of three plates brings us to another very important concept in experiment design.

Figure 1.6: Cell Growth Data

	Incubation	Batch	Plate	Chemical	Y
1	short	1	1	A	3.7
2	short	1	1	B	3.8
3	short	2	2	A	•
4	short	2	2	B	1.4
5	short	3	3	A	5.1
6	short	3	3	B	6.1
7	medium	1	4	A	5.1
8	medium	1	4	B	7.9
9	medium	2	5	A	3.9
10	medium	2	5	B	•
11	medium	3	6	A	•
12	medium	3	6	B	8.1
13	long	1	7	A	4.3
14	long	1	7	B	8.3
15	long	2	8	A	2.7
16	long	2	8	B	•
17	long	3	9	A	5
18	long	3	9	B	6.8

Key Terminology

Block A group of experimental units that are similar in some fashion, distinguishing them from other groups of experimental units. In the cell viability example, each level of Plate defines a block consisting of two half-plates. In the cell growth example, Batch defines blocks with three plates each.

When designing and analyzing your experiments, there are several reasons why you might want to create blocks or batches. Experimental units may naturally occur in clusters or groups. For example, in the cell growth data above, plates naturally group half-plates, and the incubation chamber size restriction requires us to use batches of three plates. Another type of blocking example is different shipments or lots of a raw material or reagent from a supplier. Blocking typically enables you to control for known sources of variability and obtain more precise inferences on the fixed effects.

A good question to ask is: What features of your experiment are the key sources of variability and covariability in the responses from one observation to the next? Identify these features and then model their effects by, for example, creating batches to con-

solidate the groups of similar runs. Now any differences between the batches will give you a measurement of the variability that comes from this source, and you can better estimate uncertainty around your estimates for the treatment effects once you've accounted for it. This is the purpose of a well-thought-out design—partition out the variability that can't be controlled (but can be explained), and then you will have more precision for the rest of your conclusions.

Blocking factors can have many different names. Blocks could be called *lots* or *batches*, or they could be called by the effect they are modeling, such as person or school. *Strata* is another popular term. The point is simply that you are grouping the experimental units so that they are more similar within groups and less similar (often due to effects that you cannot control) between groups.

Occasionally blocks can account for significant dissimilarity. A classic example is a litter of pups from a common dam who are competing for a fixed amount of resources (food or familial attention). In this type of situation, we might expect the measurements within a block to be negatively correlated, and in the mixed model analysis, the estimated variance component would be negative.

Blocking can often be very effective when applied in two different directions. Common examples include rows and columns in a field plot or 96-well plate. Chapter 4 contains an example of a *row-column design* (Latin Square) implementing this concept. This type of blocking also relates to general considerations of *spatial variability*; see Chapter 7.

Another type of blocking can occur when you observe repeated measurements on the same experimental units, for example, in a longitudinal or time series study on individuals. Repeated measures are usually no longer independent of each other, and a mixed model is a great way to handle this source of correlation. New complexities arise, as there can be different levels of measurement and different types of covariance structures; see Chapters 5 and 6.

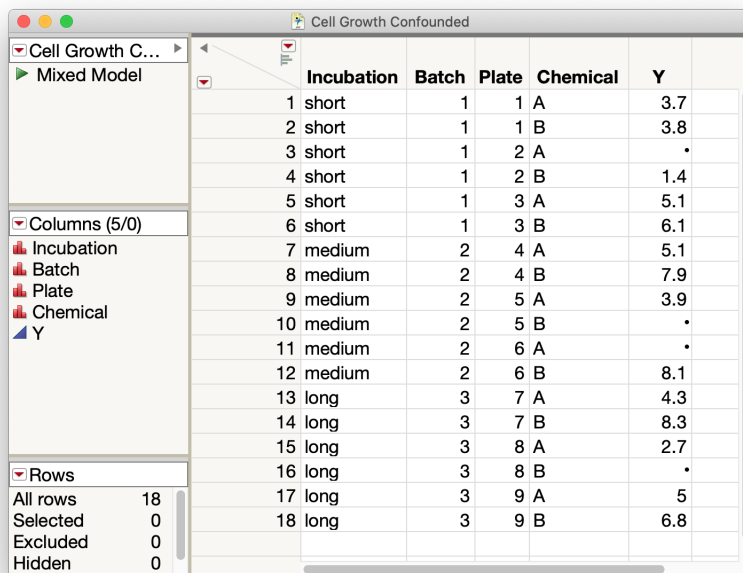
In general the goal of blocking is to control for specific sources of variability and thereby achieve more accurate and precise inferences. Blocking can often make considerable difference in modeling results. Because blocks are usually assumed to arise from a population of effects, you typically will want to declare all blocking factors as random effects in your mixed model analysis.

1.6 Confounding

Suppose the data for the cell growth data are the altered version in Figure 1.7.

The final three columns in Figure 1.7 are identical to the Cell Growth data in Figure 1.6, but the first two columns are different. Can you spot the problem?

Figure 1.7: Cell Growth Confounded Data



	Incubation	Batch	Plate	Chemical	Y
1	short	1	1	A	3.7
2	short	1	1	B	3.8
3	short	1	2	A	•
4	short	1	2	B	1.4
5	short	1	3	A	5.1
6	short	1	3	B	6.1
7	medium	2	4	A	5.1
8	medium	2	4	B	7.9
9	medium	2	5	A	3.9
10	medium	2	5	B	•
11	medium	2	6	A	•
12	medium	2	6	B	8.1
13	long	3	7	A	4.3
14	long	3	7	B	8.3
15	long	3	8	A	2.7
16	long	3	8	B	•
17	long	3	9	A	5
18	long	3	9	B	6.8

Key Terminology

Confounding The indistinguishability of two or more effects, given a model and data set. Such effects are *confounders* of each other. In the example data above, Incubation and Batch are *complete* confounders. Confounding can also be *partial*, in which portions of effects are unable to be disentangled from portions of other effects given the model and data.

Confounding is a danger lurking in many statistical and structural models of data, and you should be constantly on the lookout for it. Sometimes it is innocuous, as in the case we have already encountered involving the parameters μ , χ_1 , and χ_2 in the simple mixed model for the cell viability data. We resolve this by imposing the sum-to-zero constraint $\chi_1 = -\chi_2$. Similar constraints typically work for complex linear effects and their interactions, ensuring identifiability of the model.

The preceding example, however, is much more serious. The data provide no way to separate the effects due to Incubation and Batch. Worse would be a case where only one of the variables is observed, leading to likely incorrect conclusions about the true

magnitude of that effect. The mistake here is in the experiment design, and a principal goal of good design is to avoid confounding like this. Note in the original Cell Growth data in Figure 1.6, Incubation and Batch are nicely *orthogonal*, with each incubation level occurring exactly once within each batch.

In many data sets, especially observational ones, partial confounding is unavoidable. When confounding happens, the best you can do is understand its precise nature, determine exactly how effects are aliased, and limit your conclusions appropriately. Refer to [Hernán and Robins \(2020\)](#) for helpful insights on confounding in the context of causal inference, including such difficulties as unmeasured confounders.

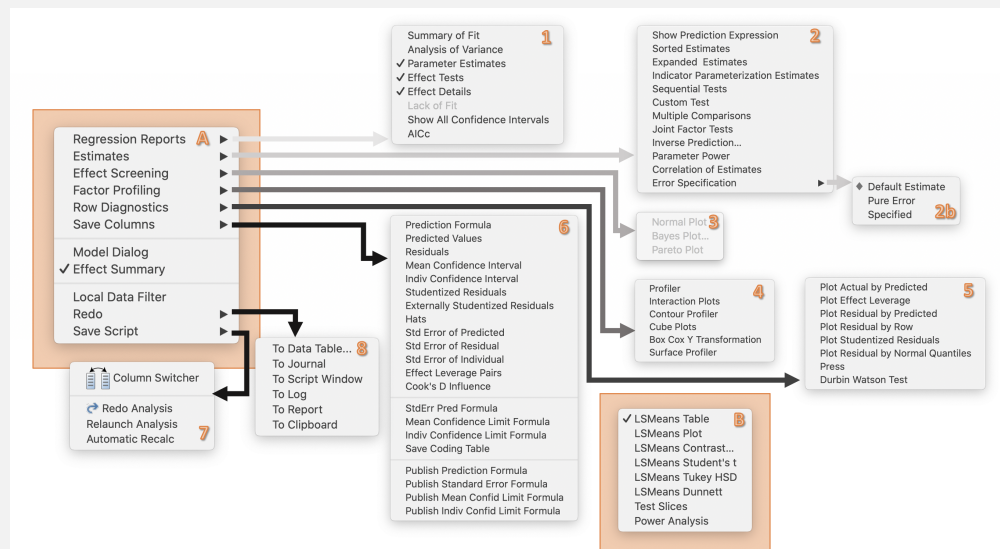
1.7 JMP and JMP Pro

A more advanced version of JMP is available as *JMP Pro*. The *Fit Model* platform in JMP Pro adds a dedicated *Mixed Model* personality, and specifying mixed models in it is a bit different from the *Standard Least Squares* personality we use above in the cell viability and growth examples. You can optionally still use Standard Least Squares in JMP Pro. Certain mixed model functionality is only available in JMP Pro, including random coefficient models (Chapter 5), repeated measures models (Chapter 6), and spatial models (Chapter 7).

The following two boxes provide breakdowns of functionality available in the Standard Least Squares and Mixed Model personalities.

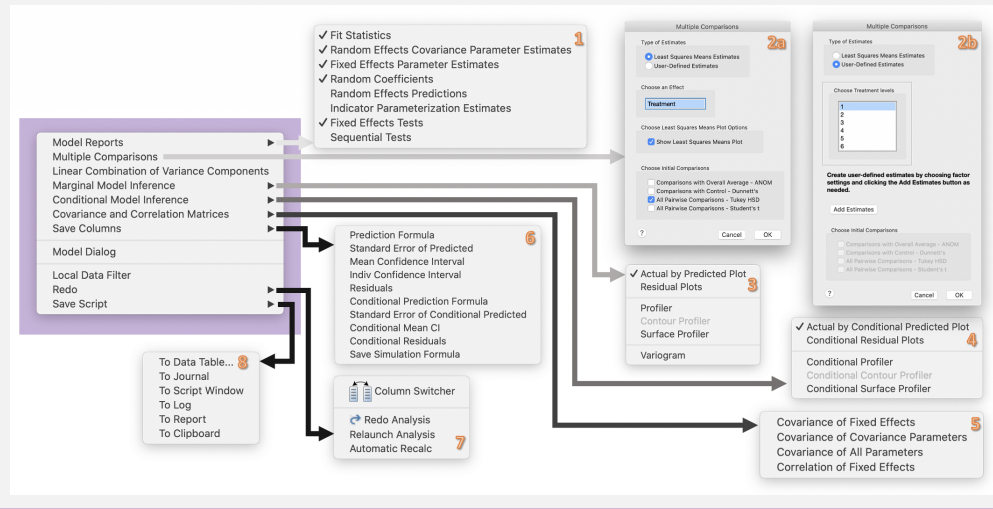
Using JMP

Once you run an analysis in JMP, you will find many drill-down options from the red triangle at the top left of the results report window. When using the *Standard Least Squares* personality, you will find the following options from the drill-down menu in the results report. For more information, go to <https://www.jmp.com/help> and search in the *Fitting Linear Models* section.



Using JMP Pro

Once you run an analysis in JMP Pro, you will find many drill-down options from the red triangle at the top left of the results report window. When using the *Mixed Model* personality, you will find the following options from the drill-down menu in the results report. For more information, go to <https://www.jmp.com/help> and search in the *Mixed Models* section.



1.8 Exercises

1. In the output for the mixed model analysis of the cell viability data in Section 1.2, review each box of results and briefly describe the statistics displayed in each. Include question marks for the ones you do not yet understand.
2. In the cell viability data from Section 1.2, exclude the largest outlier for Chemical B and rerun the analysis. What type of difference does this outlier removal make in the results? Under what conditions would such exclusion be justified?
3. In the cell viability data, exclude or delete the rows with missing Y values and rerun the analysis. How do results change? Now analyze the data with *Analyze > Specialized Modeling > Matched Pairs*. How do the results from this analysis compare the previous one?
4. Fit an appropriate mixed model to the cell growth data from Section 1.5 and interpret key results.
5. Fit a mixed model to the confounded cell growth data from Section 1.6 and compare results with the previous exercise. What is the effect of confounding?

Ready to take your SAS® and JMP® skills up a notch?



Be among the first to know about new books,
special events, and exclusive discounts.

support.sas.com/newbooks

Share your expertise. Write a book with SAS.

support.sas.com/publish

Continue your skills development with free online learning.

www.sas.com/free-training



sas.com/books
for additional books and resources.



SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.
Other brand and product names are trademarks of their respective companies. © 2020 SAS Institute Inc. All rights reserved. M2063821 US.1120

Copyright 2021, SAS Institute Inc., Cary, NC, USA. All Rights Reserved.
Reproduced with permission of SAS Institute Inc., Cary, NC