Title:
Statistical Methods for Dose-Response Assays

Author:
Tong, Frances Poyen

Acceptance Date:
2010

Series:
UC Berkeley Electronic Theses and Dissertations

Degree:
Ph.D., StatisticsUC Berkeley

Advisor(s):
Speed, Terence P

Committee:
Huang, Haiyan, Ngai, John, Nolan, Deborah

Permalink:
http://escholarship.org/uc/item/9295t422

Abstract:

Statistical Methods for Dose-Response Assays

by

Frances Poyen Tong

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Terence P. Speed, Chair
Professor Haiyan Huang
Professor John Ngai
Professor Deborah Nolan

Spring 2010

Statistical Methods for Dose-Response Assays

Abstract

Statistical Methods for Dose-Response Assays

by

Frances Poyen Tong
Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Terence P. Speed, Chair

Dose-response assays are a common and increasingly high throughput method of assessing the toxicity of potential drug targets on test populations of cells. Such assays typically involve serial dilutions of the compounds in question applied to cell samples to determine the level of cell activity across a broad range of concentrations. Another factor in such experiments may be the change in activity under different enzyme combinations and the use of controls to adjust for interassay variations. Typically, the decreasing number in the population of cells due to increasing concentrations of the drug can be modeled with a logistic curve. Since the appropriate range of concentrations of the drug to test in order to see these reactions cannot be predetermined fully, frequently the data available for a given experimental unit may not be enough to fit such curves on their own successfully. Instead, the assay data as a whole can be successfully analyzed with methods such as constrained fitting and mixed effects models, where each set can borrow strength from each other in order to be fitted while still taking into account individual significances of the specific experiment.

This dissertation illustrates variations of such methods on three major datasets from the Joe Gray lab at Lawrence Berkeley National Laboratory, the Douglas Clark lab at the Department of Chemical Engineering at UC Berkeley, and Bionovo, Inc. of Emeryville, California. The first dataset from breast cancer cell line testing involves the estimation of the National Cancer Institute concentration parameter called GI50, the concentration of the drug at which it inhibits the growth of the population of cells by half. We develop a method utilizing replicate data to estimate this parameter. The second dataset involves the estimation of a more commonly used concentration parameter called the IC50, which doesn't take into account the initial cell population, on special assays that mimic the liver metabolism in the body. The methods involve mixed effects models that incorporate the specific enzyme conditions and types of cells important to the experiment. The third dataset, involving different effects of plant compounds on an osteosarcoma cell line, illustrates the usage of negative and positive controls to appropriately adjust the observations for interassay variation.

To my parents

Contents

Acknowledgments

Foremost thanks goes to my advisor Terry Speed, who has helped me to become a better statistician than I had ever hoped to be. I am continually in awe of the way he is able to see right through the most complex problems lightning fast and suggest clear and to the point solutions that ultimately benefit the progress of science. It has truly been an honor to learn from him these past six years.

I'm also grateful to my committee members, Professors Haiyan Huang, John Ngai, and Deb Nolan, for taking time out of their busy schedules to advise and guide me the past couple of years. I'd especially like to thank Deb for having such faith in me to include me in several of her wonderful statistical computing education projects. I've learnt tremendously from her and Duncan Temple Lang and look forward to working with them this summer and beyond.

A humble thanks to all the collaborators who entrusted a young naive graduate student with their precious data: Paul Spellman from Lawrence Berkeley National Laboratory, Jessica Ryan from Douglas Clark's lab, and Xiaoyue Zhao of Bionovo, Inc. Their patience in allowing me to learn from their data and waiting for results through the years was truly appreciated.

I have known for the past six years how lucky I am to have been part of the statistics department and be surrounded by the greatest minds in the field. The cherry on top is the fact that everyone in this department is also extremely nice. I am eternally honored that the late Erich Lehmann and his beautiful wife Juliet Shaffer were so kind to me throughout my years in the department.

Life in graduate school was not possible without the caring and dependable support of the staff. Much appreciation goes out to all the current and former staff members throughout the years who have been a smiling and encouraging face in the front office, who cheerfully set up tea time every Tuesday afternoon for us hungry graduate students, who plan wonderful welcome and holiday parties, and who do so much for me in the background so that I don't have to worry about much except doing my research. It felt great to be so well taken care of. Special thanks to Debbie Haaxman, Denise Yee, Ryan Lovett, and Phil Spector who have gone above and beyond to help me. I'd also like to reminisce here of all the times when I would get frustrated about graduate school, Pat Hardy would always be there to give me a hug and tell me what I was doing was amazing.

The most important person whom I will miss in the department is Angie Fong. I can not even begin to enumerate all the things that she has done for me these past six years. Knowing that she always took good care of me lessened the stress of grad school considerably. I truly enjoyed all of our wonderfully interesting chats; Angie always knew how to make me feel better.

I would like to thank my two former UCLA professors who got me started on this path to a PhD in statistics. In 1999, Dr. David Kan taught me the beauty of mathematics and computing at a time when I was considering double majoring in human genetics and economics. In 2002, Dr. Nicolas Christou taught me the beauty and importance of statistics at a time when I had naively refused to take any statistics courses in my undergraduate career. Nicolas has been a wonderful mentor to me all these years.

I definitely would not have survived all these years without the love and friendship of the nicest and smartest people I could ever hope to meet in this lifetime.

Margaret Taub is one of the most important reasons of how I have been able to get through graduate school and be lucky enough to be writing acknowledgments for a completed dissertation at this very moment. I feel so blessed to have been her friend from the moment I started. The department did a great thing in having me stay with Margie when I visited; she sure could reel in all the new students with her sweetness. I truly miss being down the hall from Margie and hearing her laughter, and being able to chat with her whenever and about whatever. She gave me such great support and encouragement throughout our graduate school years, and I am so happy that I will soon be able to dance at her wedding to Kasper!

I have the best officemate and friend ever in Irma Hernandez. Who could not love Irma from the first time you meet her? It has been wonderful receiving daily hugs and being able to work next to her all these years. Irma has been there to support me through all the ups and downs of graduate school. I am extremely grateful that she and her husband Erick so generously allowed me to stay with them everytime I flew back to Berkeley this past year to meet Terry on one of his short trips back to the Bay Area. I could not have been better cared for during this time of stress. I am looking forward to more fun times with such a sweet couple.

I regret that I realized so late what an amazing person Victoria Wang is. Victoria is such a sweet and hilarious girl; I have loved hanging out with her in Berkeley and hope to see her soon in Boston.

It has been such a pleasure knowing Allan Sly, such a funny guy under the shy exterior! Allan has been there for me for so many things, and I can always count on him to continue to be the super nice and smart guy that he is now. Hanging out with him has always been so much fun, and it will surely continue no matter where we are in the future.

The first two years of graduate school would not have been the same without the close friendships with Trinh Pham, her husband Keiji Oenoki, and Jamie Hollander. They are the nicest people I have ever met, and I hope to remain in touch for years to come.

Other fellow graduate student friends have made life in Berkeley fantastic. I am happy and proud to have known them: Charlotte Wickham, Cathy Tuglus, Richard Liang, Brad Luen, Daniel Ting, Harry Kim, Guilherme Rocha, Peter Ralph, Kasper Hansen, Vince Vu, Greg Hather, Soyeon Ahn, and David Rosenberg.

My best friends from high school and college have been there to support me for so many years. Tammy Chen and I have been through so much together since high school and UCLA. I am in awe of the amazing job she is doing as an emergency room doctor; it makes writing this dissertation look like a piece of cake. Cynthie Wong has been the sweetest friend ever since I met her in the bathroom in junior high. I have enjoyed all our chats through the years and wish her good luck and congratulations in finishing her PhD this summer. Linda Ho has been the glue holding all of us friends together. She has been wonderful in checking up on me to see how I have been doing all the years in graduate school. Florence Tse is one of the coolest friends I have; she has been fantastic in always visiting me everywhere I move to. Now that I am finally about to finish school after 25 years, I can't wait to travel with them!

It has been really great having a brother so close in age to me. Calvin Tong is the best little brother one could ever hope for. He has been the best in helping me while I have been a poor stressed out graduate student. Without his expert advice and constant IT support, the work in this dissertation would have hit much rockier roads. I am looking forward to spending more time with him on trips in the future.

I would be nowhere in this life without the sacrifices that my parents have made to ensure that I have the best possible opportunities in life. They have worked extremely hard their entire lives so that I could put all my efforts into studying and pursuing my interests. Every time I have called to whine about school, they have put things in perspective for me and made me feel better. I have been waiting anxiously for the moment when I will be able to start taking care of them and let them finally relax after all these years. I dedicate this dissertation to them.

Without a doubt, the best thing I have gotten out of graduate school is Shankar Bhamidi. He is the best partner in life I could ever have dreamed of, and I feel so extremely lucky that I have such an amazing and wonderful guy to spend the rest of my life with. It is quite a feeling of excitement, peace, and love that I have when I realize that we will be together for everything that is good and bad in this life. He shows such enthusiasm for so many things that I take for granted, and he makes everything so much more fun! I hope we will always be excited about figuring out ways to help the world, and that we actually do someday. Graduate school would not have been the same without him stalking my office all the time. I thank him tremendously for taking such good care of me as I try to finish this little thing called a PhD. You're the bestest!

Chapter 1

Introduction

> *All substances are poisons: there is none which is not a poison. The right dose differentiates a poison and a remedy.*
>
> - Paracelsus [1493-1541]

The pharmacological treatment of cancer and other diseases is complex. One drug does not fit all, not even most. There are potentially hundreds of thousands of compounds out there that can help in some way, but how do we weed through all the ones that don't and how do we determine who gets what, when, and how much?

For a compound to survive the long process to become an useful drug, years and years of research and an enormous amount of money are required to test it and to ensure its efficacy and most importantly, safety. The beginning stages of finding potential drugs to focus on is often laborious and disappointing. Many seemingly good candidates can later prove to be toxic to people. This step of the drug discovery process is continually being sped up with better and better technologies that can determine the reactions of many compounds on different cell samples at the same time. The majority of these methods involve dose-response assays. These assays are typically either plates or chips that can interrogate the effects of varying concentrations of these compounds with a variety of cell types under different conditions.

Dose response assays can be described as mini laboratories condensed onto small, typically rectangular plates. While a good many varieties produced by biotech companies and university research labs exist, the common purpose of such a tool is to determine the status quo of a living organism outside of itself. In a way, it allows testing of something hidden that may otherwise cause discomfort or injury to the organism, a human patient in our case.

1.1 Drugs

A drug is simply defined as a substance that affects normal bodily function when absorbed in a living organism. More specifically for the pharmacological purposes of this dissertation, a drug is a chemical substance that is supposed to prevent, diagnose or treat an illness or a disease.

If the drug causes a response, then it is known as an agonist. The compounds involved in reducing hot flashes of menopausal women in Chapter 4 are agonists. On the other hand, if it does not cause a response but blocks an agonist-mediated response, it is called an antagonist. Cancer drugs (Chapter 2) are often antagonists that try to inhibit the activation of signalling pathways in cells. Other antagonists are involved in the toxicity testing done in Chapter 3, where a large variety of compounds are tested with liver cells to see whether they will become toxic inside the body. For example, Tolcapone is a drug used to treat Parkinson's that can cross the blood-brain barrier to inhibit a specific enzyme.

1.1.1 Tykerb

An example of an antagonist drug involved in one of the datasets of this dissertation is Tykerb, a drug produced by GlaxoSmithKline. Also known as lapatinib, it is a small molecule dual inhibitor of both EGFR and HER2 that was recently approved as a first-line treatment of metastatic breast cancer patients who are categorized as triple positive, meaning ER+ / EGFR+ / HER2+.

The family of signalling ERBB/HER receptors is an important target of many cancer drugs. Found in the cell membrane, they are four tyrosine kindase receptors, ERBB1-4, also known as HER1-4, in which their basic functional unit in signalling is a receptor dimer, formed from two of the same or different receptors after ligand binding. One receptor can activate its partner through an allosteric mechanism which rearranges and stabilizes both receptors.

While ERBB2 does not bind to any ligands, it is the preferred heterodimeric partner of the other three receptors. As illustrated by Figure 1.1, lapatinib competitively binds to the ATP binding pocket of the EGFR/HER2 protein kinase domain, thus preventing the receptors' self-phosphorylation, which in turn, disrupts the activation of two major signalling pathways responsible for major cell activities such as survival, growth, and proliferation.

1.2 Dose-response curves

A good informative assay will have the sufficient range of concentrations that cover three areas: low concentrations to show the level at no inhibition, concentrations where the reaction is happening, high concentrations to show the level at 100% inhibition. Usually such data will follow a curve that has a sigmoidal “S” shape (Figure 1.2).

The steps from the drug binding to a receptor and the final response that we measure on these assays are numerous and complex. However despite this, systems

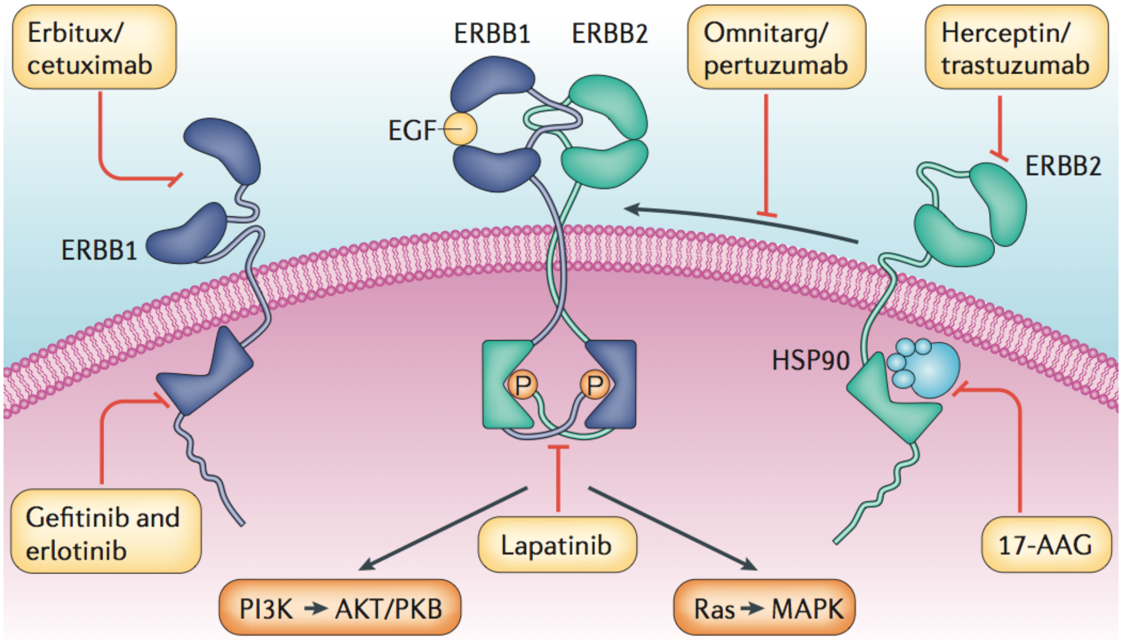


Figure 1.1: Copyright ©Nature Publishing Group [7]. Diagram of various cancer drugs and how they affect the signalling pathways in the cell.

commonly produce dose-response curves that are sigmoidal. The use of logistic curves as an empirical model works sufficiently well in a majority of the cases. This perhaps can be explained simply by the fact that there is a direct link between receptor binding and the response so what we are seeing is a response proportional to receptor binding, assuming that each intermediate step follows the law of mass action.

These "S" shaped curves are typically defined by 4 or fewer parameters: the lower and upper asymptotes define the minimum and maximum response while two more parameters, the slope and the x value at the midpoint between the lower and upper asymptotes define the linear portion of the activity. While the logistic curve is the most popular, another called the gompertz curve is also sometimes used [4]. The difference is that while the logistic curve is symmetric about the mid point, the gompertz curve can be asymmetric and sometimes that is a more flexible option for the data (Figure 1.3).

1.3 EC50, IC50, GI50

In typical dose response data like in Figure 1.2, the standard summary of the data points is to find the concentration of the drug at which the response is effected or inhibited by half. The EC50 is the concentration at which half the maximal effect is observed; this is the usual measure for agonist assays. The IC50 is the concentration of

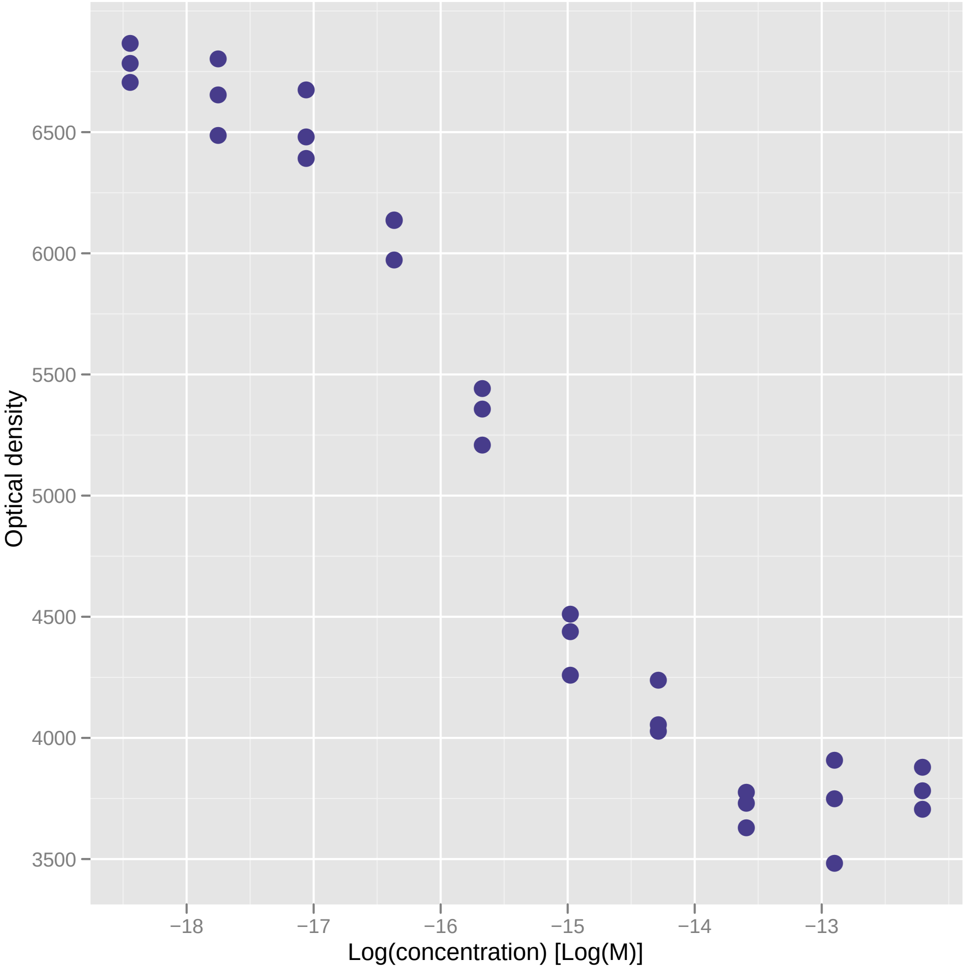


Figure 1.2: Example of a unit of data. As the concentrations of the drug increase (to the right), the more the growth is inhibited in the cell population, as measured by the optical density from the scanner.

the compound or drug that is required to effectively inhibit biological or biochemical function (in vitro) by half. The GI50 is similar to the IC50; it is the concentration at which the growth of the cell population is inhibited by half.

1.4 Outline of the thesis

This dissertation develops and illustrates a few key methods in the analysis of dose-response data. Table 1.1 is a reference to the different characteristics of each of the three main datasets involved in this dissertation.

All plots in this dissertation except figures obtained from external sources were

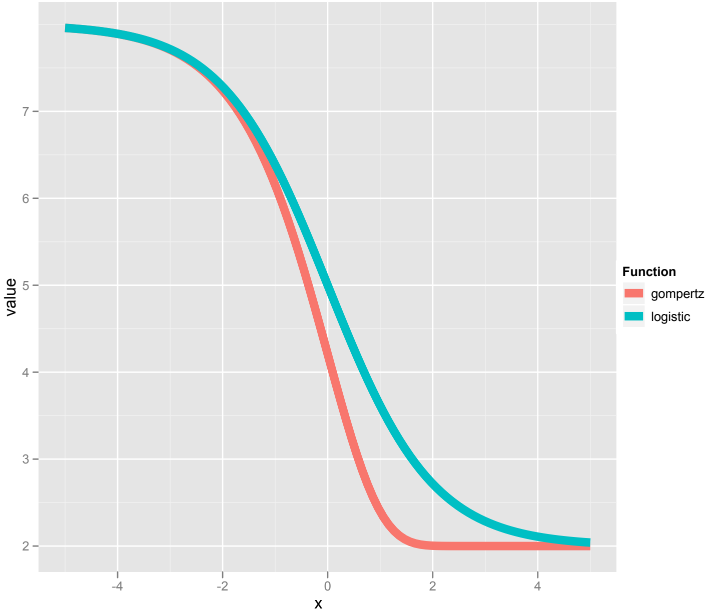


Figure 1.3: Example of gompertz and logistic curve with the same parameter values, $\phi_0 = 2$ (lower asymptote), $\phi_1 = 8$ (upper asymptote), $\phi_2 = 0$ (IC50), and $\phi_3 = -1$ (slope).

generated by the ggplot2 package by Hadley Wickham [21].

Chapter 2

Chapter 2 discusses the estimation of a specific National Cancer Institute measure called the GI50, the concentration of a compound required to inhibit half the growth of the cell population. The dataset on breast cancer cell lines involved in this chapter is provided by the Joe Gray laboratory at the Lawrence Berkeley National Laboratory. This chapter serves to illustrate the method of constraining parameters in a model to fit all replicate curves simulataneously.

Chapter 3

Chapter 3 introduces the usage of nonlinear mixed effects models to fit larger sets of data according to specific common factors such as enzyme conditions or chips. The dataset on testing compounds for toxicity involved in this chapter is provided by the Douglas Clark laboratory in the Chemical Engineering Department of UC Berkeley.

Chapter 4

Chapter 4 extends nonlinear mixed effects models to incorporate external information about the dilution series data such as controls to assess the intra- and inter-plate variation. The methods in this chapter are illustrated with data provided from Xi-

	LBL	MetaChip	Bionovo
Doses	9	9	10
Dilution fold	2 - 5	3, 4	variable
Replicates per dose	3	5	3
Size of dataset	variable	variable	540 obs.
Plate / chip size	96 wells	1080 spots	96 wells
Cell origins	Human breast cancer cell lines	Human and rat liver tissue	Human osteosarcoma cell line
Control values	on separate plates	none	on same plate
Type of controls	positive	none	negative and positive
Enzymes	none	yes	none
Goal summary	GI50	IC50	EC50

Table 1.1: Table of the different characteristics of each of the main datasets. (LBL = Lawrence Berkeley National Laboratory).

aoyue Zhao of the Bionovo Corporation in Emeryville, California.

Chapter 5

Chapter 5 reflects upon the analyses in the three core chapters, Chapters 2, 3, and 4, as a whole and discusses their similarities and differences in methods and data. Advice and issues regarding the general usage of such methods on dose-response data is also included.

Chapter 2

Estimation of GI50: Constrained fitting of replicate data

2.1 Introduction

A common measure of drug effectiveness is the IC50. This is the inhibitory concentration of a drug that will eliminate half the population of cells. Since the IC50 does not take into consideration the initial cell population at time zero, the National Cancer Institute developed three new special concentration parameters to improve the measure of drug effectiveness [1].

In a typical cell viability experiment, the initial samples from cell lines are given drugs and then measured after 72 hours. For a given cell line and drug, the GI50 (growth inhibition) is the concentration of the drug that kills half of the growth from the initial population of cells, while the TGI (total growth inhibition) is the concentration that eliminates the all the growth. If the drug continues to kill into the initial cell population, the LC50 (lethal concentration) is the concentration of drug that, in addition to killing off all the growth, eliminates half of the initial population. The difference between the calculation of these measures and the more common measure of IC50 is that additional information is required, such as the initial cell count at time zero and the cell count with no drug after some time t. The specific formulas are included in Table 2.1, where the three measures are the concentrations in the log scale that satisfy the equations.

Current methods of dealing with replicate data in growth inhibition analysis that others use involve computing summary statistics of independently calculated GI50s from each replicate, i.e. taking the mean or median of the resulting GI50s [1]. This chapter discusses methods to include all replicate data in one model in order to generate a growth inhibition curve that will be more robust than fitting each replicate curve separately.

Growth inhibition		
GI50 : the concentration of drug that causes 50% growth inhibition	$\dfrac{f(x_{\mathrm{GI50}})-\alpha}{\phi_1-\alpha}=0.5$	(2.1)
TGI : the concentration of drug that causes 100% growth inhibition	$\dfrac{f(x_{\mathrm{TGI}})-\alpha}{\phi_1-\alpha}=0$	(2.2)
Lethal concentration		
LC50 : the concentration of drug that kills off 50% of the initial population	$\dfrac{f(x_{\mathrm{LC50}})-\alpha}{\alpha}=-0.5$	(2.3)
f = function relating dilution to growth α = initial cell count at time zero ϕ_1 = cell count with no drug after time t		

Table 2.1: Table of NCI definitions of GI50, TGI, and LC50; they are the values of x such that the equations on the right are satisfied.

2.2 Data

Significant work in the testing of scores of compounds on scores of breast cancer cell lines has been conducted in the lab of Joe Gray at the Lawrence Berkeley National Laboratory (LBL). All data for the breast cancer cell line viability is procured using the Promega CellTiter-Glo® Luminescent Cell Viability Assay. The basis of such an assay relies on the fact that living cells contain and require ATP to remain metabolically active [8]. A chemical reaction involving ATP and luciferin, a compound that lights firefly tails, will produce light, and thus enables the quantitation of the amount of ATP (Figure 2.1). The observed response is the optical density of bioluminescence that is proportional to the amount of ATP in each well. Thus in reality, the number of viable cells is measured only relatively (Figure 2.2). An example of the correspondence between the intensities on the image and the quantitative measure of the intensities is included in Figure 2.3.

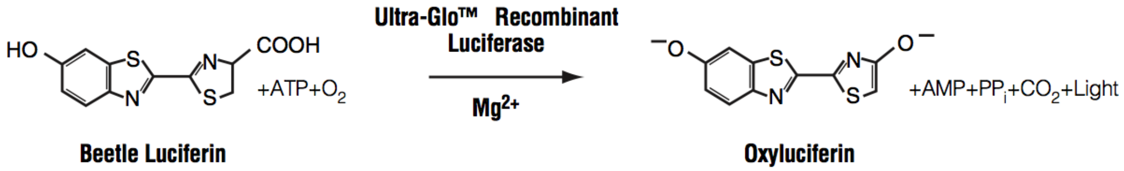


Figure 2.1: The luciferase reaction: Mono-oxygenation of luciferin is catalyzed by luciferase in the presence of Mg2+, ATP and molecular oxygen. [17]

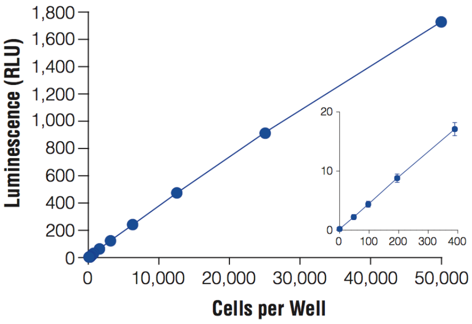


Figure 2.2: Cell number correlates with luminescent output. See [17] for more details.

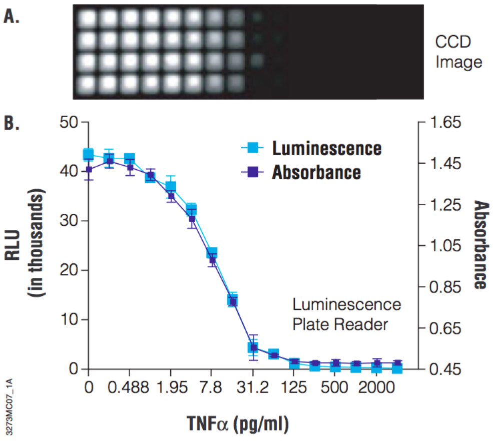


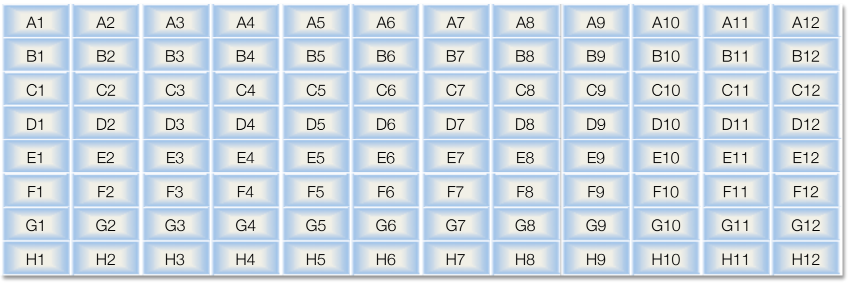
Figure 2.3: Example of scanned image with plot of intensities.

The data consists of two corresponding groups of 96-well plates. The first group are the plates where the drugs are tested on cell line samples. One sample is typically tested with three drugs on one of these drug plates. For each combination of cell line and drug on a plate, a series of 9 dilutions of the given drug is administered with 3 replicates each. With three types of drug, these observations of the relative cell counts 72 hours after drug administration account for 81 wells on the plate. The other 15 wells are used to measure 3 replicates of vehicle controls, the number of living cells in media after growing for 72 hours without any drugs; 4 replicates of untreated control, which is only media; 2 blanks; and 6 dilutions for standardized ATP controls, which are used to estimate the ATP standard curve. In this chapter, we refer to each set of 27 observations of a specific replicate of a cell line and drug combination as units. The plot of the raw values of an example unit is included in Figure 2.5.

The serial dilutions can be of any fold-difference apart, but are usually constant within a unit. A lab may typically start with a larger range of concentrations first to see where the range of activity may lie, and then repeat the experiment with a narrow and more focused range of concentrations to maximize the information. The dataset here involves dilutions of two- and five-fold.

Corresponding to each of these drug plates is a plate from the second group, called the T0 plate, as it contains the data for the number of cells at time 0 hours. Along with 2 blanks and 6 dilutions for standardized ATP controls, the rest of the 88 wells are used for T0 values.

The locations of each of the units' measurements are randomized throughout the data plate, although each plate follows the same mapping. This mapping, shown in Figure 2.4, reveals how the 96 wells are assigned to the 81 drug-related measurements as well as the various control values and ATP standards. The first column of the T0 plate is reserved for the six dilutions of the ATP standard, while the remaining 88 interrogate the same thing.



A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12
B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12
C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12
G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12
H1	H2	H3	H4	H5	H6	H7	H8	H9	H10	H11	H12

	Drug 1			Drug 2			Drug 3		
Dilution	**#1**	**#2**	**#3**	**#1**	**#2**	**#3**	**#1**	**#2**	**#3**
1	D4	F10	H12	C12	E12	E4	B9	A12	E2
2	D3	D10	G7	F9	G12	E9	C10	F8	B5
3	H7	A8	B4	B2	E3	B8	E7	H9	C4
4	F3	E10	H3	C7	G2	A11	B11	D2	A4
5	E5	F5	A10	A5	F4	C3	F11	A6	E8
6	F7	B3	H4	D7	A9	F12	A7	A2	D11
7	D12	C2	D6	F6	G9	B6	G11	G4	D5
8	G6	E6	G8	G3	B12	H5	H6	F2	H11
9	D8	B7	H2	C6	B10	C5	C11	G10	C9

Blank	A1	B1				
Vehicle control	D9	H8	A3			
Untreated	H10	G5	C8	E11		
ATP	C1	D1	E1	F1	G1	H1

Figure 2.4: Mapping of the drug plate.

A PostGreSQL database on the LBL servers currently contains data for 6,656 plates and 16752 cell line by drug combinations. All data can be easily accessed and analyzed through the R package cellvia (currently still under improvements).

2.2.1 Concentration series

To convert the information in a table like Table 2.2 for use in working with the replicates of any specific cell line and drug combination, we need to calculate the starting concentration of the drug used. This can be found by multiplying the stock concentration by the stock dilution factor, α. This starting concentraton, x_0, will be the highest concentration in the series and is plotted as the last dilution on any plots of the raw data. It then is subsequently and serially diluted by this dilution factor eight times to result in nine concentrations. Thus the entire series in the usual increasing order for the combination is obtained

$$x_i = x_0 \alpha^{i-9}, \quad i = 1, \ldots 9 \tag{2.4}$$

For the cell line UACC812 (id 10722) and the drug GSK_Tykerb (id 33), the drug group id determines which concentration series was used in each of their replicates.

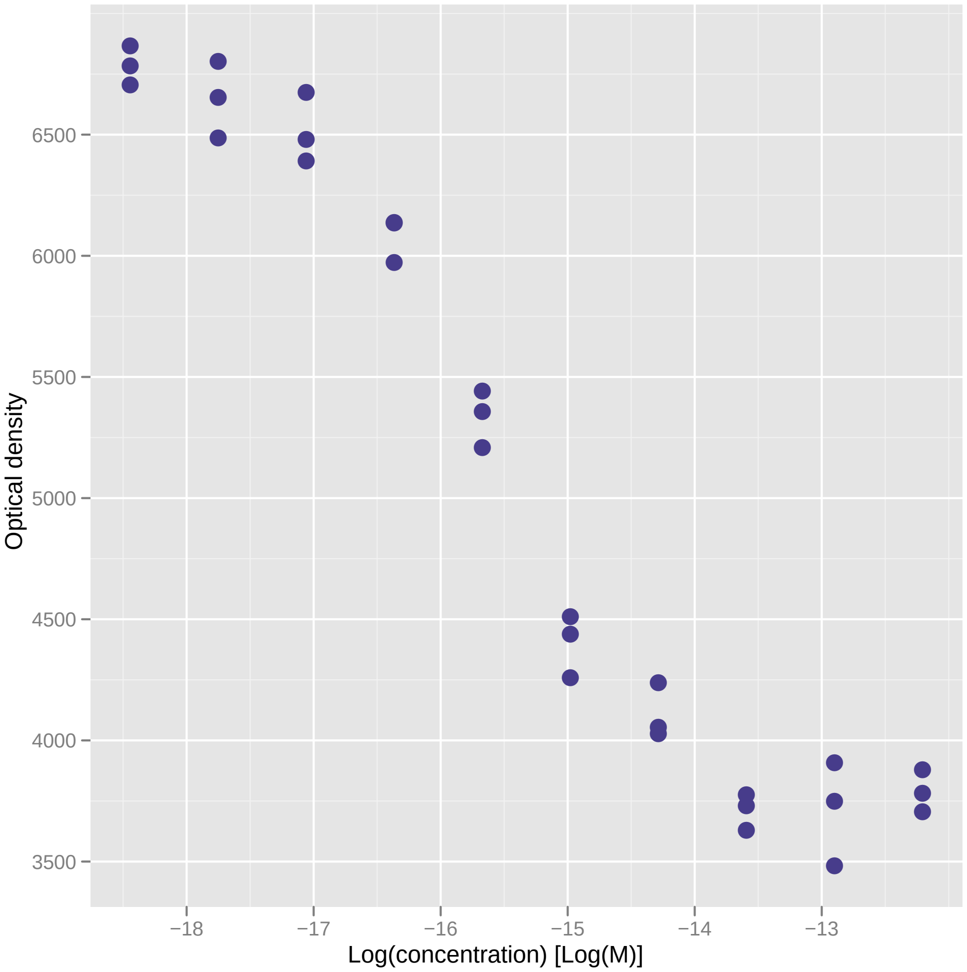


Figure 2.5: Example of a unit of data. As the concentrations of the drug increase (to the right), the more the growth is inhibited in the cell population, as measured by the optical density from the scanner. The values for the vehicle control are included at the zero dilution.

In the first replicate listed in the table, the stock concentration of 5 mM was diluted 1/300th to 1.67 M. All nine dilutions result in the following concentrations:

1	2	3	4	5
4.3×10^{-11} M	2.1×10^{-10} M	1.1×10^{-9} M	5.3×10^{-9} M	2.7×10^{-8} M
6	7	8	9	
1.3×10^{-7} M	6.7×10^{-7} M	3.3×10^{-6} M	1.7×10^{-5} M	

Plate replicate	cell_line_id	drug_id	druggroup_id	dilution_factor	final_dilution	stock_conc	stock_conc_units
1	10722	33	98	5	1:300	5.00	mM
2	10722	33	110	2	1:300	1.50	mM
3	10722	33	111	2	1:300	1.50	mM
4	10722	33	114	2	1:300	0.30	mM
5	10722	33	115	2	1:300	0.30	mM
6	10722	33	116	2	1:300	0.30	mM
7	10722	33	162	2	1:300	1.50	mM
8	10722	33	116	2	1:300	0.30	mM
9	10722	33	180	2	1:300	1.50	mM
10	10722	33	111	2	1:300	1.50	mM

Table 2.2: Example information table on stock concentrations, starting concentrations, and dilution factors for the cell line UACC812 and the drug GSK_Tykerb.

2.2.2 ATP curves

As described previously, each type of plate includes a set of 6 measures for the ATP standard curve. These measures can be utilized for checking whether a plate is functioning properly and compared with each other to identify any noticeable plate effects that may bias the data. Although the dilution series data is randomized on each 96-well plate, the ATP standards are assigned to consecutive wells in the first column of each plate. In general, in our data, plots of the standard curves show close matches between the data and T0 plate pairs (Figure 2.6), indicating little evidence for plate effects. In cases where the curves may potentially differ significantly, it may be possible for later techniques to use the differences between the curves to adjust all the observations appropriately.

2.2.3 NCI data

The National Cancer Institute (NCI) has a long running program called DTP (Developmental Therapeutics Program), involving a cancer screening project that tests thousands of compounds on cancer cell lines [1]. Their experiments typically involved 5 10-fold serial dilutions with samples incubated 48 hours. They compute GI50, TGI, and LC50 from all these experiments and upload them online to be available for download. Biologists frequently use their values as is in their own research papers [3].

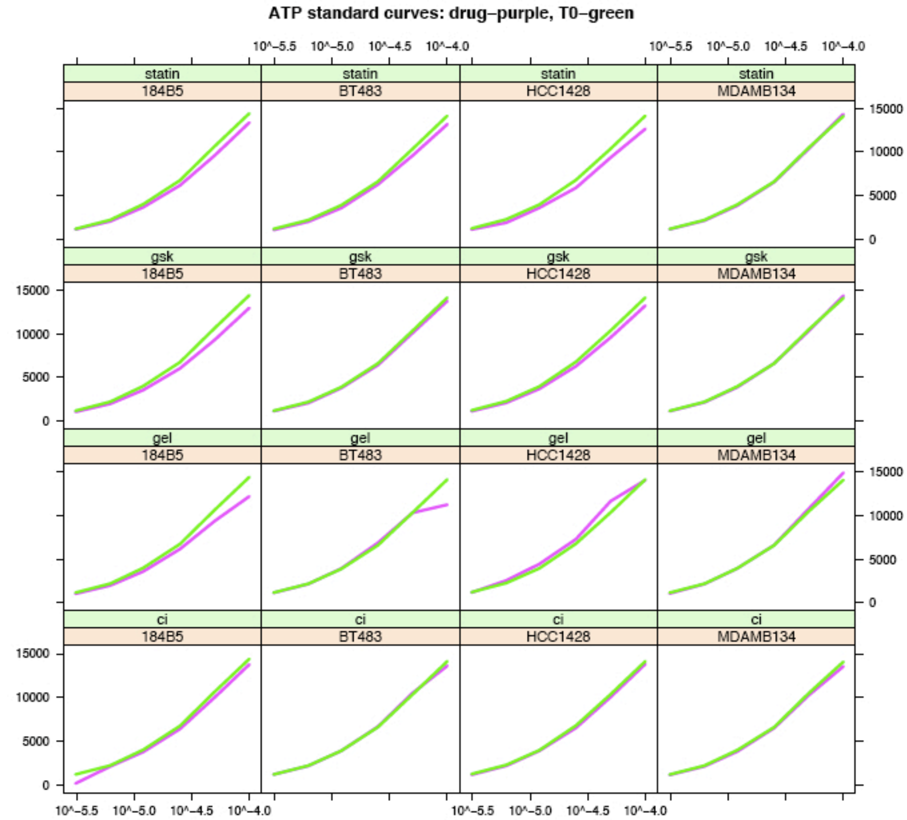
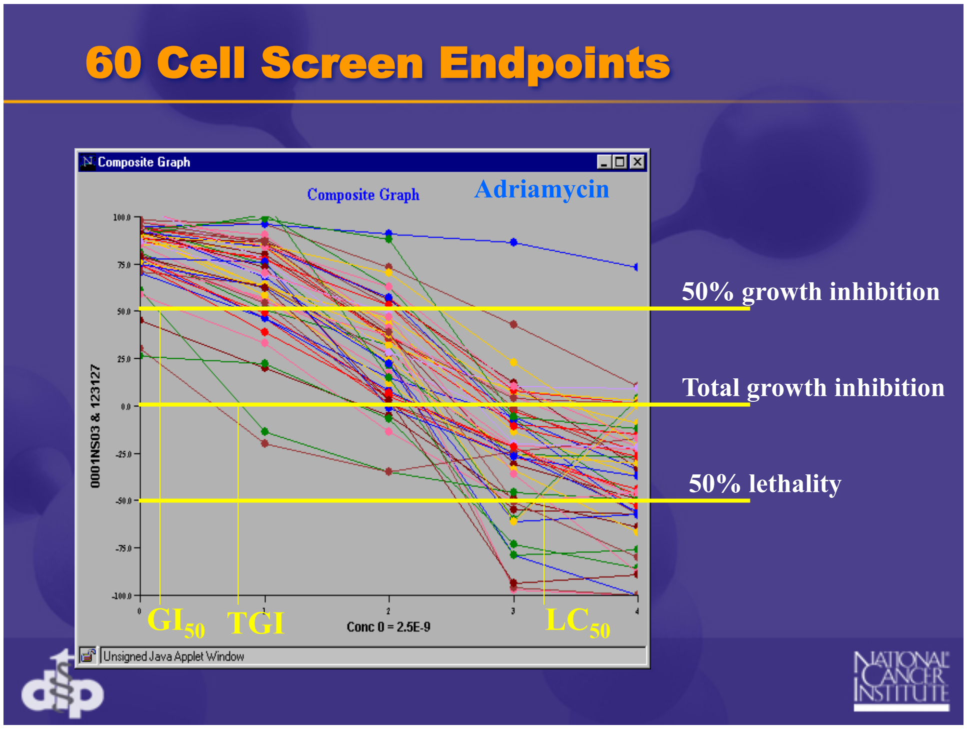


Figure 2.6: Comparison of ATP standard curves on drug plate (purple) and T0 plate (green).

The data includes many replicates of different cell lines and compound combinations; however they analyze each replicate by itself. Their method of computing the special concentration parameters is to first convert their data into a GI% scale according to the formula (Table 2.1), essentially subtracting the observations by initial cell count at time zero and then scaling by the difference in the cell counts from time 0 to time t. They then linearly interpolate between the two data points closest to the 50% horizontal line; Figures 2.7 and 2.8 illustrate this. This means that they are not utilizing a large majority of their data when they only need two data points to calculate a GI50. In addition, with this method, they are unable to calculate any errors associated with their estimate.

Unfortunately, the raw data is unavailable for public use; thus we cannot attempt any analysis that requires the untransformed data.



60 Cell Screen Endpoints

50% growth inhibition

Total growth inhibition

50% lethality

Figure 2.7: National Cancer Institute's method of estimating GI50, TGI, and LC50: piecewise linear interpolation.

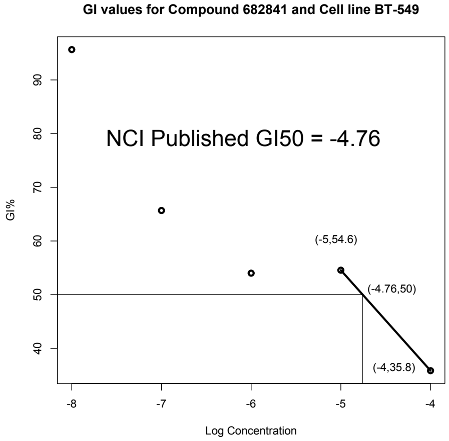


Figure 2.8: Example of how the National Cancer Institute estimates GI50 values.

2.3 Methods

The goal is to summarize the observations from each unit into three important measures: GI50, TGI, and LC50.

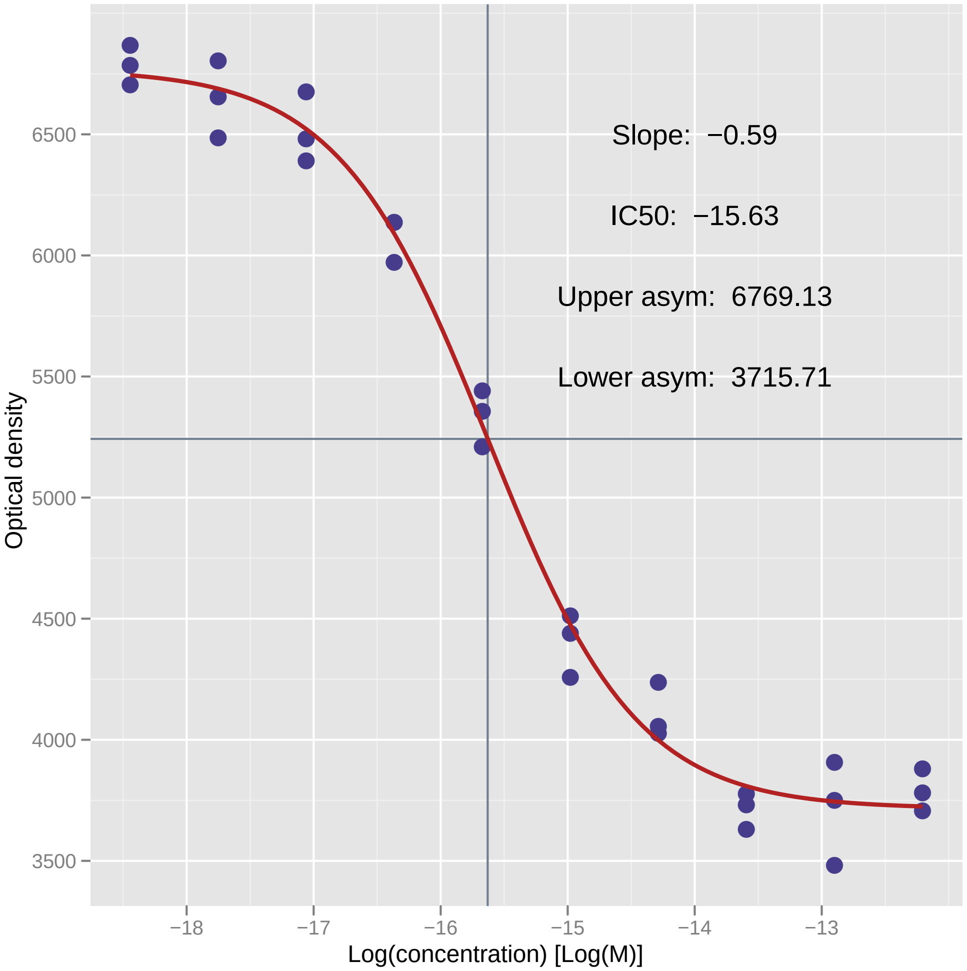


Figure 2.9: Plot of fitted logistic curve on an example unit. The parameter estimates are used later to obtain the GI50, TGI, and LC50 values.

These measures are estimated from a logistic curve that is fitted on these observations. For our purposes, a general formula for a four-parameter logistic curve that is decreasing is

$$f(x) = \phi_0 + \frac{\phi_1 - \phi_0}{1 + \exp\left(-\frac{x-\phi_2}{\phi_3}\right)} \tag{2.5}$$

where ϕ_0 is the lower asymptote, ϕ_1 is the upper asymptote, ϕ_2 is the inflection point, and ϕ_3 is the slope or growth rate. For the independent fitting of each dilution series unit, the following model can be used:

$$y_{ijk}(r,s) = \phi_{0,ijk} + \frac{\phi_{1,ijk} - \phi_{0,ijk}}{1 + \exp\left(\frac{x_{ijk}(r) - \phi_{2,ijk}}{\phi_{3,ijk}}\right)} + \epsilon_{ijk}(r,s) \tag{2.6}$$

where $y_{ijk}(r,s)$ is the number of living cells at replicate s of dilution r, $x_{ijk}(r)$ (log scale), of kth replicate of the cell line i and drug j. In addition, $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$. We include the three control values into the fitting as data at dilution 0, as they contain no drug and will help anchor the estimate of the upper asymptote $\phi_{1,ijk}$. This model can be easily fitted with nls() in R or other common nonlinear estimation programs. See Figure 2.9 for the fitted curve from the example.

Once the curve parameters have been estimated, we can then transform each fitted logistic curve into a GI (growth inhibition) curve by subtracting an estimate of α_{ijk} (the number of cells at time 0 hours) and scaling by the difference between $\phi_{1,ijk}$ (the number of cells at time 72 hours) and α_{ijk}. The estimate of α_{ijk} here is the median of its 88 observations. Similarly, an LC (lethal concentration) curve is produced by subtracting and then scaling by $\widehat{\alpha_{ijk}}$. Consequently, the three special concentration parameters, GI50, TGI, and LC50, can be estimated by solving for the dilution x such that their three formulas hold (Table 2.1).

$$\widehat{x}_{\text{GI50}} = \widehat{\phi}_2 - \widehat{\phi}_3 \log\left(\frac{\widehat{\phi}_1 - \widehat{\alpha}}{\widehat{\phi}_1 + \widehat{\alpha} - 2\widehat{\phi}_0}\right) \tag{2.7}$$

$$\widehat{x}_{\text{TGI}} = \widehat{\phi}_2 - \widehat{\phi}_3 \log\left(\frac{\widehat{\phi}_1 - \widehat{\alpha}}{\widehat{\alpha} - \widehat{\phi}_0}\right) \tag{2.8}$$

$$\widehat{x}_{\text{LC50}} = \widehat{\phi}_2 - \widehat{\phi}_3 \log\left(\frac{2\widehat{\phi}_1 - \widehat{\alpha}}{\widehat{\alpha} - 2\widehat{\phi}_0}\right) \tag{2.9}$$

This method is simple but is only appropriate for the cases where there is only one set of dilution series data, and thus only one fitted curve, for a given cell line / drug combination. For cases with replicate data, fitting them each independently will generate different GI50s, leading to the problem of how to summarize them into one representative GI50. Figure 2.10 shows the resulting GI curves from the individual curves in the original example. As we can see from the plot, the GI curves of the same cell line and drug vary widely as the concentration of the drug increases. Due to the nature of the transformation, any differences are then magnified; the spread in the TGI estimates is much wider than the spread in the GI50 estimates (where the curves intersect the gray horizontal lines at GI = 0 and 0.5, respectively). Biologically, assuming the cells grow in satisfactory conditions and are not allowed to reach confluence, any replicate of a particular combination of cell line and drug should ideally lead to exactly the same GI curve, regardless of the size of the initial cell population.

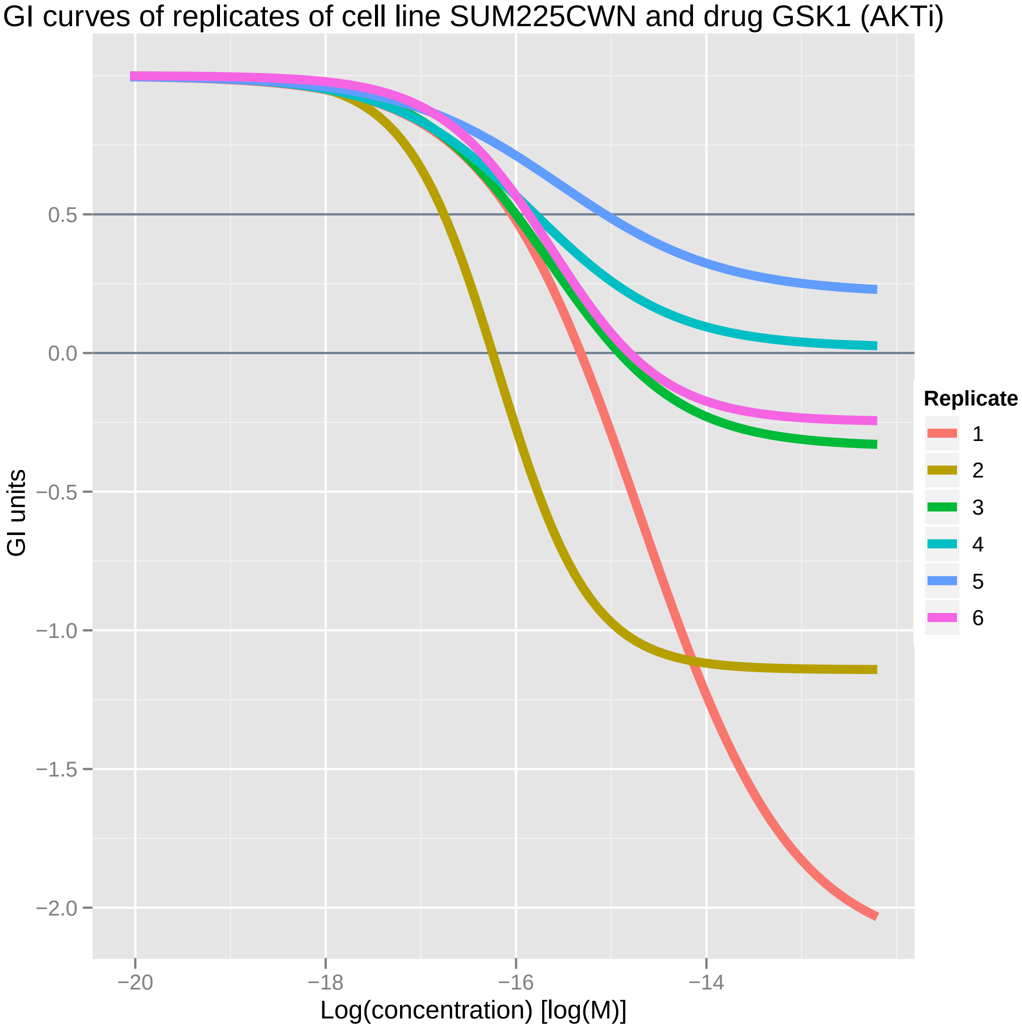


Figure 2.10: Individual GI curves of the replicates of the cell line SUM225CWN and drug GSK1 (AKTi). The separate GI50 and TGI estimates are found at the intersections of the curves at the horizontal lines at the y-intercepts of 0.5 and 0, respectively. Notice that two of the TGI estimates will end up being nonexistent.

2.3.1 Joint estimation across replicates

We now discuss the steps involved to produce a model that will be able to fit all replicate data together while enforcing a common GI curve. At present for a replicate of a given cell line / drug combination, we have a combined set of 27 observations for 9 drug dilutions and 3 observations for the number of cells at time 72 hours. Corresponding to these 30 observations from the drug plate, there are 88 observations of the number of cells at time zero. To try to use all the data in one model, we first need to establish a relationship between the number of cells at time 72 hours given no drug treatment (ϕ_1) and the number of cells at time 0 hour (α). We can assume the relationship is a basic exponential growth equation where at time t, the population is an exponential function of the population at time 0 with a parameter δ controlling the growth rate.

$$N(t) = N(0)\exp(\delta t) \tag{2.10}$$

Since the time t is constant at 72 hours, ϕ_1 is a product of α and a scale factor, β_1, that is common across replicates of a given cell line, i.e.

$$\phi_1 = \beta_1 \alpha \tag{2.11}$$

Two additional assumptions are made: that the parameters ϕ_2 and ϕ_3 are equal across the replicates of a particular combination. Without loss of generality, we shall show results for the case of two replicates, and so, $\phi_2 = \phi_2'$ and $\phi_3 = \phi_3'$. Relationships between the parameters ϕ_0 and ϕ_0', and ϕ_1 and ϕ_1' can be revealed if we set the GI curves of the two replicates equal to each other. If we let $z = 1/(1+\exp(-(x-\phi_2)/\phi_3))$, then

$$\begin{aligned}
\text{GI: } \frac{\phi_0 + (\phi_1 - \phi_0)z - \alpha}{\phi_1 - \alpha} &= \frac{\phi_0' + (\phi_1' - \phi_0')z - \alpha'}{\phi_1' - \alpha} && (2.12)\\
&= \frac{\frac{1}{\alpha'}\left\{\phi_0' + (\phi_1' - \phi_0')z - \alpha'\right\}}{\frac{\phi_1'}{\alpha'} - 1} && (2.13)\\
&= \frac{\frac{1}{\alpha'}\left\{\phi_0' + (\phi_1' - \phi_0')z - \alpha'\right\}}{\frac{\phi_1}{\alpha} - 1} && (2.14)\\
&= \frac{\frac{\alpha}{\alpha'}\left\{\phi_0' + (\phi_1' - \phi_0')z - \alpha'\right\}}{\phi_1' - \alpha} && (2.15)\\
&= \frac{\frac{\alpha}{\alpha'}\phi_0' + \frac{\alpha}{\alpha'}(\phi_1' - \phi_0')z - \alpha}{\phi_1 - \alpha} && (2.16)
\end{aligned}$$

Thus, $\phi_0 = \dfrac{\alpha}{\alpha'}\phi_0'$. Since this holds for any arbitrary pair of replicates, we can extend this to a variable number of replicates, n, to say that the ratio $\dfrac{\phi_{0,k}}{\alpha_k}$ for every replicate k is equal. Substitution of $\phi_{0,k} = \alpha_k \beta_0$ and $\phi_{1,k} = \alpha_k \beta_1$ into the basic logistic model leads to a model that involves all the data for a given cell line / drug combination.

$$y_k = \alpha_k \left(\beta_0 + \frac{\beta_1 - \beta_0}{1 + \exp\left(-\frac{x-\beta_2}{\beta_3}\right)} \right) + \epsilon_k \tag{2.17}$$

The resulting model is also a logistic curve; each fitted replicate curve is composed of a base curve, common to all replicates, that is scaled by its own value of the initial cell population, α_k. That each fitted curve will produce the same GI curve is proved easily, as the replicate-specific α_k parameter will be canceled out during the transformation to the GI curve.

$$\mathrm{GI}(x) = \frac{1}{\beta_1 - 1}\left(\beta_0 + \frac{\beta_1 - \beta_0}{1 + \exp\left(\frac{x - \beta_2}{\beta_3}\right)} - 1\right) \tag{2.18}$$

Similarly, the LC (lethal concentration) curve is the same for all replicates.

$$\mathrm{LC}(x) = (\beta_1 - 1)\mathrm{GI}(x) = \left(\beta_0 + \frac{\beta_1 - \beta_0}{1 + \exp\left(\frac{x - \beta_2}{\beta_3}\right)} - 1\right) \tag{2.19}$$

The three special concentration parameters can be estimated from this model as

$$\widehat{x}_{\mathrm{GI50}} = g_1(\widehat{\boldsymbol{\beta}}) = \widehat{\beta}_2 - \widehat{\beta}_3 \log\left(\frac{\widehat{\beta}_1 - 1}{\widehat{\beta}_1 + 1 - 2\widehat{\beta}_0}\right) \tag{2.20}$$

$$\widehat{x}_{\mathrm{TGI}} = g_2(\widehat{\boldsymbol{\beta}}) = \widehat{\beta}_2 - \widehat{\beta}_3 \log\left(\frac{\widehat{\beta}_1 - 1}{1 - \widehat{\beta}_0}\right) \tag{2.21}$$

$$\widehat{x}_{\mathrm{LC50}} = g_3(\widehat{\boldsymbol{\beta}}) = \widehat{\beta}_2 - \widehat{\beta}_3 \log\left(\frac{2\widehat{\beta}_1 - 1}{1 - 2\widehat{\beta}_0}\right) \tag{2.22}$$

Thus, to estimate the special concentration parameters, we need to estimate $n+4$ parameters such that the constrained fitting of the n sets of replicate data will result in the same growth inhibition and lethal concentration curves.

2.3.2 Likelihood

To initiate the discussion of how the parameters are estimated, we present the likelihood of the data as follows. All the data involved come from three sources across two different plates: the measurements for the dilution series, for the vehicle control and for the T0 values. The first set provides the information for how the different levels of drug are affecting the cells, the second set on the population in the absence of the drug, and the third on the population at time zero. We assume the following model for our measurements:

$$y_{ijk}(r, s) \sim \mathcal{N}(f(x_{ijk}(r), \boldsymbol{\beta}_{ij}, \alpha_{ijk}), \sigma_{ij}^2) \tag{2.23}$$

$$y_{ijk}(0, s) \sim \mathcal{N}(\alpha_{ijk}\beta_{1,ij}, \gamma_{ij}^2) \tag{2.24}$$

$$a_{ijk}(s) \sim \mathcal{N}(\alpha_{ijk}, \tau^2) \tag{2.25}$$

where $\boldsymbol{\beta}_{ij} = \begin{bmatrix} \beta_{0,ij} \\ \beta_{1,ij} \\ \beta_{2,ij} \\ \beta_{3,ij} \end{bmatrix}$, a_{ijk} are the replicates of the corresponding T0 values for the kth replicate of the cell line i and drug j, f is the four-parameter logistic curve function, σ_{ij} is the standard deviation of the errors $\epsilon_{ij}(r, s)$, γ_{ij} is the standard deviation of the vehicle controls $y_{ijk}(0, s)$, and τ is the standard deviation of the a_{ijk} values. Recall that r and s specify the sth replicate of the rth dilution for each unit.

Thus, the likelihood as a function of the parameters is

$$\mathcal{L}(\boldsymbol{\beta}_{ij}, \alpha_{ijk}, \sigma^2_{ij}, \gamma^2_{ij}, \tau^2_{ij} | y_{ijk}(r, s), a_{ijk}(s)) = \tag{2.26}$$
$$\prod_{k=1}^{n_{ij}} \prod_{s=1}^{3} P(y_{ijk}(0, s)|\beta_{1,ij}, \alpha_{ijk}, \gamma_{ij}) P(a_{ijk}(s)|\alpha_{ijk}, \tau_{ij}) \prod_{r=1}^{9} P(y_{ijk}(r, s)|\boldsymbol{\beta}_{ij}, \alpha_{ijk}, \sigma_{ij})$$

Maximizing this likelihood is equivalent to minimizing the following penalized sum of squares.

$$h(\boldsymbol{\beta}_{ij}, \alpha_{ijk}, \sigma_{ij}, \gamma_{ij}, \tau_{ij}) = \sum_{k=1}^{n_{ij}} \sum_{r=1}^{9} \sum_{s=1}^{3} A_{ijkrs} + \sum_{k=1}^{n_{ij}} \sum_{s=1}^{3} B_{ijks} + \sum_{k=1}^{n_{ij}} \sum_{t=1}^{88} C_{ijkt} \tag{2.27}$$

where

$$A_{ijkrs} = \frac{(y_{ijk}(r, s) - f(x_{ijk}(r), \boldsymbol{\beta}_{ij}, \alpha_{ijk}))^2}{\sigma^2_{ij}} \tag{2.28}$$
$$B_{ijks} = \frac{(y_{ijk}(0, s) - \alpha_{ijk}\beta_{1,ij})^2}{\gamma^2_{ij}} \tag{2.29}$$
$$C_{ijkt} = \frac{(a_{ijk}(t) - \alpha_{ijk})^2}{\tau^2_{ij}} \tag{2.30}$$

These sums of squares can again be minimized with the function optim() in R or other common optimization tools. The four logistic curve parameters, $\beta_{0,ij}$, $\beta_{1,ij}$, $\beta_{2,ij}$, $\beta_{3,ij}$, and the n_{ij} scale factors, $\alpha_{ijk}, k = 1, \ldots, n$ will be estimated at alternate steps from the variance parameters, σ^2_{ij}, γ^2_{ij}, τ^2_{ij}, until convergence.

Initial values and iterative estimation

Reasonable starting values can be easily found by first setting α^0_{ijk} to be the medians of the 88 observations for each replicate k. Then $\beta^0_{0,ij}$, $\beta^0_{1,ij}$, $\beta^0_{2,ij}$, $\beta^0_{3,ij}$ are estimated from the ordinary nonlinear least squares fitting of the logistic curve to one of the k replicate data using $\widehat{\alpha}^0_{ijk}$. Once optim() provides the next set of these estimates, the variance parameters can be then be updated.

2.3.3 Heteroscedasticity and robustness

The variance of observations from dose response assays commonly increases as a function of the observations. To ensure the equal contribution of points, except for outliers, to the fitted curve, we include weights in the sums of squares that will be updated at each iteration of the method. This gives

$$h_w(\boldsymbol{\beta}_{ij}, \alpha_{ijk}, \sigma, \gamma, \tau) = \sum_{k=1}^{n_{ij}} \sum_{r=1}^{9} \sum_{s=1}^{3} w_{1,ijk}(r,s) A_{ijkrs} + \sum_{k=1}^{n_{ij}} \sum_{s=1}^{3} w_{2,ijk}(s) B_{ijks} + \sum_{k=1}^{n_{ij}} \sum_{s=1}^{3} w_{3,ijk}(t) C_{ijkt} \tag{2.31}$$

where

$$w_{1,ijk}(r,s) = \frac{b_{ijk}(r,s)}{\bar{y}_{ijk}(r)^2}, r \neq 1 \tag{2.32}$$

$$w_{2,ijk}(s) = \frac{b_{ijk}(0,s)}{\bar{y}_{ijk}(0)^2} \tag{2.33}$$

$$w_{3,ijk}(s) = \frac{c_{ijk}(t)}{\bar{a}_{ijk}(t)^2} \tag{2.34}$$

The Tukey biweights, $b_{ijk}(r,s)$ and $c_{ijk}(t)$, in each iteration are calculated from the residuals of the prior iteration's fitted values.

2.3.4 Standard errors

Standard errors of the GI50 estimates are calculated with the delta method. The results from using the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method in optim() include an approximation to the Hessian, the matrix of second derivatives of the objective function $h(\boldsymbol{\beta}_{ij})$ (Eqn 2.27) at the minimum. Since the Fisher information matrix is the negative Hessian and the inverse of the covariance matrix, the covariance matrix of $\boldsymbol{\beta}_{ij}$, $\Sigma_{\boldsymbol{\beta}_{ij}}$, is thus found as the inverse of the negative Hessian.

$$\Sigma_{\boldsymbol{\beta}_{ij}} = \left[- \left(\frac{\partial^2 h_w}{\partial \boldsymbol{\beta}_{ij}^2} \right)^T \left(\frac{\partial^2 h_w}{\partial \boldsymbol{\beta}_{ij}^2} \right) \right]^{-1} \tag{2.35}$$

Then the standard errors of $\widehat{x}_{\mathrm{GI50}}$, $\widehat{x}_{\mathrm{TGI}}$, and $\widehat{x}_{\mathrm{LC50}}$ are the square roots of

$$\left(\left. \frac{\partial g_l}{\partial \boldsymbol{\beta}_{ij}} \right|_{\widehat{\boldsymbol{\beta}}_{ij}} \right)^T \widehat{\Sigma}_{\boldsymbol{\beta}_{ij}} \left(\left. \frac{\partial g_l}{\partial \boldsymbol{\beta}_{ij}} \right|_{\widehat{\boldsymbol{\beta}}_{ij}} \right), \qquad l = 1, 2, 3 \tag{2.36}$$

2.4 Results

2.4.1 Influence of T0 values

Attempts to fit the model on various sets of replicate data resulted in fitted curves whose lower and upper asymptotes do not follow the observations of the dilution series closely (Figure 2.12). This is due to the inclusion of the T0 values. We can see that a relationship between the two variables is clear but not strong enough that the use of a direct relationship between $\phi_{1,ijk}$ and α_{ijk} (Eqn 2.11) results in smaller discrepancies in the fitted curves from the observations. The fact that there are 88 a_{ijk} values means considerably more weight is given to the estimates of α_{ijk} being the robust mean of these 88 values than from the observations in the dilution series near where the upper asymptotes, $\phi_{1,ijk}$ are estimated. As a result, both sets of lower and upper asymptotes estimates are constant multiples of the mean of the a_{ijjk} values.

How the dilution series data are adjusted by the T0 values is seen in Figures 2.11 and 2.12. The first plot shows the medians of the T0 values for each replicate plotted against the values of the upper asymptotes as fitted on the dilution series observations only. The distances and locations of these points relative to the best fit line corresponds to the distances and locations of the fitted curves to their respective set of observations. For instance, the replicates whose points appear above the best fit line have fitted curves that underestimate the observations and similarly, replicates whose points are below the best fit line have fitted curves that overestimate the observations.

The fitting of the dilution series data without the corresponding T0 values results in curves that fit the observations as well as individual fitting can (Figure 2.15). Although we are unable to judge whether the original fitted curves are in any way better than these, both sets of estimates from constrained fitting with and without the T0 values end up with IC50 and slope estimates that are quite similar, -15.72 vs. -15.63 and -0.72 vs. -0.66, in addition to hitting quite close to the weighted means of the individual estimates, -15.62 and -0.61 (Tables 2.3 and 2.4). This is because it is in effect averaging the ratios of $\phi_{1,ijk}$ and α_{ijk}; thus, the estimates of the common parameters, $\boldsymbol{\beta}_{ij}$ will be still be quite close.

More complex methods, such as mixed effects models, may be better at addressing the balance between observations from the two different plates. Forcing the direct relationship between them may not be as appropriate as modeling them more flexibly. Techniques from later chapters can be adapted here for future analysis. Because of this, we continue to focus on fitting the dilution series data separately from the T0 data.

The LBL database currently contains data for up to 14 replicates per cell line and drug combination. We illustrate some results of the constrained fitting method on the set of six replicates for the cell line SUM225CWN and the drug GSK1 (AKTi), one of the few sets where every single replicate can be fit independently. Comparisons of the

Replicate	$\widehat{\phi}_0$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$
1	910.80	1710.28	-15.72	-0.72
2	1797.30	3374.93	-15.72	-0.72
3	1843.20	3461.12	-15.72	-0.72
4	768.15	1442.41	-15.72	-0.72
5	2008.80	3772.08	-15.72	-0.72
6	3889.80	7304.18	-15.72	-0.72

Table 2.3: Table of estimates from constrained fitting of cell line SUM225CWN and drug GSK1(AKTi) using T0 values

resulting estimates from both the models show very close values (Figures 2.13, 2.14) , while plots of the fitted curves show that the logistic curve fits the observations quite well (Figures 2.15 and 2.16) and the residual plots do not reveal anything amiss in our assumptions (Figure 2.17). Because of the way the constrained model is constructed, these resulting estimates (Table 2.4) will lead to a single GI curve and thus, one GI50 and TGI value for this set. However the individual fittings will lead to six different GI curves and the issue of how to summarize across these curves in an appropriate way, especially when they look like the curves in Figure 2.10.

	Individual				Constrained			
Replicate	$\widehat{\phi}_0$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	$\widehat{\phi}_0$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$
1	590.69	1206.32	-14.70	-0.80	682.94	1258.20	-15.63	-0.66
2	1457.27	2469.72	-16.18	-0.48	1305.04	2404.30	-15.63	-0.66
3	1704.56	3064.60	-15.65	-0.68	1664.30	3066.18	-15.63	-0.66
4	873.56	1893.68	-15.83	-0.73	960.02	1768.66	-15.63	-0.66
5	2942.99	5549.59	-15.54	-0.84	3002.05	5530.74	-15.63	-0.66
6	3715.71	6769.13	-15.63	-0.59	3688.39	6795.18	-15.63	-0.66
	Weighted mean		-15.62	-0.61				

Table 2.4: Table of individual and constrained (without using T0 values) nonlinear least squares estimates for cell line SUM225CWN and drug GSK1(AKTi)

Since the results of the constrained fitting match well with the individual estimates, one may question why joint estimation is even necessary when we can simply fit each replicate individually and then take the mean. First of all, even if we obtain the same estimates from both methods, the error involved in summarizing the individual cases would be greater than the error of the estimate obtained using all the data at once. More importantly is that for practical reasons, individual fitting of the replicates will not suffice in general. There are too many cases where the data of one single unit is not good enough to be able to illuminate the entire curve. This is true for all the real and large datasets that we have come across. A majority of the time, it is a tricky problem where the biologist does not know exactly in which range of concentrations the main activity will lie, but requires such information in order to run an experiment to figure this out. Thus, valuable spots on the experimental plates will not be used optimally, and there will be insufficient information in some areas.

Figure 2.18 displays an example of a set of data where the individual model is unable to fit 5 out of the 12 replicates. The first two replicates have a much wider range of concentrations used than the others and must have been a guide for later replicates that focus more on the area where the IC50 and slope parameters are fitted. Even so, a few of the newer sets do not have enough information to cover the estimation of either or both of the asymptotes. However, together they have more than enough to cover the important range of the curve where the IC50 and slope lies.

Figure 2.20 shows that the constrained model is able to fit all 12 of the replicates. We can tell from these fitted plots and the residual plots (Figure 2.22) that the majority of the replicates are fit quite reasonably and that replicate 9 is an outlier; we can check from the database whether this is due to experimental conditions such as who performed the work or when.

Both sets of estimates from individual and constrained fitting are included in Table 2.5. The weighted mean of the possible estimates of IC50 from individual fitting is -14.13 while the estimate from the constrained fitting on all the data is -14.31. The slope estimates are close as well: -0.70 vs -0.86. We compare the lower and upper asymptote estimates in Figures 2.23 and 2.24 to see that the constrained fit is not forcing the asymptote values to be too different from the individual fits.

2.4.2 Goodness-of-fit

Although we have seen visually that the constrained fitting of the data works well, we can try to justify this model quantitatively with a goodness-of-fit test. Because it requires fewer parameters, the constrained model is the null one, while the individual model is the alternative. The F test determines whether the larger model provides a significantly better fit than the smaller model.

$$F = \frac{((SS_{null} - SS_{alt})/(df_{null} - df_{alt}))}{SS_{alt}/df_{alt}} \tag{2.37}$$

Does the closer fit of the observations we get from fitting each replicate separately justify the extra parameters? The number of parameters for the constrained model varies additively as the number of replicates varies ($n + 4$), while the number of parameters for the individual model varies multiplicatively ($4n$). We calculate the F statistic [Eqn 2.37] to the example with 6 replicates (180 observations) in this chapter that was able to have all replicates successfully fit independently.

	Individual				Constrained			
Replicate	$\widehat{\phi}_0$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$	$\widehat{\phi}_0$	$\widehat{\phi}_1$	$\widehat{\phi}_2$	$\widehat{\phi}_3$
1	NA	NA	NA	NA	2418.44	6230.58	-14.31	-0.86
2	NA	NA	NA	NA	1958.98	5046.89	-14.31	-0.86
3	1395.21	4831.63	-14.00	-0.90	1852.95	4773.73	-14.31	-0.86
4	2191.20	5448.42	-13.99	-0.59	2203.93	5677.93	-14.31	-0.86
5	2211.77	5301.63	-14.53	-0.79	2042.42	5261.86	-14.31	-0.86
6	NA	NA	NA	NA	2198.20	5663.17	-14.31	-0.86
7	2020.91	5688.89	-14.36	-0.79	2194.96	5654.84	-14.31	-0.86
8	279.70	5172.66	-13.09	-1.90	1886.70	4860.67	-14.31	-0.86
9	NA	NA	NA	NA	1364.13	3514.38	-14.31	-0.86
10	1372.78	4667.71	-14.40	-1.24	1704.62	4391.57	-14.31	-0.86
11	NA	NA	NA	NA	1189.79	3065.24	-14.31	-0.86
12	1945.40	4387.32	-14.35	-0.53	1748.53	4504.70	-14.31	-0.86
	Weighted mean		-14.13	-0.70				

Table 2.5: Table of individual and constrained nonlinear least squares estimates for cell line HCC1954 and drug GSK1 Tykerb.

$$F = \frac{(8558306 - 7782805)/[(180 - 10) - (180 - 24)]}{7782805/(180 - 24)} = 1.11 \tag{2.38}$$

The p-value we obtain from this test statistic and the $F_{14,156}$ distribution is 0.35, too large to consider rejecting the null hypothesis that the constrained model is the correct one. It can fit all the data almost as well as the individual fitting can, while using fewer parameters, enforcing the summarization of all replicate data at once.

Although this shows that the constrained model is the preferred one for this specific dataset, we cannot easily generalize this for all other data without looking at a few more results of this test. However, this requires sets of individual estimates to be complete, something that is less common than we expected.

2.5 Discussion

Because we are working with replicate data where the goal is just one set of special concentration estimates, constrained fitting makes sense and by sharing strength across the units, all the data can be used. The units that fail to be fitted by individual methods are usually not a complete loss. They often contribute much information to the pool and only need to be helped a little by other units in order to be fitted reasonably. This is why we cannot settle for doing the simplest method of fitting each unit independently, somehow summarizing their results, and ignoring the ones that cannot contribute in this way.

Through some key informative assumptions, we are able to formulate a model that can tackle all replicate data as a whole to provide the most intuitive way to ultimately calculate the GI50, TGI, and LC50 values.

More work is needed to incorporate the T0 values in an appropriate way; we suggest the use of mixed effects models, which we introduce in the next chapter, that can model this data more flexibly.

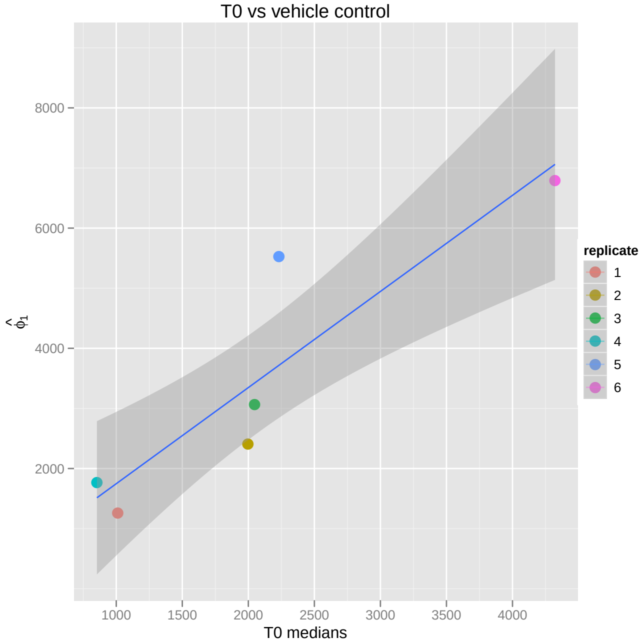


Figure 2.11: Plot of individually estimated upper asymptote $\widehat{\phi}_1$ values against the median of T0 values, including the best fit line for cell line SUM225CWN and drug GSK1 (AKTi).

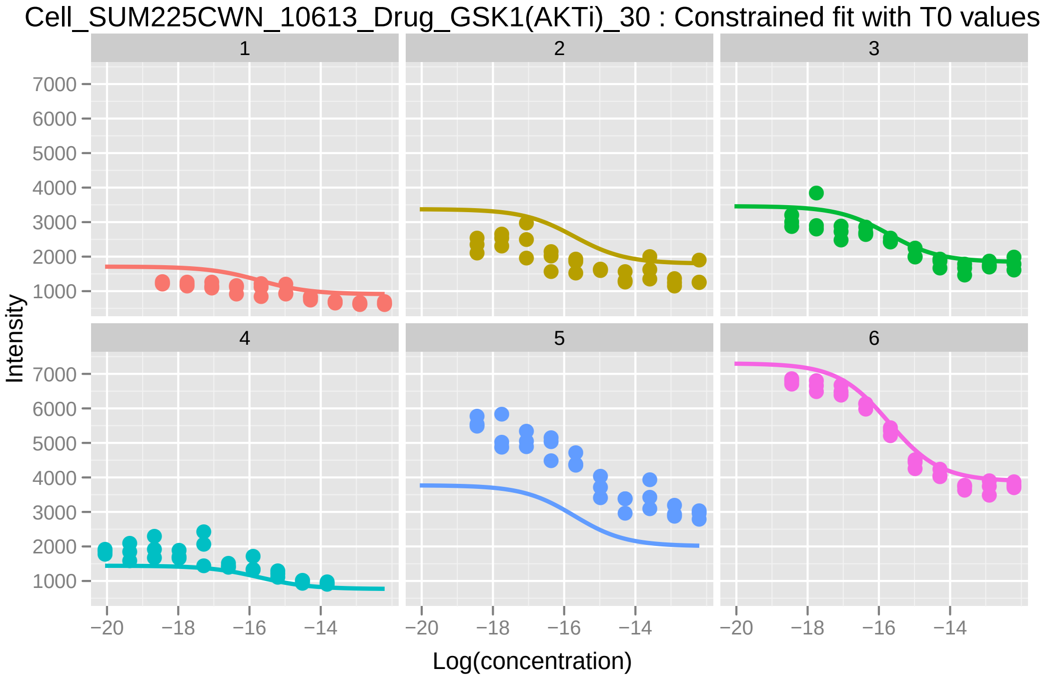


Figure 2.12: Fitted curves from the constrained fitting of cell line SUM225CWN and drug GSK1(AKTi) using T0 values.

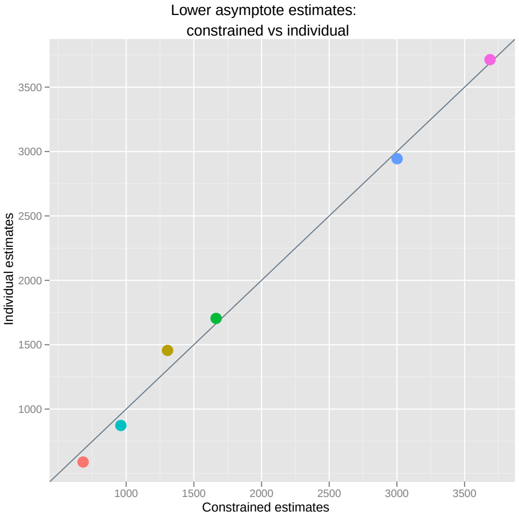


Figure 2.13: Cell line SUM225CWN and drug GSK1 (AKTi) : Constrained vs individual estimates of lower asymptote. A gray line with intercept 0 and slope 1 is included for reference.

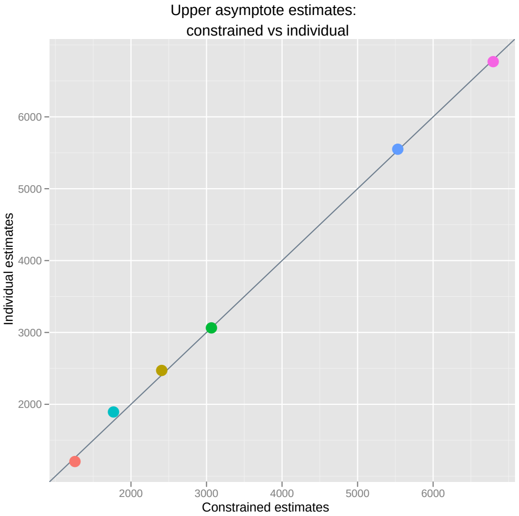


Figure 2.14: Cell line SUM225CWN and drug GSK1 (AKTi) : Constrained vs individual estimates of lower asymptote. A gray line with intercept 0 and slope 1 is included for reference.

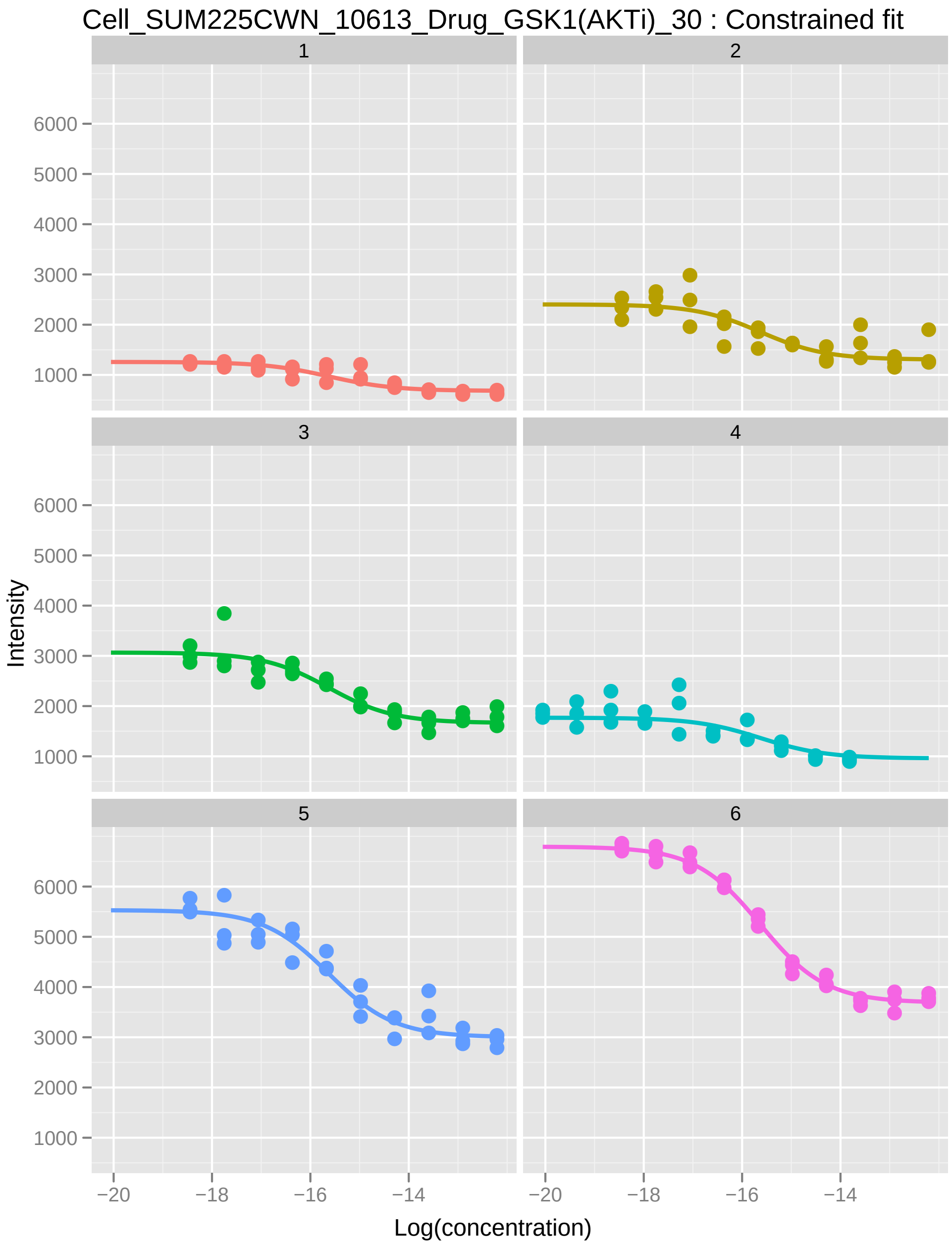


Figure 2.15: Fitted curves from constrained fitting (no T0 values used) for the six replicates of cell line SUM225CWN and drug GSK1(AKTi).

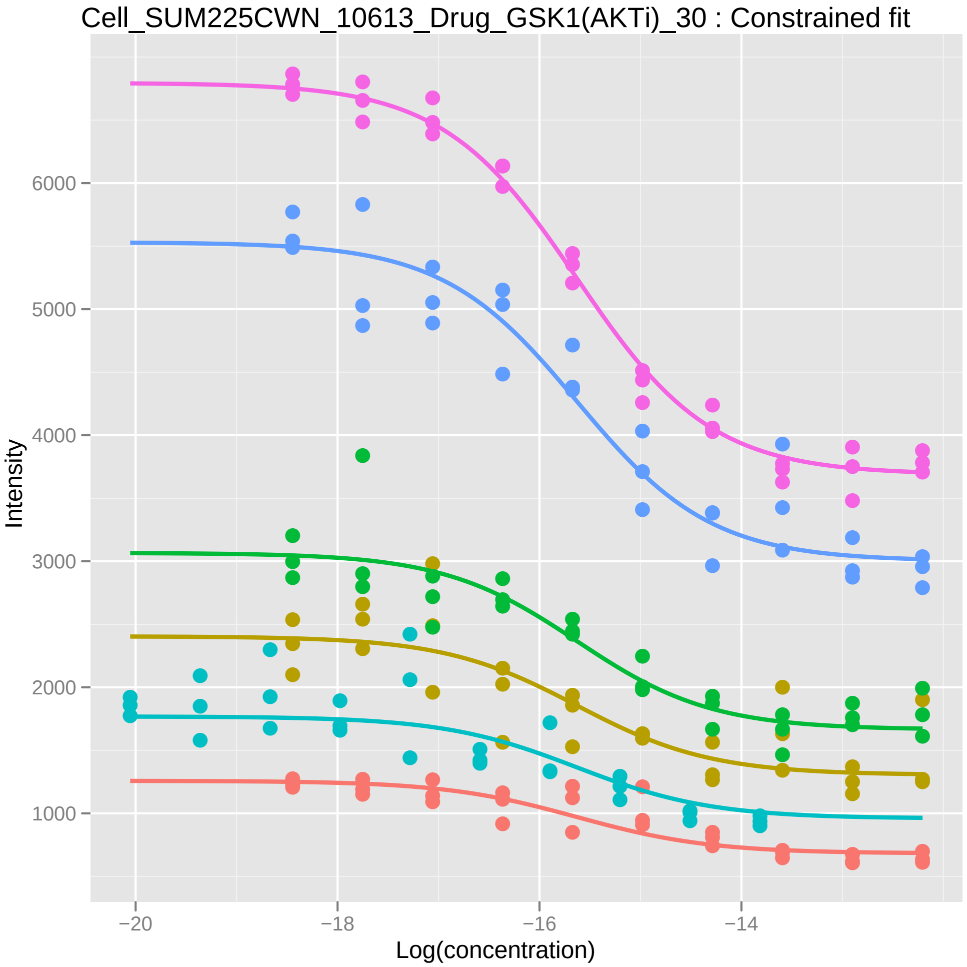


Figure 2.16: Fitted curves from constrained fitting for the six replicates of cell line SUM225CWN and drug GSK1(AKTi).

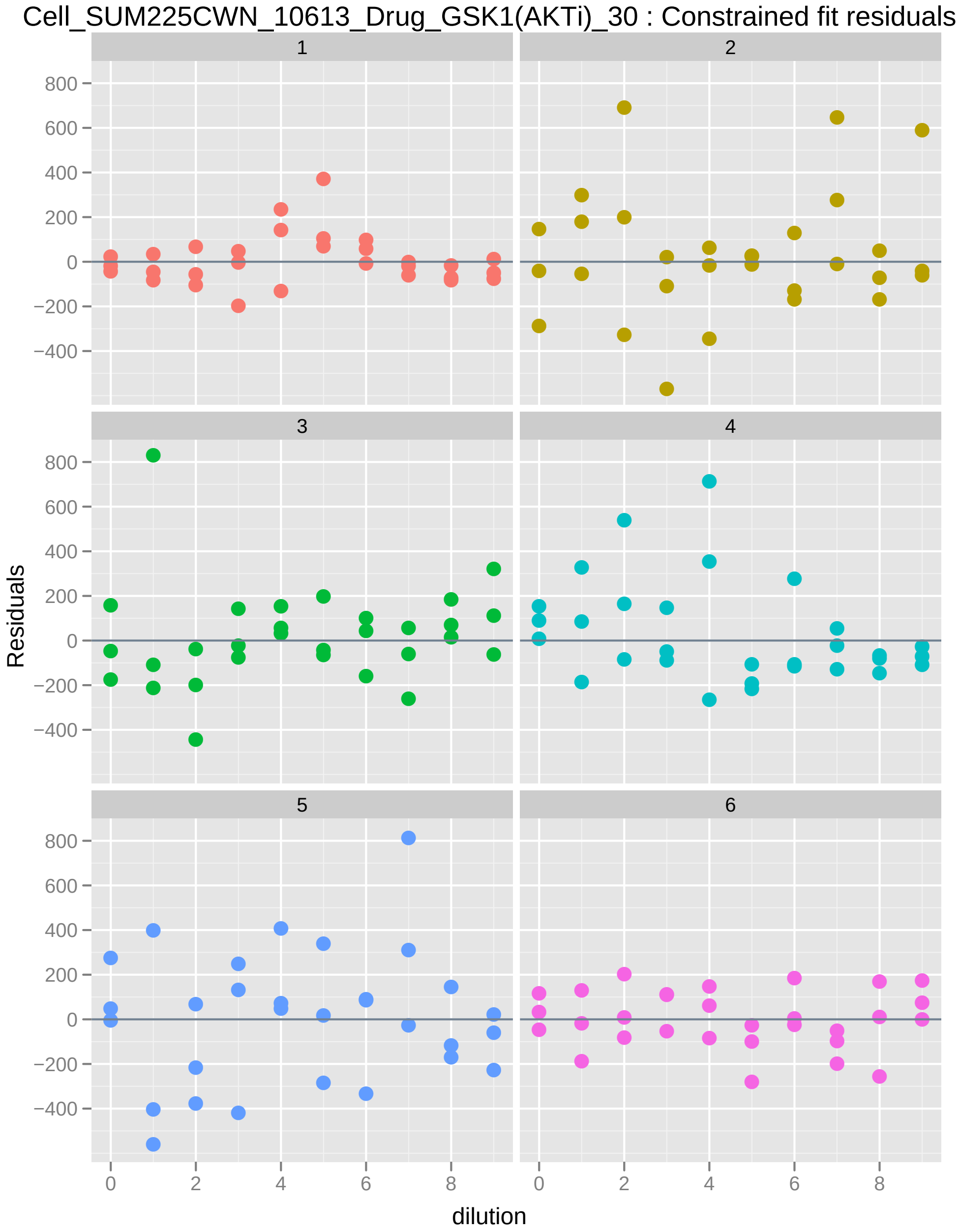


Figure 2.17: Residuals from constrained fitting for the six replicates of cell line SUM225CWN and drug GSK1(AKTi).

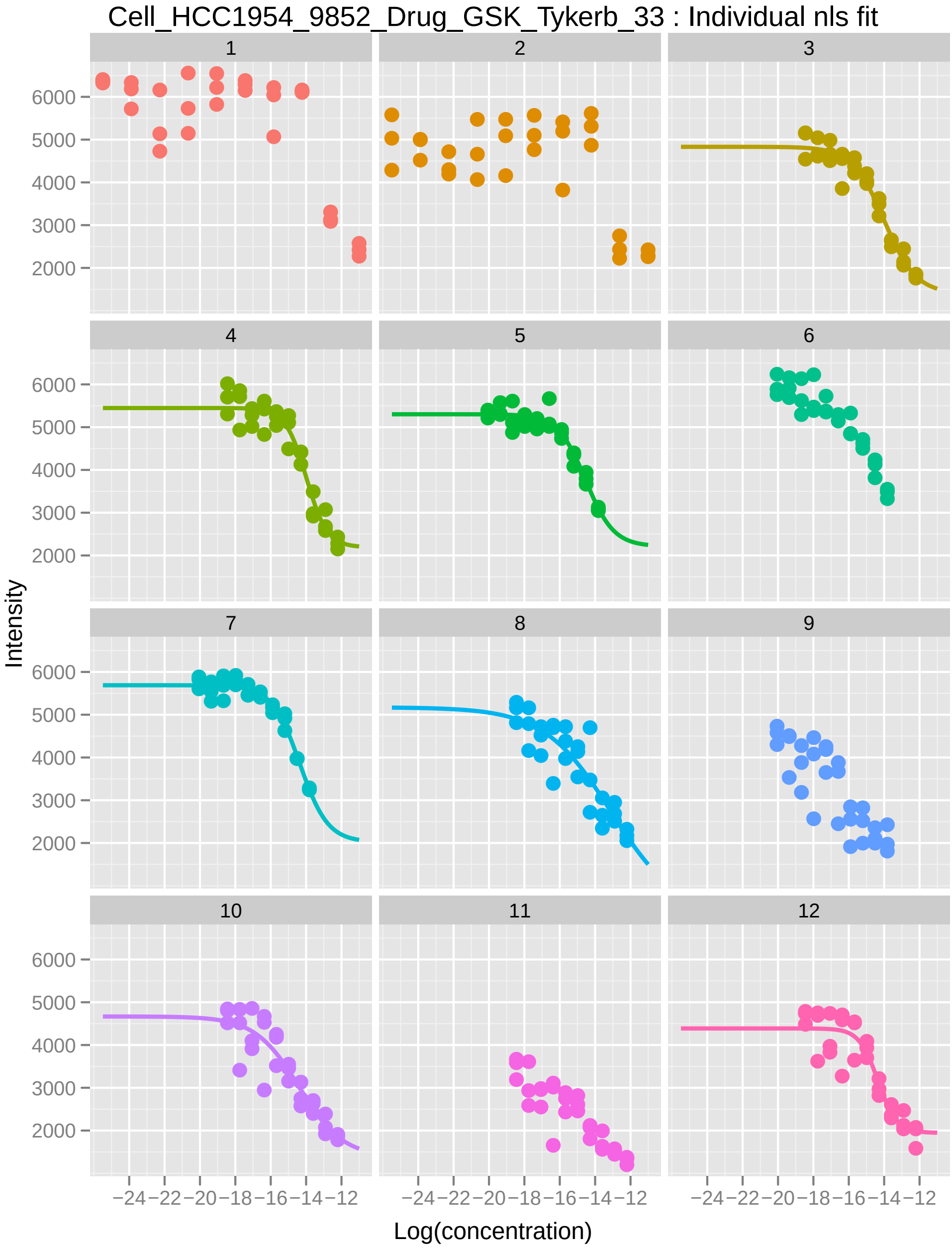


Figure 2.18: Fitted curves from individual fitting for the 12 replicates of cell line HCC1954 and drug GSK Tykerb.

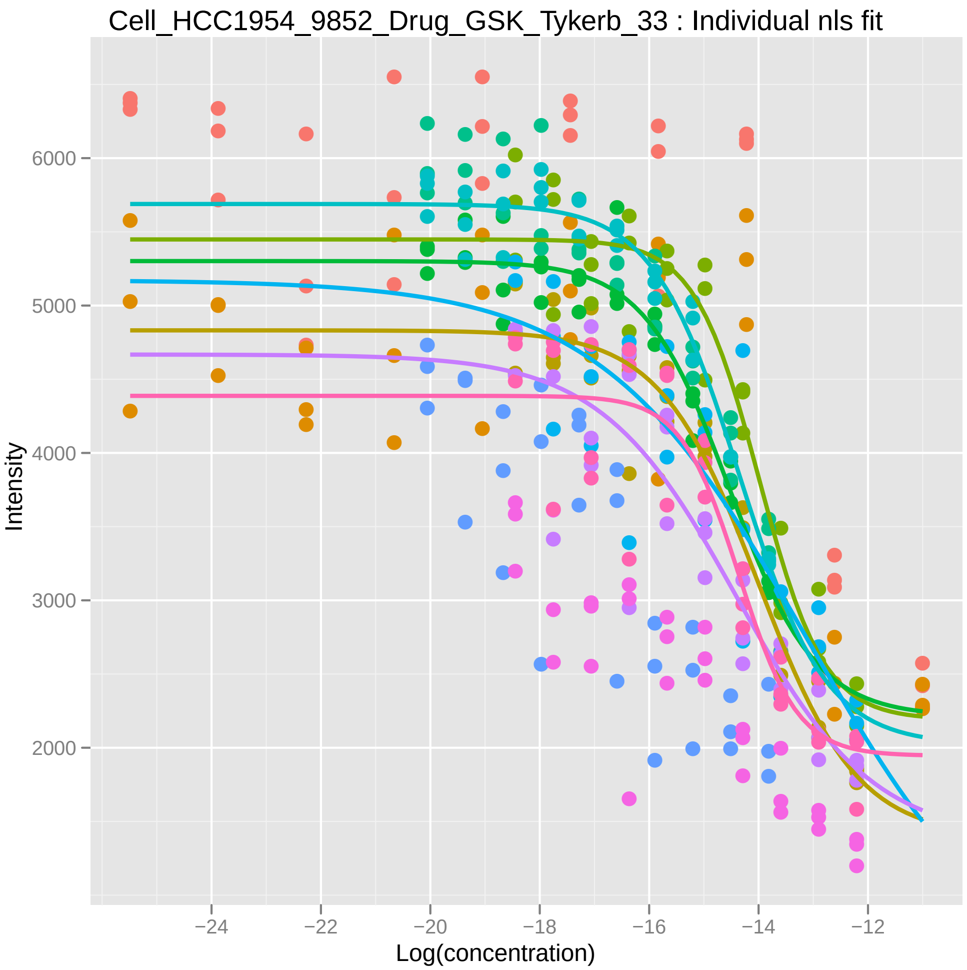


Figure 2.19: Fitted curves from individual fitting for the 12 replicates of cell line HCC1954 and drug GSK Tykerb.

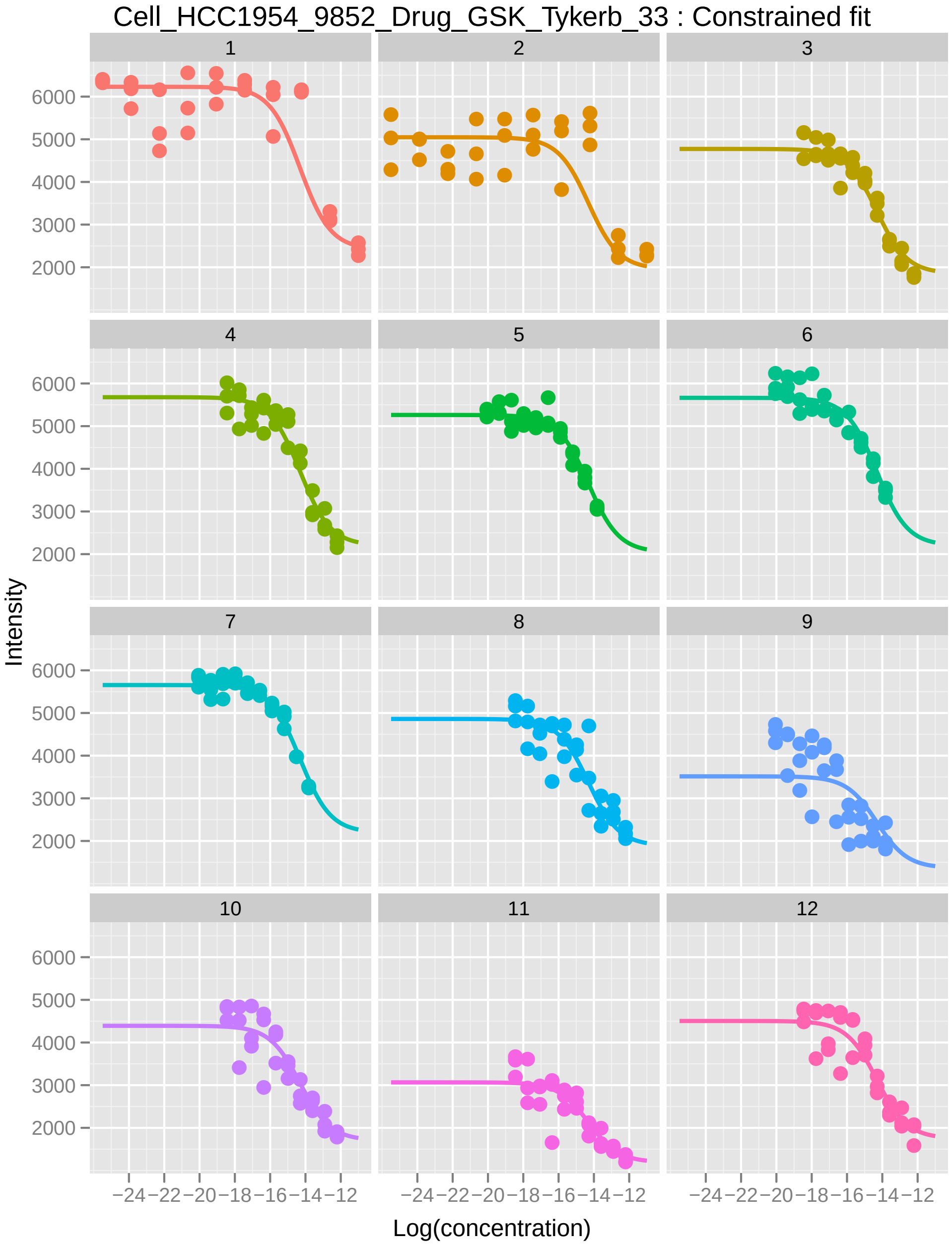


Figure 2.20: Fitted curves from constrained fitting for the 12 replicates of cell line HCC1954 and drug GSK Tykerb.

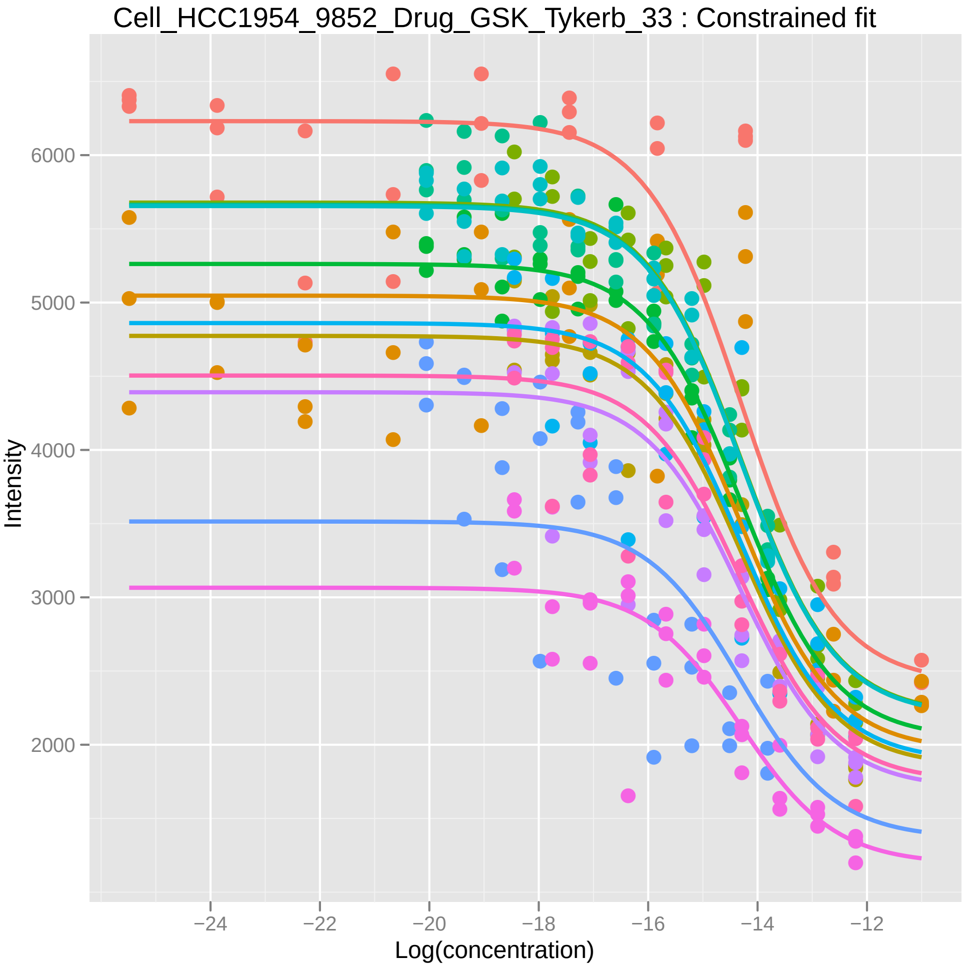


Figure 2.21: Fitted curves from constrained fitting for the 12 replicates of cell line HCC1954 and drug GSK Tykerb.

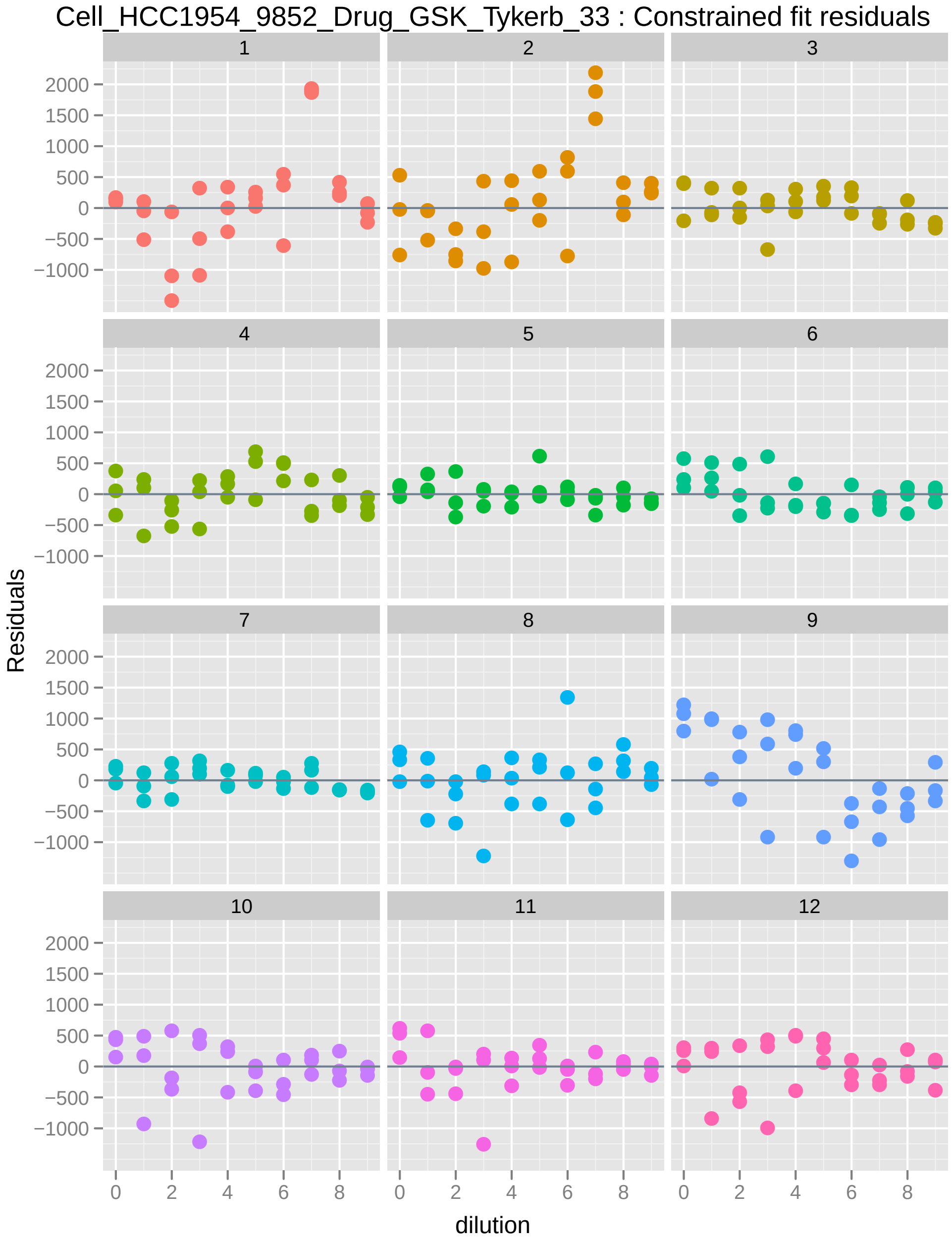


Figure 2.22: Residuals from constrained fitting for the 12 replicates of cell line HCC1954 and drug GSK Tykerb.

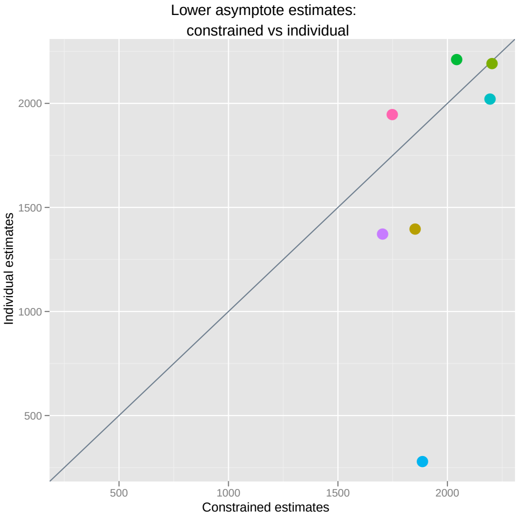


Figure 2.23: Cell line HCC1954 and drug GSK Tykerb : Constrained vs individual estimates of lower asymptote.

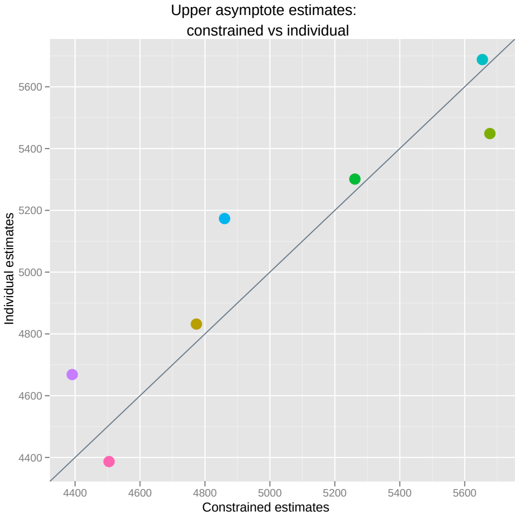


Figure 2.24: Cell line HCC1954 and drug GSK Tykerb : Constrained vs individual estimates of upper asymptote.

Chapter 3

Estimation of IC50: Crossed and nested mixed effects models

3.1 Introduction

The majority of drug screening assays assess the effects of toxicity on cell cultures only; however, in reality, inside the human body, the liver may metabolize these compounds into substances that can be more toxic than the parent compound. Typically, the enzymes responsible for the initial clearance of drugs from the body, called P450, will make potentially harmful substances more water soluble to be eliminated from the body easily; however, at other times, these enzymes can convert some compounds, like the common pain reliever acetaminophen, into a toxic chemical. Other compounds are prodrugs, which P450 enzymes must activate in order to be effective.

Current methods to screen for such effects involve using primary hepatocyte cultures, sometimes even slivers of liver, and generating metabolites using liver microsomes for subsequent toxicity screening. This process tends to be inconsistent as well as costly and time-consuming. So, as with the progress of biotechnology, the goal is the same: high throughput and automated assays that are more cost-effective and consistent.

The Clark lab at the Department of Chemical Engineering at UC Berkeley has been working on developing a miniaturized in-vitro assay system that can assess the toxicity of compounds and their metabolites and differentiate between them. The system consists of the coupling of two types of chips: one that houses the cell cultures and the other the contains the metabolizing enzymes.

3.1.1 DataChip and MetaChip

The DataChip (Data Analysis Toxicology Assay Chip) is a miniaturized 3D cell culture assay, which houses 1,080 individual cultures of Hep3B, human hematoma cells. In a mirrored format, the MetaChip (Metabolizing Enzyme Toxicology Assay

Chip) houses the immobilized metabolic enzymes encapsulated in a sol-gel, which are supposed to emulate the metabolic reactions in the human liver. In this way, many drug candidates can be tested simultaneously. Once the DataChip has been prepped and incubated on its own for 18 hours, it is stamped with the accompanying MetaChip and incubated together for 6 hours. Afterwards, the MetaChip is removed and discarded while the DataChip is placed in medium and incubated for an additional 18 hours before it is stained so a scanner can read the relative levels of living and dead cells in each well. For more details, see [14], [11], [13], and [12].

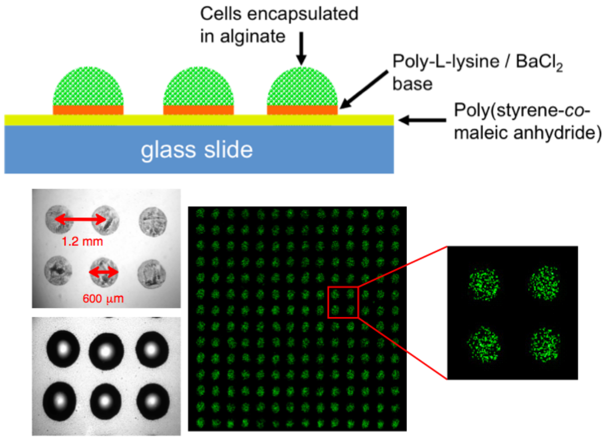

Cells encapsulated in alginate
Poly-L-lysine / BaCl$_2$ base
Poly(styrene-*co*-maleic anhydride)
glass slide
1.2 mm
600 µm

Figure 3.1: DataChip: 3D array of cells encapsulated in alginate gels.

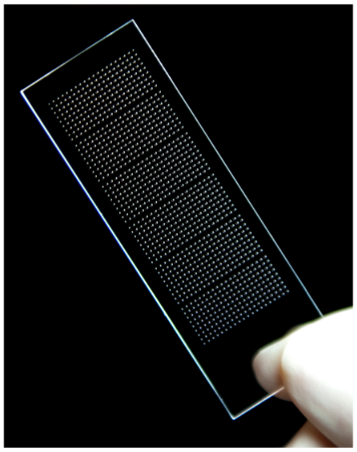


Figure 3.2: DataChip slide: A single chip can support 1080 nL-scale metabolic toxicity assays - the equivalent of eleven 96-well plates.

Enzyme conditions

Experiments involve four enzyme conditions:

- No enzymes
- Mix of P450s
- Mix of Phase II
- Mix of P450s and Phase II, also referred to as the “All Mix”

The isoforms of P450 are the most important as they catalyze first-pass (Phase I) reactions that can lead either to activation or inactivation of the drug. Whereas Phase I reactions can often result in active metabolites, enzymes that are involved in Phase II are typically detoxifying in nature and eventually enhance excretion of the compound [2].

Chip setup

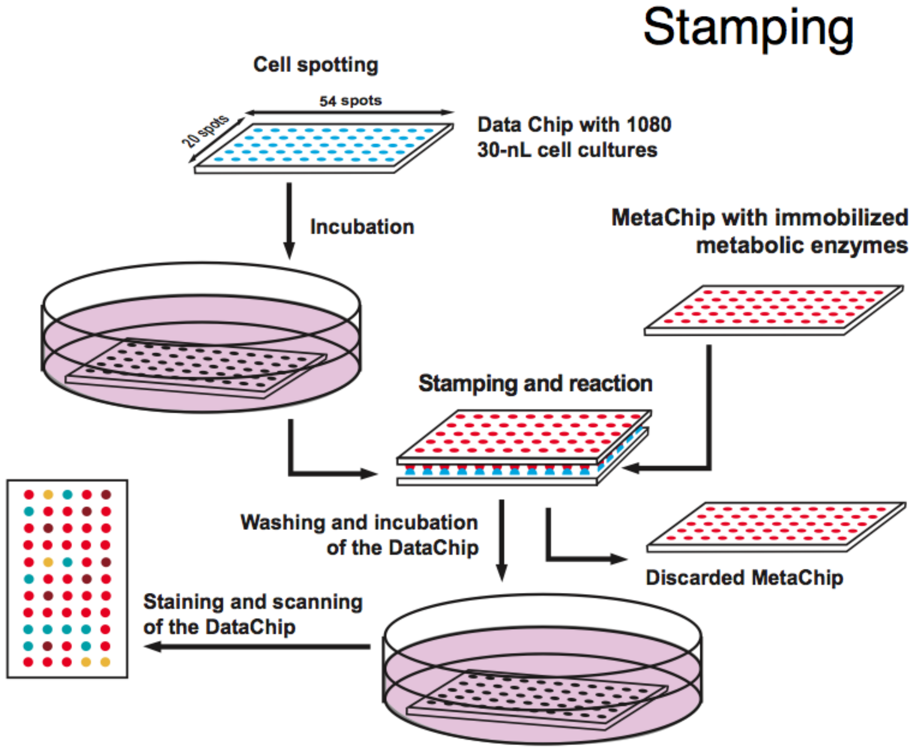


Figure 3.3: Procedure for DataChip and MetaChip to obtain measures of toxicity. The DataChip contains the cell cultures, while the MetaChip contains the metabolic enzymes. The two chips are then stamped together to produce a reaction. Afterwards, the MetaChip is removed, and the DataChip is incubated once more, stained and then scanned to obtain intensity values.

For the analysis, the layout of the chip consists of 1,080 observations. Each chip can include all combinations of 6 different compounds and 4 different enzyme conditions. Some chips may have a different format with 12 different compounds and 2 different enzyme conditions.

Each pairing of a compound and enzyme condition will be referred to hereafter as a unit; thus a chip will contain 24 units, which are blocks of 9 serial dilutions with 5 replicates each, for a total of 45 observations (Figure 3.4). As a check, 24 blocks multiplied by 45 observations equals 1,080.

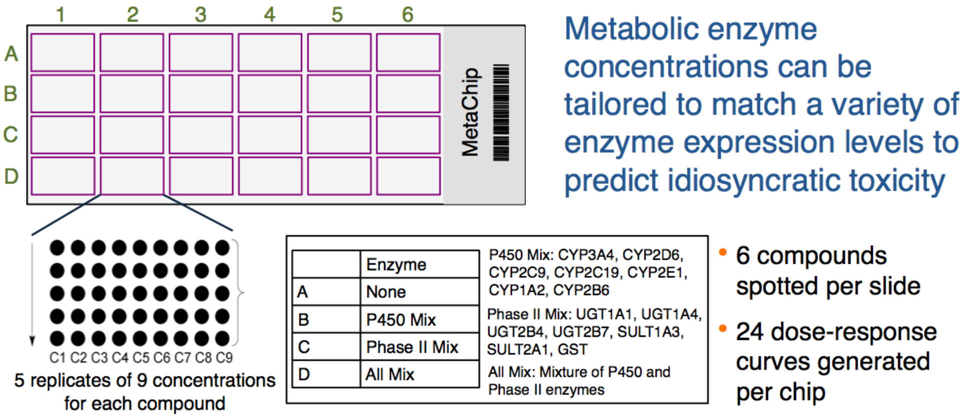
All observations from each chip get outputted to an Excel file (Figure 3.5), where it will be processed by R to be in a better format for analysis.

The example Excel file is easily read and converted to a data frame, where the first 6 lines are shown in Figure 3.6. Table 3.1 provides explanations for all the columns.

The data on a typical chip looks like the plot shown in Figure 3.7, where there are 24 units corresponding to 24 unique combinations of compound and enzyme condition. Ideally, an unit would have optimal range of concentrations of the compound in order

Name	Description	Values
compoundID	unique identifier for compounds	$1, \dots, 6$
enzymeRep	identifier for enzyme - replicate combination	$1, \dots, 20$
dilution	increasing number means increasing concentration	$1, \dots, 9$
value	intensity measuring cell viability	positive real numbers
compound	compound name	factor
startConc	starting concentration	positive real numbers
concentration	absolute concentration	positive real numbers
logconc	logged concentration (base 10)	real numbers
enzyme	enzyme condition	factor
replicate	identifier for replicate	$1, \dots, 5$
units	concentration units	concentration units
species	specifies which species	human, rat
file	full file name with path	string
sfile	shortened file name	string
funit	unit ID within a chip	$1, \dots, 24$
fileID	numeric identifier for file (chip)	positive integers
unit	unit identifier unique across entire batch of chips	positive integers
runit	unit identifier for unique combinations of compound and enzyme	positive integers
enz.comp	for use in nested modelling	factor
weight	weights for robust fitting	$[0, 1]$

Table 3.1: The representation of the columns in the data frame for the complete dataset.


1
2
3
4
5
6
A
B
C
D
MetaChip
C1 C2 C3 C4 C5 C6 C7 C8 C9
5 replicates of 9 concentrations for each compound

Metabolic enzyme concentrations can be tailored to match a variety of enzyme expression levels to predict idiosyncratic toxicity

	Enzyme
A	None
B	P450 Mix
C	Phase II Mix
D	All Mix

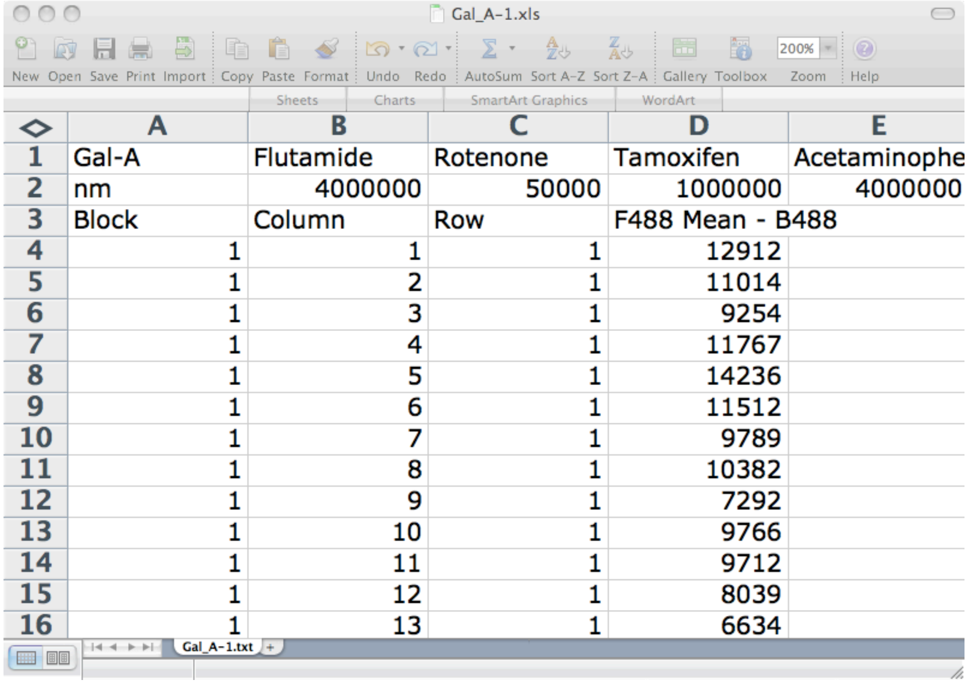
P450 Mix: CYP3A4, CYP2D6, CYP2C9, CYP2C19, CYP2E1, CYP1A2, CYP2B6

Phase II Mix: UGT1A1, UGT1A4, UGT2B4, UGT2B7, SULT1A3, SULT2A1, GST

All Mix: Mixture of P450 and Phase II enzymes

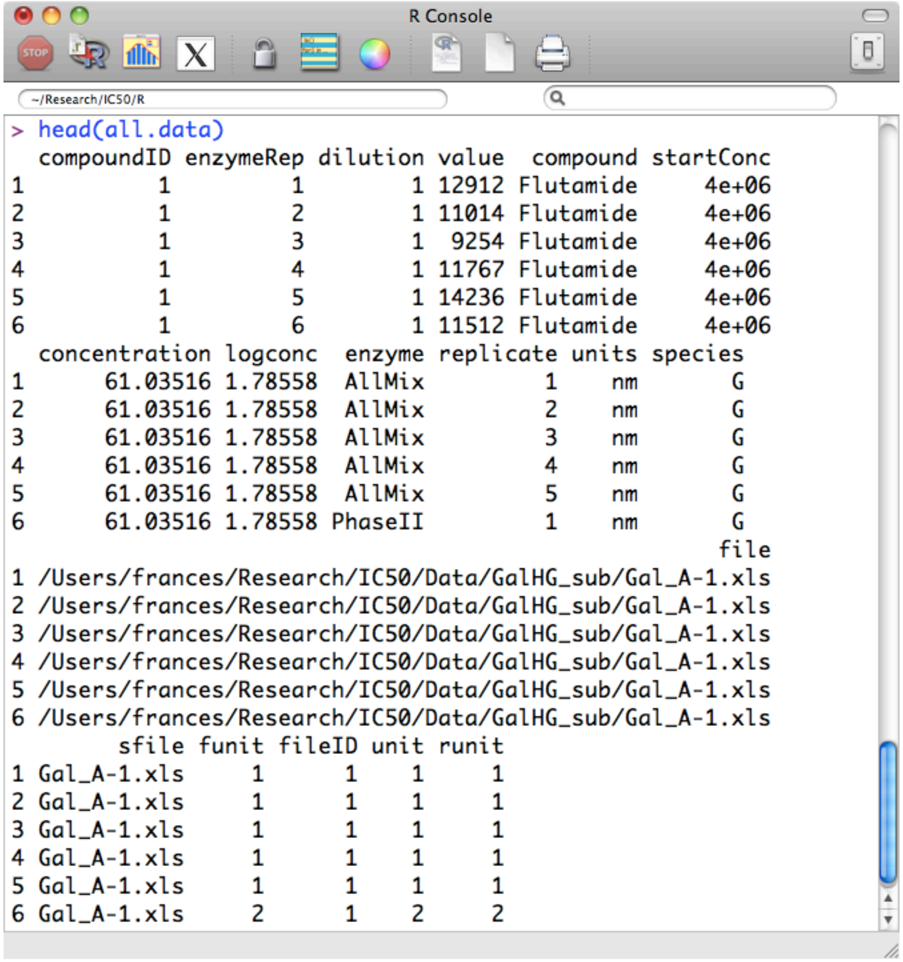
- 6 compounds spotted per slide
- 24 dose-response curves generated per chip

Figure 3.4: Layout of the MetaChip: The chip can test 6 compounds by 4 enzyme conditions for a total of 24 dose-response curves that each consists of 5 replicates of 9 four-fold dilutions.



	A	B	C	D	E
1	Gal-A	Flutamide	Rotenone	Tamoxifen	Acetaminophe
2	nm	4000000	50000	1000000	4000000
3	Block	Column	Row	F488 Mean - B488	
4	1	1	1	12912	
5	1	2	1	11014	
6	1	3	1	9254	
7	1	4	1	11767	
8	1	5	1	14236	
9	1	6	1	11512	
10	1	7	1	9789	
11	1	8	1	10382	
12	1	9	1	7292	
13	1	10	1	9766	
14	1	11	1	9712	
15	1	12	1	8039	
16	1	13	1	6634	

Figure 3.5: Direct output of data from the scanner consists of 4 columns of information: 1) Block refers to one of the 6 compounds; 2) Column refers to the enzyme and replicate (the order is determined by the set labeling of blocks on all chips); 3) Row indicates the dilution; 4) The important stuff: the intensity values as measured by the scanner. The first two lines in the file are added in to specify the compounds, enzymes, starting concentrations, and units.



```
> head(all.data)
  compoundID enzymeRep dilution value  compound startConc
1          1         1        1 12912 Flutamide     4e+06
2          1         2        1 11014 Flutamide     4e+06
3          1         3        1  9254 Flutamide     4e+06
4          1         4        1 11767 Flutamide     4e+06
5          1         5        1 14236 Flutamide     4e+06
6          1         6        1 11512 Flutamide     4e+06
  concentration logconc  enzyme replicate units species
1      61.03516 1.78558  AllMix         1    nm       G
2      61.03516 1.78558  AllMix         2    nm       G
3      61.03516 1.78558  AllMix         3    nm       G
4      61.03516 1.78558  AllMix         4    nm       G
5      61.03516 1.78558  AllMix         5    nm       G
6      61.03516 1.78558 PhaseII         1    nm       G
                                                   file
1 /Users/frances/Research/IC50/Data/GalHG_sub/Gal_A-1.xls
2 /Users/frances/Research/IC50/Data/GalHG_sub/Gal_A-1.xls
3 /Users/frances/Research/IC50/Data/GalHG_sub/Gal_A-1.xls
4 /Users/frances/Research/IC50/Data/GalHG_sub/Gal_A-1.xls
5 /Users/frances/Research/IC50/Data/GalHG_sub/Gal_A-1.xls
6 /Users/frances/Research/IC50/Data/GalHG_sub/Gal_A-1.xls
        sfile funit fileID unit runit
1 Gal_A-1.xls     1      1    1     1
2 Gal_A-1.xls     1      1    1     1
3 Gal_A-1.xls     1      1    1     1
4 Gal_A-1.xls     1      1    1     1
5 Gal_A-1.xls     1      1    1     1
6 Gal_A-1.xls     2      1    2     2
```

Figure 3.6: Screenshot of R console where the first six lines of the working dataset is displayed. See Table 3.1 for column descriptions.

to detect any activity. Sometimes there will be no effect as evidenced by a majority of the units with no trend on the example chip shown.

3.2 Current working dataset

Experiments are typically run in batches of several chips, and data is continually being generated on a massive scale. A method of analysis that can estimate IC50 values reasonably and quickly would be ideal. From viewing the data from the example chip, we can expect that we may be unable to fit every curve independently by unit. Nevertheless, we need a method that will be run on a set of several chips at a time. We thus focus on analyzing a working dataset that is representative of the data that is continually being produced in the lab. There are eight chips in this set, with a total of 8,640 observations for 22 different compounds across 4 different

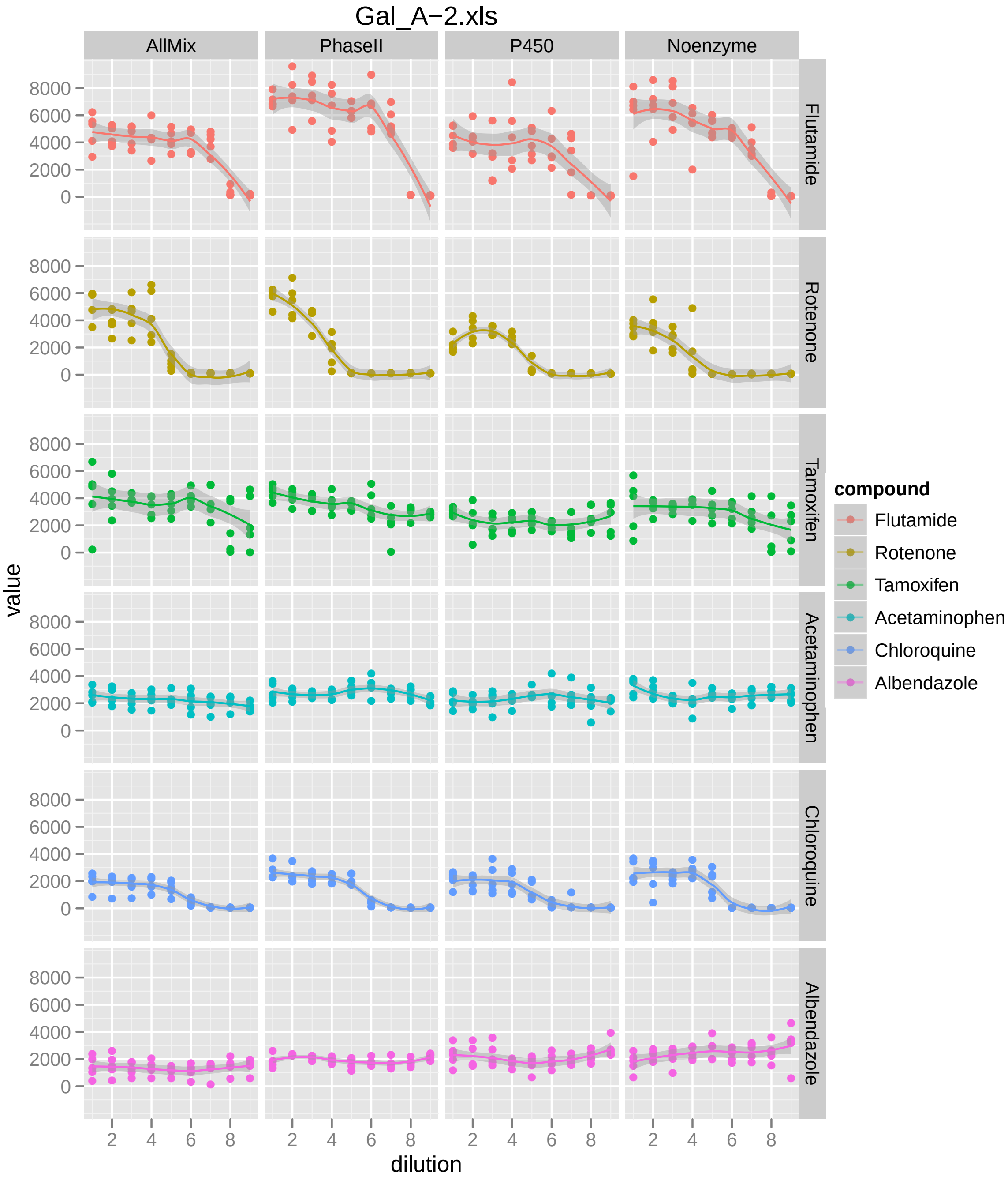


Figure 3.7: Plots of raw data from a sample Excel data file, organized such that the columns are by enzymes while the rows are by compounds. Smoothed lines through the data are provided for easier trend viewing.

enzyme conditions. Further structure of the dataset is displayed in Figure 3.8, where the blue boxes indicate the locations of the compounds across the chips, and Figure 3.9, which indicates the frequencies of specific compound-enzyme combinations.

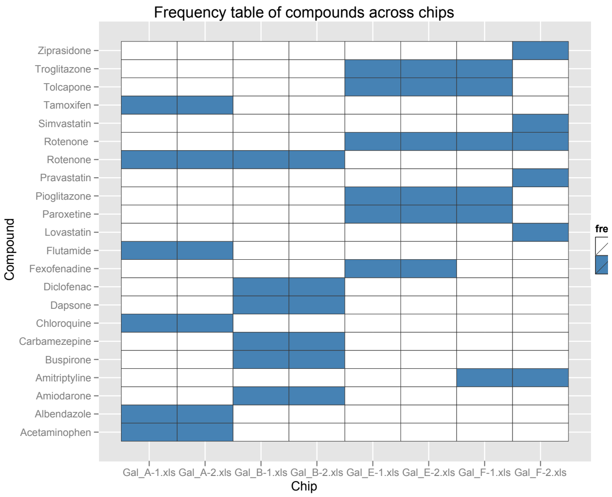


Figure 3.8: Mapping of 22 unique compounds across 8 chips in current working dataset. Blue indicates the presence of the compound.

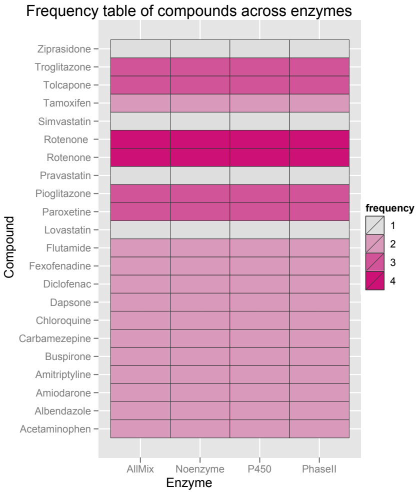


Figure 3.9: Frequencies of compound-enzyme combinations in the current dataset of 8 chips.

Edge problems

As the diagram of the chip (Figure 3.4) shows, the spots are arranged in the same, highly organized pattern on every single chip. What this means is that if anything occurs to affect or damage any area on the chip, it may effectively cause the loss of all information on a specific dilution since all the replicates are lined up next to each other. One of these problems is edge bleeding of the chip. Caused by inefficient drying of the chips in one of the experimental steps, edge bleeding means that the spots located near the edges of the chip will leak outside of their normal locations and thus, effectively diminish the intensity levels in those spots. This is clearly shown in the first two dilutions of the unit shown in Figure 3.13; the scanner is unable to pick up any signal from all 10 of spots. The obvious and common solution for this issue originated in the microarray world years ago: randomize the spots! Unfortunately, this data is not so, and we will discuss a fix for this in the analysis later.

3.3 Individual nonlinear least squares

The goal of the analysis of such data is to estimate IC50 values, the concentrations of the compound required to inhibit half of the cell growth. The first step in cases like this is to simply try fitting a logistic curve to each unit on its own and see what happens.

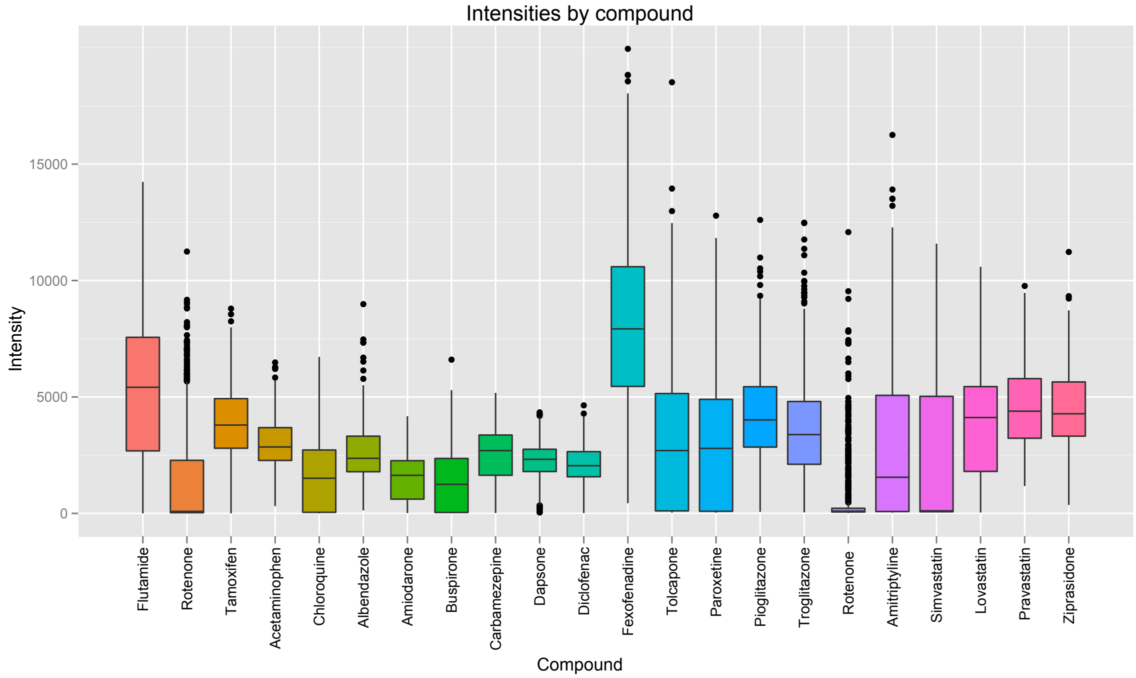


Figure 3.10: Boxplots of raw intensity values of the current working dataset by compound.

We randomly pick an unit where we can see a clear decreasing trend. This unit, as shown in Figure 3.14, is labelled 132, with compound Paroxetine (more popularly known as Paxil, an antidepressant), no enzymes, and on the chip Gal E-1. On visual inspection of all units of data, the lowest values all tend to zero, so we will use a three parameter logistic curve here, where we estimate the upper asymptote ϕ_1, IC50 ϕ_2, and slope ϕ_3.

$$y_{rs} = \frac{\phi_1}{1 + \exp\left(-\frac{x_r - \phi_2}{\phi_3}\right)} + \epsilon_{rs} \tag{3.1}$$

where r and s reference the dilutions and replicates, respectively, and $\epsilon_{rs} \sim \mathcal{N}(0, \sigma^2)$.

The parameters, $\phi_k, k = 1, 2, 3$, can be estimated by minimizing the nonlinear least squares:

$$\sum_{r=1}^{9} \sum_{s=1}^{5} (y_{rs} - f(x_r, \phi_k))^2 \tag{3.2}$$

where f is the three parameter logistic curve function. Nonlinear least squares has been implemented in R [18] as the nls() function, whose default is the Gauss-Newton algorithm. Initial values can also be provided automatically with its self starting function for the three parameter logistic curve, SSlogis().

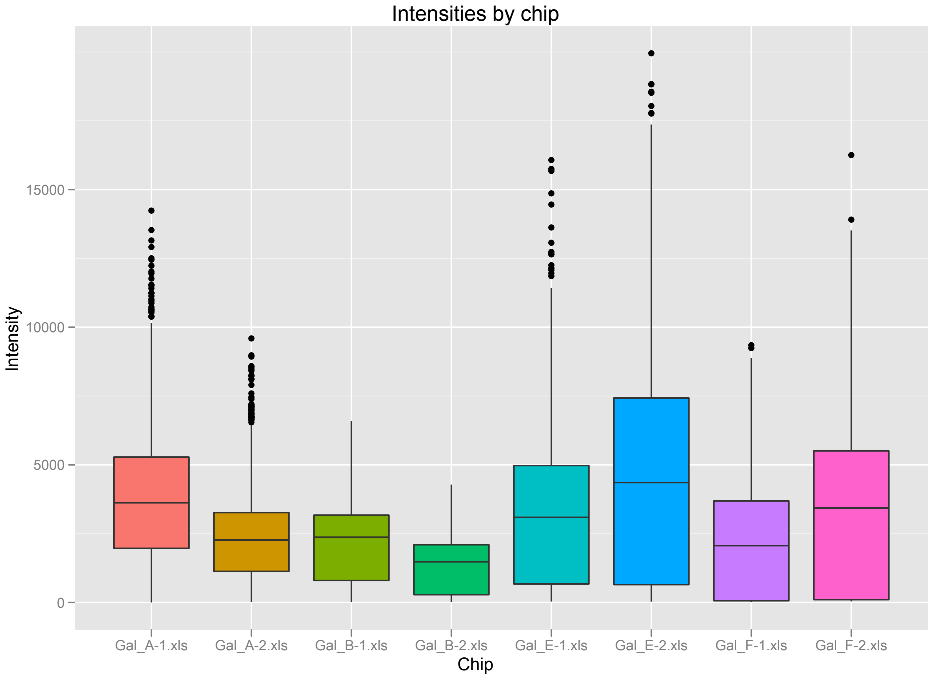


Figure 3.11: Boxplots of raw intensity values of the current working dataset by chip.

This function calculates initial values through a method that takes advantage of the partial linearity of the logistic curve function. We can rearrange the equation such that the right hand side is linear.

$$\log\left(\frac{y/\phi_1}{1-y/\phi_1}\right) = \frac{1}{\phi_3}x - \frac{\phi_2}{\phi_3} \tag{3.3}$$

Then, if we let $z = \log\left(\frac{y/\phi_1}{1-y/\phi_1}\right)$, we can rearrange once more such that ϕ_2 and ϕ_3 become linear parameters:

$$x = \phi_2 + \phi_3 z \tag{3.4}$$

Since we do not know what ϕ_1 is, we substitute ϕ_1 with $\tilde{\phi_1} = 1.05 * \max(y)$ so that we can fit the linear model 3.4 to obtain rough values for ϕ_2 and ϕ_3, $\tilde{\phi_2}$ and $\tilde{\phi_3}$.

These values, $\tilde{\phi_2}$ and $\tilde{\phi_3}$, will then be supplied as initial values for the fitting of a two-parameter logistic curve with the original observations. The "plinear" algorithm in nls() will estimate ϕ_2 and ϕ_3, as well as ϕ_1, the linear parameter in the logistic curve. The majority of the time, these initial values are extremely close to the ultimate estimates.

By running nls(), we obtain the estimates, $\hat{\phi_1} = 4936.25$, $\hat{\phi_2} = 5.41$ and $\hat{\phi_3} = -0.46$.

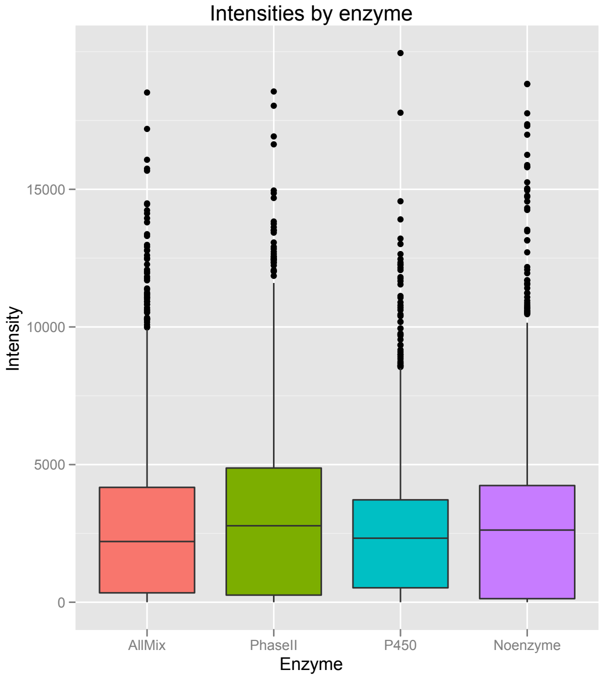


Figure 3.12: Boxplots of raw intensity values of the current working dataset by enzyme.

Robust fitting of the data is available with robust M-estimators, using iterated reweighted least squares (IWLS), through the nlrob() function in robustbase package. Three ψ functions are available for use: Huber's (default), Hampel, and the Tukey bisquare. With the default ψ function, we obtain the following robust estimates, which are quite similar to the original ones. The slope estimate seems the most dramatically different.

Upper asym	IC50	Slope
4870.71	5.37	-0.35

Both the ordinary and robust fitted curves are plotted in Figure 3.15, where they seem to fit the points reasonably and similarly.

3.3.1 Results on working dataset

We continue with fitting the rest of the 192 units on the 8 chips that we are currently dealing with, where even with robust fitting, there are over half of the units that cannot be fitted.

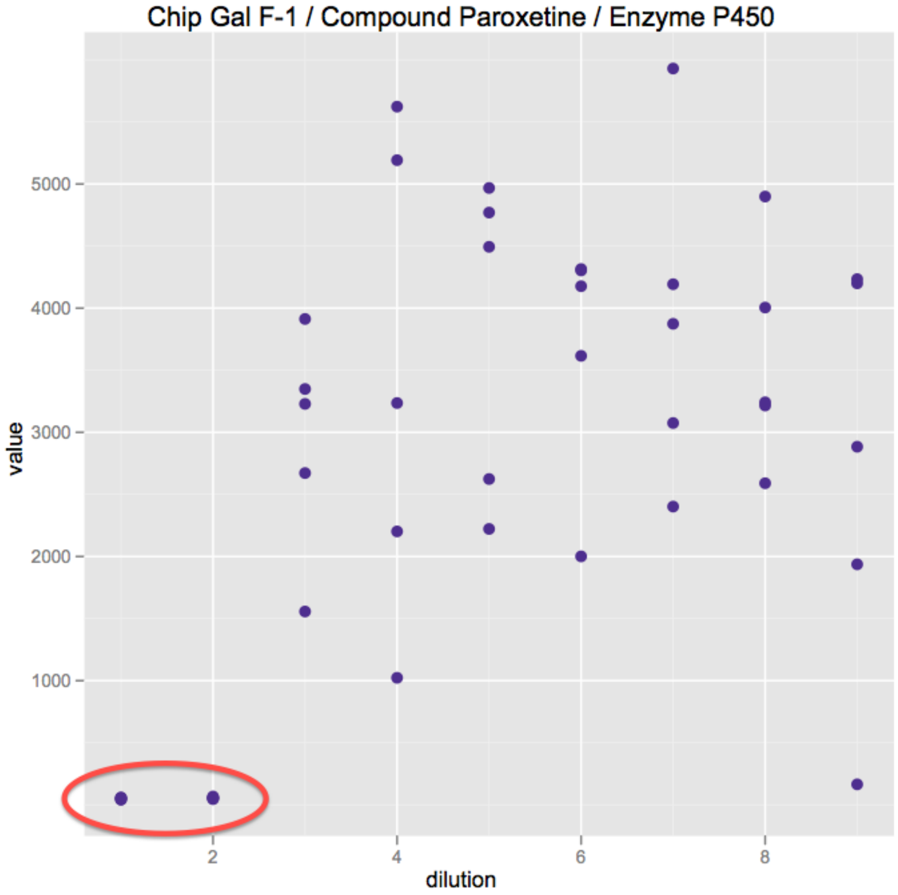


Figure 3.13: Example of experimental problem of edge bleeding.

	Fitted	Failed	Percent success
Ordinary	72	120	37.5%
Robust	93	99	48.4%

Figure 3.16 displays the results for the first of the 8 chips, where robust nonlinear least squares can fit successfully only 6 out of the 24 units. Some units are so flat that it is obvious why we are unable to fit a curve, while other units are a bit more difficult to explain, but a majority of these cases occur because of identiafiability issues with the IC50 and slope parameters. There just isn't sufficient information in the area where these two parameters have to be estimated from. Clearly, this is not the best method of analyzing such data. We cannot rely on sufficient data for each unit to be even able to fit a logistic curve on more than half of the dataset. The next step is methods that involve the sharing of information throughout the entire dataset.

3.4 Two-stage methods

In the rest of this chapter, we focus on methods that take advantage of the entire dataset to fit logistic curves to each unit. These methods are generally done in two stages: the first stage being individual nonlinear least squares and subsequent summarization of its results, and the second stage being the utilization of these results in a mixed effects model. One of these methods for nested structures called the global two-stage method has been well developed and implemented by Davidian and Giltinan [9]. We fit our data in a similar vein.

3.4.1 Model 0: Grouped by unit

We now discuss a naive approach as an intermediate step on the way to discussion and implementation of more complex and complete models. In this situation, the simplest way to try and fit curves on every unit by sharing information is to consider the unit factor as a random effect. Instead of looking at each unit on its own, we now specify that these units are a random sample from a much larger population of units, which in a sense, is true in this case since we are not limited to this current working dataset.

The first stage involves the fitting of each unit independently with nonlinear least squares as shown in the previous section. The resulting estimates are then used as data to robustly estimate the covariance components needed in a mixed effects model that will fit the units together. Mixed effects here can be considered a way to balance the individual fitting of each unit where some estimates can be extreme or nonexistent, and the fitting of a common curve to every unit where there is sufficient data but also loss of information from units individually.
Let us first define the notation:

y		observations
x		dilutions (log scale)
ϕ		linear combination of fixed and random effects
β		fixed effects
b		random effects
ϵ		error
σ^2		variance of error
σ_b^2		variance of the random effects
n		total number of observations
i	$1,\ldots,n_{comp}$	indexes compound
j	$1,\ldots,n_{enz}$	indexes enzyme conditions
k	$1,\ldots,n_{chip}$	indexes chip
l	$1,\ldots,n_{ijk}$	indexes any replicates of a compound-enzyme combination within a chip
r	$1,\ldots,9$	indexes dilution
s	$1,\ldots,5$	indexes replicates within a dilution

Since this current model will only be referring to the data by unit, we shall try to simplify things by defining i^* to index units, which in later sections will be implied by any unique combination of the original indices i, j, k, l.

$$\begin{aligned} y_{i*}(r,s) &= f(x_{i^*}(r,s), \phi_{i^*}) + \epsilon_{i^*}(r,s) && (3.5)\\ &= \frac{\phi_{1,i^*}}{1+\exp\left(-\frac{x_{i^*}(r,s)-\phi_{2,i^*}}{\phi_{3,i^*}}\right)} + \epsilon_{i^*}(r,s) && (3.6)\\ &= \frac{\beta_1 + b_{1,i^*}}{1+\exp\left(-\frac{x_{i*}(r,s)-(\beta_2+b_{2,i^*})}{\beta_3+b_{3,i^*}}\right)} + \epsilon_{i^*}(r,s) && (3.7)\\ \epsilon_{i^*} &\sim \mathcal{N}(0,\sigma^2) && (3.8) \end{aligned}$$

This model fits a logistic curve to each unit i^* with three parameters, ϕ_{1,i^*} (upper asymptote), ϕ_{2,i^*} (IC50), and ϕ_{3,i^*} (slope). These are only parameters in the sense that each are composed of a parameter that is common across all units called a fixed effect and a random variable that takes on a value specific to the unit called a random effect. The fixed effects parameter, β, can be viewed essentially as the mean vector of the parameters, while the random effects are the deviations from the mean.

$$\phi_{i^*} = \beta + b_{i^*} \tag{3.9}$$

$$\begin{bmatrix} \phi_{1,i^*} \\ \phi_{2,i^*} \\ \phi_{3,i^*} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} b_{1,i^*} \\ b_{2,i^*} \\ b_{3,i^*} \end{bmatrix} \tag{3.10}$$

The random effects are assumed to be normally distributed as such:

$$b_{i^*} \sim \mathcal{N}(0, \sigma^2\Sigma) \tag{3.11}$$

where Σ is a $q \times q$ relative covariance matrix. The dimension q is the total number of random effects to be predicted, which in this case is the number of units (192) multiplied by 3 parameters (576).

For the random effects, b_{i^*}, in the same unit i^*, the covariance matrix is

$$G = \sigma^2 \begin{bmatrix} \sigma_{b,11} & \sigma_{b,12} & \sigma_{b,13} \\ \sigma_{b,21} & \sigma_{b,22} & \sigma_{b,23} \\ \sigma_{b,31} & \sigma_{b,32} & \sigma_{b,33} \end{bmatrix} \tag{3.12}$$

where the diagonal elements are the variances of the random effects of the upper asymptote, IC50, and slope, and the off-diagonal elements are the covariances. This is true for all units, and the covariance of any two random effects that are not within the same unit is zero.

Thus the structure of Σ in this case is elucidated as $\sigma^2\Sigma$ is just the Kronecker product of G and the identity matrix I_q.

Note that if $\sigma_{b,11} = 0$, this is equivalent to fitting a common upper asymptote to all units; this similarly applies to all other parameters.

3.4.2 Likelihood

In order to eventually find the conditional means of these random effects given the data, we first discuss the likelihood of this model. Recall that $\mathbf{y}$ is the $n \times 1$ vector of observations. Given the vector of random effects $\mathbf{b}$,

$$\mathbf{y}|\mathbf{b} \sim \mathcal{N}(f(\boldsymbol{\beta}, \mathbf{b}), \sigma^2 I_n) \tag{3.13}$$

Further the random effects have distribution

$$\mathbf{b} \sim \mathcal{N}(0, \sigma^2\Sigma) \tag{3.14}$$

The likelihood of the data as a function of the parameters is given by

$$L(\boldsymbol{\beta}, \Sigma, \sigma^2|\mathbf{y}) = p(\mathbf{y}|\boldsymbol{\beta}, \Sigma, \sigma^2) \tag{3.15}$$

where $p(\mathbf{y}|\boldsymbol{\beta}, \Sigma, \sigma^2)$ is the density of the random vector $\mathbf{y}$ given the parameters $\boldsymbol{\beta}, \Sigma, \sigma^2$. Now note that

$$p(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta}, \Sigma, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{[\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{b})]^T[y - f(\boldsymbol{\beta}, \mathbf{b})])}{2\sigma^2}\right) \tag{3.16}$$

$$p(\mathbf{b}|\Sigma, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{q/2}|\Sigma|^{q/2}} \exp\left(-\frac{\mathbf{b}^T\Sigma^{-1}\mathbf{b}}{2\sigma^2}\right) \tag{3.17}$$

Then

$$p(\mathbf{y}|\boldsymbol{\beta}, \Sigma, \sigma^2) = \int p(\mathbf{y}|\mathbf{b}, \boldsymbol{\beta}, \sigma^2) p(\mathbf{b}|\Sigma, \sigma^2) db \tag{3.18}$$

$$= \frac{1}{(2\pi\sigma)^{(n+q)/2}|\Sigma|^{q/2}} \int \exp\left(-\frac{[\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{b})]^T[\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{b})] + \mathbf{b}^T\Sigma^{-1}\mathbf{b}}{2\sigma^2}\right) db \tag{3.19}$$

Given Σ, we then find the values of $\boldsymbol{\beta}$ and $\mathbf{b}$ that minimize the penalized sums of squares [20]

$$[\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{b})]^T[\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{b})] + \mathbf{b}^T\Sigma^{-1}\mathbf{b} \tag{3.20}$$

3.4.3 Spherical transformation of random effects

The random effects, $\mathbf{b}$, can be defined as a linearly transformed "spherical" normally distributed random variable, $\mathbf{u}$, in which the contours of the density are concentric spheres centered at zero [5]. If we let Λ be the Cholesky factor of Σ such that $\Sigma = \Lambda^T\Lambda$, then

$$\mathbf{b} = \Lambda\mathbf{u}, \qquad \mathbf{u} \sim \mathcal{N}(0, I_q) \tag{3.21}$$

And so the expression 3.20 is equivalent to

$$[\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{b})]^T[\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{b})] + \mathbf{b}^T(\Lambda^{-1})^T(\Lambda^{-1})\mathbf{b} \tag{3.22}$$

so that we now find the values of β and u that minimize

$$[\mathbf{y} - f(\boldsymbol{\beta}, \Lambda, \mathbf{u})]^T[\mathbf{y} - f(\boldsymbol{\beta}, \Lambda, \mathbf{u})] + \mathbf{u}^T\mathbf{u} \tag{3.23}$$

3.4.4 Estimation of fixed effects and covariance components

The fixed effects, β_1, β_2, and β_3, are estimated with weighted means of the individual nonlinear least squares estimates where the weights are the precisions from those fits.

$$\widehat{\beta_h} = \frac{\sum_{i^*=1}^{n_{\text{unit}}} \widehat{w_{i^*}} \times \widehat{\phi_{hi^*}}}{\sum_{i^*=1}^{n_{\text{unit}}} \widehat{w_{i^*}}}, \qquad h = 1, 2, 3 \tag{3.24}$$

where

$$\widehat{w_{i^*}} = \frac{1}{\widehat{\sigma^2_{i^*}}} \tag{3.25}$$

We do not estimate the fixed effects with the random effects for a simple reason that will be explained later.

	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$
Fixed effects estimates	4146.81	5.88	-0.37

We delay the discussion of the estimation of covariance components in a later section when it can be more complete. The robust estimates of the covariance $(\hat{\Sigma})$ and correlation matrix of the random effects for this model, calculated from the individual nonlinear least squares estimates, are

	Robust covariance			Robust correlation		
	b_1	b_2	b_3	b_1	b_2	b_3
b_1	5435311.76	-1370.79	100.37	1.00	-0.21	0.10
b_2	-1370.79	7.72	-0.65	-0.21	1.00	-0.54
b_3	100.37	-0.65	0.18	0.10	-0.54	1.00

One particular thing to note from the correlation matrix is that the IC50 and slope parameters are generally moderately correlated; depending on the dataset, we have seen around -0.4 to -0.7. The correlation between the upper asymptote and these two is much lower, around 0 to 0.2. Since the estimation of the asymptote and the IC50 and slope typically rely on data points in different areas of the logistic curve, this is consistent with expectations.

3.4.5 Penalized least squares

There are many methods available to minimize the penalized least squares 3.23. For this model, one of the simplest is to utilize the optimization function in R called optim() and one of its available methods L-BFGS-B, which allows box constraints [6].

Given values for the fixed effects and covariance matrix, the minimization is independent of the units. So the prediction of the random effects, $\mathbf{b}_{i^*}$ for the i^*th unit can be accomplished by minimizing the penalized nonlinear least squares with only the unit data.

The predicted spherical random effects $\hat{u_{i^*}}$ will minimize

$$[\mathbf{y_{i^*}} - f(\hat{\boldsymbol{\beta}}, \hat{\Lambda}, \mathbf{u_{i^*}})]^T[\mathbf{y_{i^*}} - f(\hat{\boldsymbol{\beta}}, \hat{\Lambda}, \mathbf{u_{i^*}})] + \mathbf{u_{i^*}}^T\mathbf{u_{i^*}} \tag{3.26}$$

and then the original random effects b_{i^*} are calculated

$$\hat{\mathbf{b}_{i^*}} = \hat{\Lambda}\hat{\mathbf{u}_{i^*}} \tag{3.27}$$

The fitted curves of the first chip are included in Figure 3.17 as a visual check of the results. The method was able to fit a curve to each unit, and through checking all the fitted curves of the eight files (included in the appendix), we see that a great majority of the units have fitted curves that follow the data points well. However there are a few that have completely unreasonable fits even though they seem like it would have been able to have been fitted. In addition, there are cases where the slope estimates are somewhat too steep; there is potential for improved estimates.

The advantages of this method can be seen in its computational simplicity and speed. The fact that each unit is processed one at a time means that any unit that happens to be an outlier and unable to be reasonably fitted will not be able to spoil or interrupt the fitting of the rest of the units in the batch. It is not an all or none proposition that may leave the lab with no results even if the majority of the data is close to ideal.

However, it does not allow us to analyze how the IC50 values are affected by the compounds and enzyme conditions. It does not take into account of the potential differences that data on separate chips may encounter. So while this was a step up from the failures of individual nonlinear least squares, there needs to be a balance between a simple model that can estimate the IC50 easily without regard for the inherent characteristics of the data and a complex model that interprets the IC50 in the correct way but may not be practical in reality.

3.5 Model 1: Partially crossed random effects

The previous model was able to obtain decent IC50 estimates most of the time, but we can hope to do much better by using more of the information that we have about the dataset. It is time to take a closer look at how we can utilize information about compounds, enzyme conditions, and even which chip observations lie on.

3.5.1 Design of model

This section will discuss the development of the model. In order to fit curves specific to factors in the data, we have to determine how the three parameters in the logistic curve are affected by the factors. Then we may be able to estimate the parameters through components related to each factor. Through empirical analyses of the individual nonlinear least squares estimates, we can see through scatter and boxplots whether we need to include specific random effects. And of course, as with all modelling, the decisions here are never definitive and may change during and/or after fitting. Increasing complexity in the model means longer fitting time and perhaps less chance of converging to reasonable estimates, since the same number of observations are being used to estimate more and more things. The best way is to start simply and then build up.

	Upper asymptote	IC50	Slope
Compound	yes	yes	yes
Enzyme	no	yes	no
Chip	yes	no	no

Table 3.2: Table of the inclusion of random effects by factor for the three parameters.

We suggest the entries in the table above (Table 3.2) because of the following plots of the individual nonlinear least squares estimates. They can help but because individual sets of estimates can tend to be extreme, the estimates are most likely more varied than expected. Also, the model should be general enough to be sufficient for similar datasets. Another option is to use the results from Model 0, as there are more estimates available. This can be focus of comparison later, but at present, results from individual estimates suffice. In general, we should take note of how well these estimates are doing as well as on what proportion of the dataset.

There are four sets of plots to address the upper asymptote parameter. Boxplots that compare the individual estimates by compound (Figure 3.18) show that there is a clear difference. However because of the mapping of the compounds to chips, we also have to be sure that this is not an artifact of the chips instead. Boxplots of the estimates by chip (Figure 3.19) do indeed show a similar pattern. We can take a closer look in Figure 3.20 to see that the pattern is mostly due to variation in chips.

Simple linear mixed effects models can be fit here to determine whether to include the compound factor in our full model. The inclusion of both compound and chip factors results in similar estimates of the variance. The exclusion of the compound factor increases the residual variance considerably more than the variance of the chip random effect.

Group	With compound SD	Without compound SD
Compound	1521.5	NA
Chip	1574.7	1765.5
Residual	1290.2	1937.0

We can do a simple check with a likelihood ratio test that shows that the inclusion of the compound factor does significantly improve the likelihood. It would be best to start with including both in our model.

Model	Df	AIC	BIC	logLik	χ^2	χ^2 Df	P($> \chi^2$)
Without compound	3	1695.4	1703.0	-844.7			
With compound	4	1659.0	1669.2	-825.5	38.4	1	5.8e-10

The enzyme factor does not seem to have an effect on the upper asymptote (Figure 3.21), which is reasonable since the starting population of living cells wouldn't depend on the enzyme conditions.

The model should allow for the estimation of a different IC50 for each compound - enzyme combination but not for replicates within a combination. Plot of the individual estimates indicate that we may be able to obtain unique estimates for the majority of the units (Figure 3.23), while there isn't likely a need to adjust for chip effects (Figure 3.24).

Although we would have expected that the slope parameter depends heavily on the enzyme conditions, we do not see any differences in the estimates when separated by enzyme levels (Figure 3.22). They do not vary by chip much either (Figure 3.26).

So at present, we include a compound effect for the slope parameter (Figure 3.25). The inclusion and exclusion of effects is never a set game plan; it depends on a lot of factors and can continually change due to practical constraints in the fitting.

Recall that we consider all the observations in this dataset as $y_{ijkl}(r, s)$ where

i $= 1, \ldots, n_{comp}$, indexes the compounds,
j $= 1, \ldots, n_{enz}$, indexes the enzyme conditions,
k $= 1, \ldots, n_{chip}$, indexes the chip,
l $= 1, \ldots, n_{ijk}$, indexes any replicates of a compound / enzyme combination on the same chip
r $= 1, \ldots, 9$, indexes the 4-fold serial dilutions of the compound
s $= 1, \ldots, 5$, indexes the replicates for each dilution

The specific model is a logistic curve with three parameters that define the upper asymptote ϕ_1, x-value of the midpoint ϕ_2, i.e. IC50 here, and the slope at the midpoint ϕ_3. In following Table 3.2, we include random effects for the upper asymptote by compound and by chip, for IC50 by compound and enzyme, and for slope by compound.

$$\begin{aligned} y_{ijk}(r,s) &= f(x(r,s); \boldsymbol{\phi}_{ijk}) + \epsilon_{ijk}(r,s) && (3.28) \\ &= \frac{\phi_{1,ik}}{1 + \exp\left(-\frac{(x(r,s) - \phi_{2,ij})}{\phi_{3,i}}\right)} + \epsilon_{ijk}(r,s) && (3.29) \end{aligned}$$

$$\begin{aligned} \phi_{ijk} = \begin{bmatrix} \phi_{1,ik} \\ \phi_{2,ij} \\ \phi_{3,i} \end{bmatrix} &= \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} a_{1,i} \\ a_{2,i} \\ a_{3,i} \end{bmatrix} + \begin{bmatrix} 0 \\ b_{2,j} \\ 0 \end{bmatrix} + \begin{bmatrix} c_{1,k} \\ 0 \\ 0 \end{bmatrix} && (3.30) \\ &= \beta + A_i a_i + B_j b_j + C_k c_k && (3.31) \end{aligned}$$

where A_i is the 3×3 identity matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$, $B_j = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and $C_k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$.

$$\mathbf{a_i} = \begin{bmatrix} a_{1,i} \\ a_{2,i} \\ a_{3,i} \end{bmatrix} \quad \mathbf{b_j} = \begin{bmatrix} b_{1,j} \\ b_{2,j} \\ b_{3,j} \end{bmatrix} \quad \mathbf{c_k} = \begin{bmatrix} c_{1,k} \\ c_{2,k} \\ c_{3,k} \end{bmatrix} \qquad (3.32)$$

$$a_i \sim \mathcal{N}(0, \Sigma_a) \quad b_j \sim \mathcal{N}(0, \sigma_b^2) \quad c_k \sim \mathcal{N}(0, \sigma_c^2) \qquad (3.33)$$

$$\epsilon_{ijk}(r,s) \sim \mathcal{N}(0, \sigma^2) \qquad (3.34)$$

$$\Sigma = \begin{bmatrix} \sigma^{-2}\Sigma_a \otimes I_{n_{comp}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_b^2}{\sigma^2} \otimes I_{n_{enz}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{\sigma_c^2}{\sigma^2} \otimes I_{n_{chip}} \end{bmatrix} \tag{3.35}$$

Random effects associated with different factors and with different levels of the same factor are assumed to be independent and identically distributed. Σ for this report is a $q \times q$ matrix, q being the total number of random effects and calculated here as (number of compounds × 3) + (number of enzymes × 1) + (number of chips × 1) = (22 × 3) + (4 × 1) + (8 × 1) = 78. The structure of Σ can be visualized in Figure 3.27.

We can write the parameters of this model in matrix notation as

$$\boldsymbol{\phi} = X\boldsymbol{\beta} + Z \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \\ \mathbf{c} \end{bmatrix} \tag{3.36}$$

of dimension $pn \times 1$ with design matrices: X, dimension $pn \times p$, and Z, dimension $pn \times q$, where $p = 3$, the number of fixed effects.

$\boldsymbol{\phi}$ can be put into matrix form as Φ with dimensions $n \times p$ such that the elements of ϕ fill column-wise. Each row of Φ is thus the set of {upper asymptote, IC50, slope} values for a specific observation.

3.5.2 Estimation of covariance matrix

The estimation of Σ and σ is dependent on the results of the individual fits. Because the compound factor is involved in all three parameters, a 3 × 3 covariance matrix is needed in Σ. Since two of the parameters each have random effects by two factors, the contributions have to be separated. This can be done through a linear mixed effects model for each of these parameters. For the upper asymptote,

$$\widehat{\phi}_{1ik,nls} = \beta_1 + a_{1,i} + c_{1,k} + \epsilon_{1ik} \tag{3.37}$$
$$\tag{3.38}$$

and for the IC50,

$$\widehat{\phi}_{2ij,nls} = \beta_2 + a_{2,i} + b_{2,j} + \epsilon_{2ij} \tag{3.39}$$

The estimates are considered the data here where they are decomposed into the mean values, β_1 and β_2, the effects by compound, $a_{1,i}$ and $a_{2,i}$, the effects by enzyme, $b_{2,j}$ and the effects by chip, $c_{1,k}$. If we define

$$\widehat{\mathbf{a}}_{\mathbf{3,i}} = \text{median}(\widehat{\phi}_3) - \text{median}(\widehat{\phi}_{3,i}) \tag{3.40}$$

Compound	Upper asymptote	IC50	Slope
Flutamide	1747.00	0.34	-0.14
Rotenone	51.33	-1.89	-0.10
Tamoxifen	-480.82	1.00	0.65
Acetaminophen	-1097.65	-0.24	-0.18
Chloroquine	-1755.48	-0.57	-0.00
Amiodarone	-806.88	0.52	-0.08
Buspirone	-42.36	-0.35	-0.11
Carbamezepine	-65.88	-0.67	-0.24
Dapsone	211.32	1.91	6.96
Diclofenac	-529.38	2.41	0.79
Fexofenadine	4217.70	1.55	1.58
Tolcapone	-224.16	-0.45	-0.06
Paroxetine	-809.26	-1.51	-0.25
Pioglitazone	-562.82	1.18	1.13
Troglitazone	-1400.36	1.78	0.16
Rotenone	312.32	-1.93	0.31
Amitriptyline	1793.66	-0.41	-0.19
Simvastatin	51.46	-1.37	-0.08
Lovastatin	-751.41	-1.29	-0.73
Pravastatin	939.82	0.90	4.90
Ziprasidone	-798.17	-0.91	-0.16

Table 3.3: Compound effects used to help estimate covariance matrix.

Then $\widehat{\Sigma}_a$ (Table 3.4) is defined as the robust covariance of the $n_{comp} \times 3$ matrix $\left[\begin{array}{ccc} \widehat{\mathbf{a}}_\mathbf{1} & \widehat{\mathbf{a}}_\mathbf{2} & \widehat{\mathbf{a}}_\mathbf{3} \end{array}\right]$ (Table 3.3).

To see whether these numbers in Table 3.4 are reasonable, we can look at the square of the diagonal to see that the standard deviations, 1673.84, 2.74, and 0.27, are within scale of the parameter estimates.

The estimates for the other two variances of the random effects, σ_b^2 and σ_c^2 are obtained from the linear mixed effects results.

And lastly, $\widehat{\sigma}$ is calculated from the estimated residual errors of the individual nonlinear least squares fittings.

$$\widehat{\sigma} = \sqrt{\frac{1}{n_\#} \sum_{i_\#=1}^{n_\#} \widehat{\sigma}^2_{i_\#,nls}} = \sqrt{1091311} = 1044.67 \tag{3.41}$$

where $i_\#$ indexes the $n_\#$ units that were able to be fitted with nonlinear least squares.

Thus, the estimated covariance matrix to be used in the next stage is composed of the elements in the table 3.6.

	Upper asym	IC50	Slope
Upper asymp	2801728.01	-1832.57	-150.76
IC50	-1832.57	7.50	-0.16
Slope	-150.76	-0.16	0.07

Table 3.4: Estimated covariance matrix of the random effects for compound.

	Upper asym	IC50	Slope
Upper asym	1.00	-0.40	-0.34
IC50	-0.40	1.00	-0.22
Slope	-0.34	-0.22	1.00

Table 3.5: Estimated correlation matrix of the random effects for compound.

3.5.3 Estimation of **u** and β in a linear mixed effects model

Here we will elaborate on how **u** and $\boldsymbol{\beta}$ can be estimated directly in a linear mixed effects model. This will adapted for the nonlinear model later.

The objective is to determine the values of the conditional mode $\tilde{u}$ and conditional estimate $\tilde{\beta}$ that, given Σ, minimize the residual sum of squares:

$$||\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{u}, \Lambda)||^2 + ||\mathbf{u}||^2 \tag{3.42}$$

where

$$f(\boldsymbol{\beta}, \mathbf{u}, \Lambda) = X\boldsymbol{\beta} + Z\Lambda\mathbf{u} \tag{3.43}$$

We can write this equivalently as

$$\left|\left|\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ I_q & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix}\right|\right|^2 \tag{3.44}$$

where Z and X are design matrices, of dimension $n \times q$ and $n \times p$, for the random and fixed effects respectively.

In order to obtain the minimum, we differentiate the above sum of squares with respect to $\begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix}$.

Let's first expand the sum of squares

$$S = \left(\begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ I_q & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix}\right)^T \left(\begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} - \begin{bmatrix} Z\Lambda & X \\ I_q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix}\right) \tag{3.45}$$

$$= \mathbf{y}^T\mathbf{y} - \begin{bmatrix} \mathbf{y}^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} Z\Lambda & X \\ I_q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix} - \begin{bmatrix} \mathbf{u} & \boldsymbol{\beta} \end{bmatrix} \begin{bmatrix} \Lambda^T Z^T & I_q \\ X^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{0} \end{bmatrix} \tag{3.46}$$

$$+ \begin{bmatrix} \mathbf{u} & \boldsymbol{\beta} \end{bmatrix} \begin{bmatrix} \Lambda^T Z^T & I_q \\ X^T & \mathbf{0} \end{bmatrix} \begin{bmatrix} Z\Lambda & X \\ I_q & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \boldsymbol{\beta} \end{bmatrix}$$

2801728.01	-1832.57	-150.76	0.00	0.00
-1832.57	7.50	-0.16	0.00	0.00
-150.76	-0.16	0.07	0.00	0.00
0.00	0.00	0.00	0.13	0.00
0.00	0.00	0.00	0.00	2479593

Table 3.6: Covariance components of $\widehat{\sigma}^2\widehat{\Sigma}$ for random effects. The full matrix $\widehat{\Sigma}$ is obtained by expanding each of the elements in this matrix as a diagonal matrix whose size is the number of factor levels and dividing by $\widehat{\sigma}^2$.

The derivative is

$$\frac{dS}{d\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix}} = -2\begin{bmatrix}\Lambda^T Z^T & I_q\\ X^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{y}\\ \mathbf{0}\end{bmatrix} + 2\begin{bmatrix}\Lambda^T Z^T & I_q\\ X^T & \mathbf{0}\end{bmatrix}\begin{bmatrix}Z\Lambda & X\\ I_q & \mathbf{0}\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix} \tag{3.47}$$

$$= -2\begin{bmatrix}\Lambda^T Z^T\mathbf{y}\\ X^T\mathbf{y}\end{bmatrix} + 2\begin{bmatrix}\Lambda^T Z^T Z\Lambda + I_q & \Lambda^T Z^T X\\ X^T Z\Lambda & X^T X\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix} \tag{3.48}$$

Once the derivative is set to zero, the conditional mode, $\tilde{\mathbf{u}}$ and the conditional estimate, $\tilde{\boldsymbol{\beta}}$, can found as the solutions to the following sparse, positive-definite linear system.

$$\begin{bmatrix}\Lambda^T Z^T Z\Lambda + I_q & \Lambda^T Z^T X\\ X^T Z\Lambda & X^T X\end{bmatrix}\begin{bmatrix}\tilde{\mathbf{u}}\\ \tilde{\boldsymbol{\beta}}\end{bmatrix} = \begin{bmatrix}\Lambda^T Z^T\mathbf{y}\\ X^T\mathbf{y}\end{bmatrix} \tag{3.49}$$

We can then adapt this for use in the fitting of nonlinear models where Z and X are Jacobian matrices and we solve the system above to estimate the next step.

3.6 Penalized iteratively reweighted least squares

Conditional on the covariance of the random effects $\sigma^2\Sigma$, the estimates, $\begin{bmatrix}\tilde{\mathbf{u}}\\ \tilde{\boldsymbol{\beta}}\end{bmatrix}$, jointly minimize the residual sum of squares:

$$\begin{bmatrix}\tilde{\mathbf{u}}\\ \tilde{\boldsymbol{\beta}}\end{bmatrix} = \arg\min_{\mathbf{u},\boldsymbol{\beta}}[(\mathbf{y}-\boldsymbol{\mu})'W(\mathbf{y}-\boldsymbol{\mu}) + ||\mathbf{u}||^2] \tag{3.50}$$

where $\boldsymbol{\mu} = f\left(\begin{bmatrix}\mathbf{u}\\ \boldsymbol{\beta}\end{bmatrix}\right)$ and W is a $n \times n$ weight matrix to take care of the heteroscedasticity.

We can determine $\tilde{\mathbf{u}}$ and $\tilde{\boldsymbol{\beta}}$ through penalized, iteratively reweighted least squares. Let $\begin{bmatrix}\mathbf{u}^{(0)}\\ \boldsymbol{\beta}^{(0)}\end{bmatrix}$ be the vector of initial values.

The Jacobian matrices U (dimension: $n \times q$) and V (dimension: $n \times p$) are defined as:

$$U_{(i)} = W^{1/2} \left. \frac{df}{d\phi} \right|_{\phi_{(i)}} Z\Lambda \tag{3.51}$$

$$V_{(i)} = W^{1/2} \left. \frac{df}{d\phi} \right|_{\phi_{(i)}} X \tag{3.52}$$

(The matrix $\frac{df}{d\phi}$ is of dimension $n \times pn$ and looks like the horizontal concatenation of p diagonal $n \times n$ matrices.)
where they are used in the following system of equations to obtain the increments of the parameter vector at the ith iteration

$$\begin{bmatrix} U_{(i)}^T U_{(i)} + I_q & U_{(i)}^T V_{(i)} \\ V_{(i)}^T U_{(i)} & V_{(i)}^T V_{(i)} \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta}_{\boldsymbol{u}(i)} \\ \boldsymbol{\delta}_{\boldsymbol{\beta}(i)} \end{bmatrix} = \begin{bmatrix} U_{(i)}^T W^{1/2}(\mathbf{y} - \boldsymbol{\mu}_{(i)}) - \mathbf{u}_{(i)} \\ V_{(i)}^T W^{1/2}(\mathbf{y} - \boldsymbol{\mu}_{(i)}) \end{bmatrix} \tag{3.53}$$

where $\boldsymbol{\mu}_{(i)} = f\left(\begin{bmatrix} \mathbf{u}_{(i)} \\ \boldsymbol{\beta}_{(i)} \end{bmatrix} \right)$

The system of equations can be easily solved with the Cholesky decomposition. For a system $Ax = a$, if L is the Cholesky factor of A such that $LL^T = A$, then we can first solve for y in $Ly = a$, and then solve for x in $L^T x = y$.

The new set of values for u and β can then be calculated by the addition of the delta vector to the current set of values:

$$\begin{bmatrix} \mathbf{u}_{(i+1)} \\ \boldsymbol{\beta}_{(i+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{(i)} \\ \boldsymbol{\beta}_{(i)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\delta}_{\boldsymbol{u}(i)} \\ \boldsymbol{\delta}_{\boldsymbol{\beta}(i)} \end{bmatrix} \tag{3.54}$$

However, if this update increases the residual sum of squares, the move is not beneficial and cannot be accepted. So a simple change to the method involves decreasing a step factor (typically, halving it) until the residual sum of squares from the proposed values are lower than the current sum of squares.

$$\begin{bmatrix} \mathbf{u}_{(i+1)} \\ \boldsymbol{\beta}_{(i+1)} \end{bmatrix} = \begin{bmatrix} \mathbf{u}_{(i)} \\ \boldsymbol{\beta}_{(i)} \end{bmatrix} + \lambda \begin{bmatrix} \boldsymbol{\delta}_{\boldsymbol{u}(i)} \\ \boldsymbol{\delta}_{\boldsymbol{\beta}(i)} \end{bmatrix} \tag{3.55}$$

Start with $\lambda = 1$ and halve it until

$$(\mathbf{y} - \boldsymbol{\mu}_{(i+1)})^T (\mathbf{y} - \boldsymbol{\mu}_{(i+1)}) + ||\mathbf{u}_{(i+1)}||^2 < (\mathbf{y} - \boldsymbol{\mu}_{(i)})^T W (\mathbf{y} - \boldsymbol{\mu}_{(i)}) + ||\mathbf{u}_{(i)}||^2 \tag{3.56}$$

to obtain the new set of values to be used in the next iteration until convergence.

3.6.1 Initial values

There are quite a few ways of obtaining reasonable initial values. Since the linear mixed effects models for estimating the covariance components resulted in some reasonable values for the fixed and random effects, those can be used as the initial values.

3.6.2 Convergence: Relative offset criterion

We can determine the convergence of the iterations with the orthogonality convergence criterion. In the case of a fixed effects model, convergence is reached if the following is less than a suitably chosen value:

$$\frac{||Q_1^T(y - f(\beta))||/\sqrt{p}}{||Q_2^T(y - f(\beta))||/\sqrt{n-p}} \tag{3.57}$$

where Q_1^T and Q_2^T are the first p and last $n-p$ columns respectively, of the Q matrix from the QR decomposition of the Jacobian matrix V.

Work is needed to determine the comparable criterion for the mixed effects case; we can not use 3.57 directly since random effects are not considered parameters but random variables in the model. For now, we iterate our procedure until the sum of squares is extremely small.

3.6.3 Using optim()

This method involves minimizing the sum of squares, given Λ,

$$||\mathbf{y} - f(\boldsymbol{\beta}, \mathbf{u}, \Lambda)||^2 + ||\mathbf{u}||^2 \tag{3.58}$$

directly with numerical optimization methods such as those available in the optimization function in R called optim().

3.7 Simulation testing

The task of fitting nonlinear models is generally difficult. Success or what can be considered success depends on so many factors, including whether we have the correct model, the choice of initial values, the quality of the data, and insufficient data.

So it is not surprising that the few attempts at using the nlmer() function in the lme4 package failed. It is not a failure of the author, but a failure on the user's part. We need to truly understand our data, how this specific model should be fit, and what we can do to use this methodology practically.

In this section, we fit a simpler nonlinear model on simulated data to see how well we can estimate true effects. We simulate a dataset of 25 chips where the same 6

upper asymptote	β_1	1000.00
IC50	β_2	6.85
slope	β_3	-0.40
upper asymptote, compound	$\sigma_{a,1}$	100
upper asymptote, chip	$\sigma_{c,1}$	300
IC50, compound	$\sigma_{a,2}$	0.5
error	σ	100

Table 3.7: Values for simulation.

compounds occur on each chip (Figure 3.28). The parameters are set to values that are roughly in the range of what the values in the dataset (Table 3.7).

The model in this case assumes that the upper asymptotes vary by compound and chip, while the IC50 varies by compound. The slope remains constant.

$$\begin{aligned} y_{ijk}(r,s) &= f(x(r,s);\phi_{\mathbf{ijk}}) + \epsilon_{ijk}(r,s) \qquad (3.59)\\ &= \frac{\phi_{1,ik}}{1+\exp\left(-\frac{(x(r,s)-\phi_{2,i})}{\phi_3}\right)} + \epsilon_{ijk}(r,s) \qquad (3.60)\end{aligned}$$

$$\phi_{ijk} = \begin{bmatrix} \phi_{1,ik} \\ \phi_{2,i} \\ \phi_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} a_{1,i} \\ a_{2,i} \\ 0 \end{bmatrix} + \begin{bmatrix} c_{1,k} \\ 0 \\ 0 \end{bmatrix} \qquad (3.61)$$

$$(3.62)$$

where $A_i = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$, and $C_k = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

$$\mathbf{a_i} = \begin{bmatrix} a_{1,i} \\ a_{2,i} \\ a_{3,i} \end{bmatrix} \qquad \mathbf{c_k} = \begin{bmatrix} c_{1,k} \\ c_{2,k} \\ c_{3,k} \end{bmatrix} \qquad (3.63)$$

$$a_i \sim \mathcal{N}(0,\sigma_a^2) \qquad c_k \sim \mathcal{N}(0,\sigma_c^2) \qquad (3.64)$$

$$\epsilon_{ijk}(r,s) \sim \mathcal{N}(0,\sigma^2) \qquad (3.65)$$

We obtain estimates for 574 of the 600 units through individual nonlinear least squares, where they are used to provide estimates of the covariance components.

Estimation of covariance components

To estimate and predict the fixed and random effects respectively, we first need to estimate the covariance components required for the random effects in this model.

Because the compound factor is involved in the random effects of two of the parameters, upper asymptote and IC50, there is a 2×2 covariance matrix to be estimated. However the random effects of upper asymptote are separated into two groups, one for the compound factor and the other for the chip factor. What we can do is to separate the effects with a linear mixed effects model

$$\widehat{\phi}_{1ik,nls} = \beta_1 + a_i + c_k + \epsilon_{ik} \tag{3.66}$$

where the data is the estimates from the individual nonlinear least squares to be decomposed into a mean value (β_1) and random effects by compound (a_i) and by chip (c_k). The estimated covariance matrix of the compound effects by upper asymptote and compound is the robust covariance of the $\widehat{a_i}$ and the medians of the IC50 estimates by compound (see Table 3.8 for these values and Table 3.9 for the estimated covariance and correlation matrix).

Compound	Upper asym effects	IC50 effects
A	-8.98	0.08
B	-17.74	0.26
C	109.98	-0.50
D	-41.33	-0.48
E	-126.62	-0.00
F	84.69	0.04

Table 3.8: Table of compound effects by upper asymptote and IC50. The first one was obtained through a linear mixed effects model on individual nonlinear estimates of the upper asymptote. The second was medians of the IC50 estimates by compound factor.

The chip factor is only involved in the upper asymptote parameter; we can easily estimate the variance of the random effects with the estimated variance from the linear mixed effects model 3.66, $\widehat{\sigma}^2_{1,c} = 110199.25$, ($\widehat{\sigma}_{1,c} = 331.96$).

And lastly, $\widehat{\sigma}$ is easily obtained as a summary of the estimated residual errors from

	Covariance		Correlation	
upper asym	9187.52	6.59	1.00	0.20
IC50	6.59	0.12	0.20	1.00

Table 3.9: Estimated covariance and correlation matrix for compound factor in simulation testing.

the individual nonlinear least squares fittings.

$$\widehat{\sigma} = \sqrt{\frac{1}{n_{\#}} \sum_{i_{\#}=1}^{n_{\#}} \widehat{\sigma}^2_{i_{\#},nls}} = 100.13 \tag{3.67}$$

where $i_{\#}$ indexes the $n_{\#}$ units that were able to be fitted with nonlinear least squares.

Thus the estimated relative covariance matrix of the random effects (3.29) is

$$\widehat{\Sigma} = \begin{bmatrix} \frac{1}{\widehat{\sigma}^2} \begin{bmatrix} \widehat{\sigma}^2_{a,1} & \widehat{\sigma}_{a,12} \\ \widehat{\sigma}_{a,12} & \widehat{\sigma}^2_{a,2} \end{bmatrix} \otimes I_{n_{comp}} & \mathbf{0} \\ \mathbf{0} & \frac{\widehat{\sigma}^2_{c,1}}{\widehat{\sigma}^2} \otimes I_{n_{chip}} \end{bmatrix} \tag{3.68}$$

Results

The model is fit with penalized least squares where we obtain final results very close to the true values. While the overall estimated linear combinations of the fixed and random effects match the true values well (Figures 3.31 and 3.32), the estimated fixed and predicted random effects separately are shifted from the true effects (Figure 3.30). This is due to the fact that the fixed effects are estimating the empirical means of the parameters, and it is unlikely that the simulated data that is generated will have values that are equal or even extremely close to the true values, thus the estimated fixed effects will be some distance off from the true values and the predicted random effects will compensate for this distance in the final estimates. The good results for this simulation is dependent on a balanced dataset as well as simulating data where over 95% of the units are successfully estimated with individually with just nonlinear least squares.

	Fitted estimates	True values	Initial values
$\widehat{\beta_1}$	894.77	1000.00	984.56
$\widehat{\beta_2}$	6.94	6.85	6.94
$\widehat{\beta_3}$	-0.40	-0.40	-0.40

Table 3.10: Fixed effects estimates for simulated data.

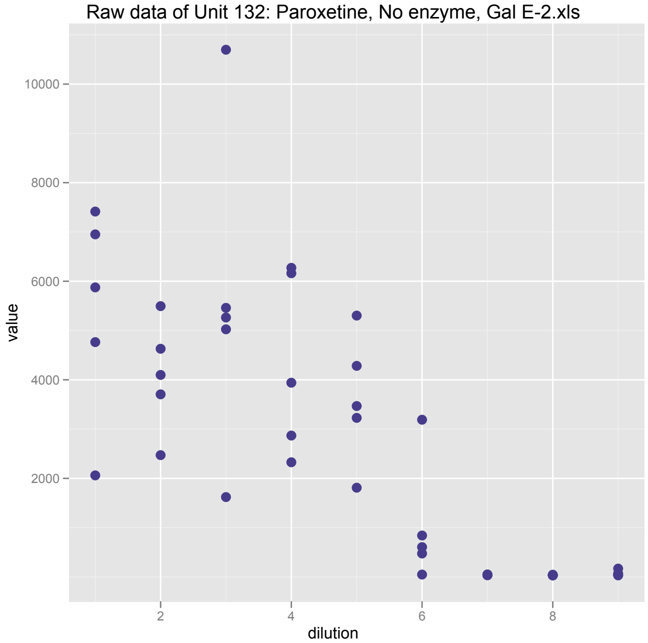


Figure 3.14: Plot of raw data from Unit 132.

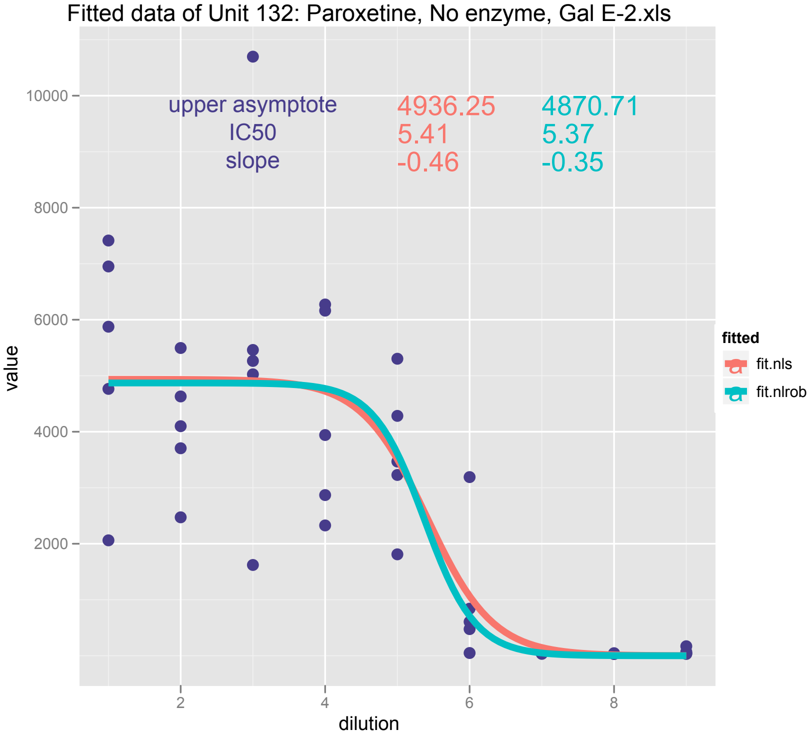


Figure 3.15: Plot of fitted curves, ordinary and robust nonlinear least squares, of Unit 132.

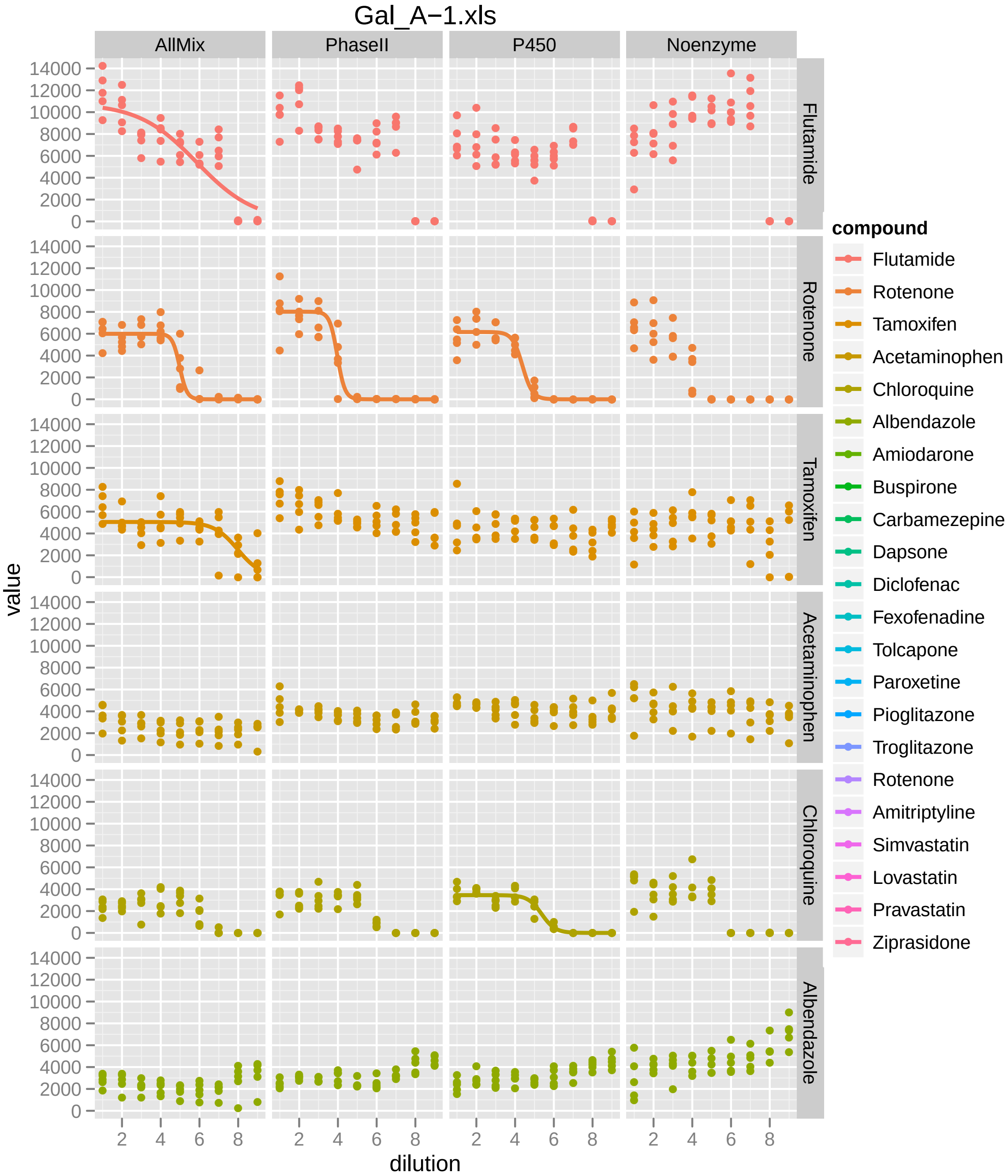


Figure 3.16: Plots of attempted robust nonlinear least squares fitting of chip Gal A-1.

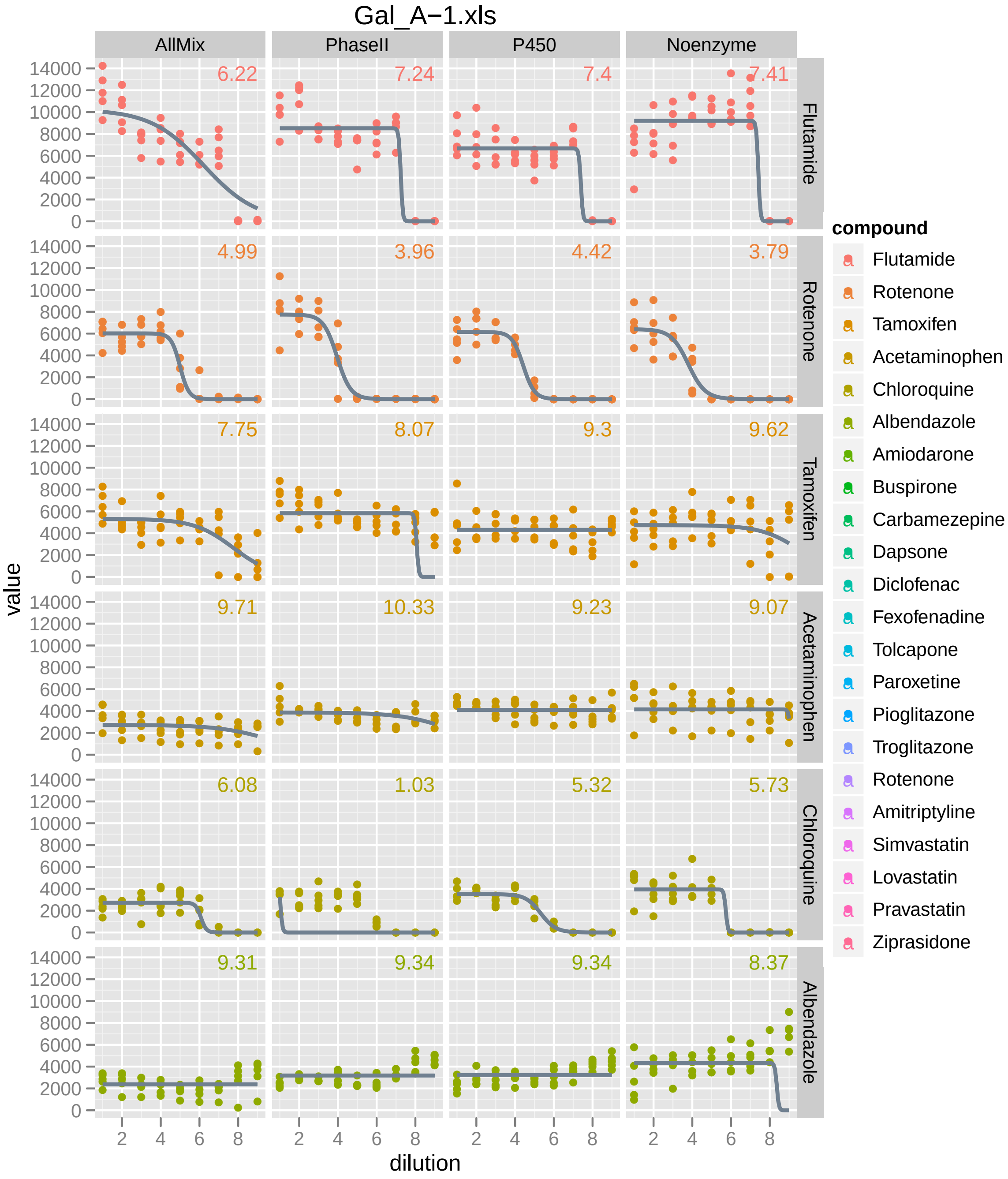


Figure 3.17: Fitted curves from naive two stage method.

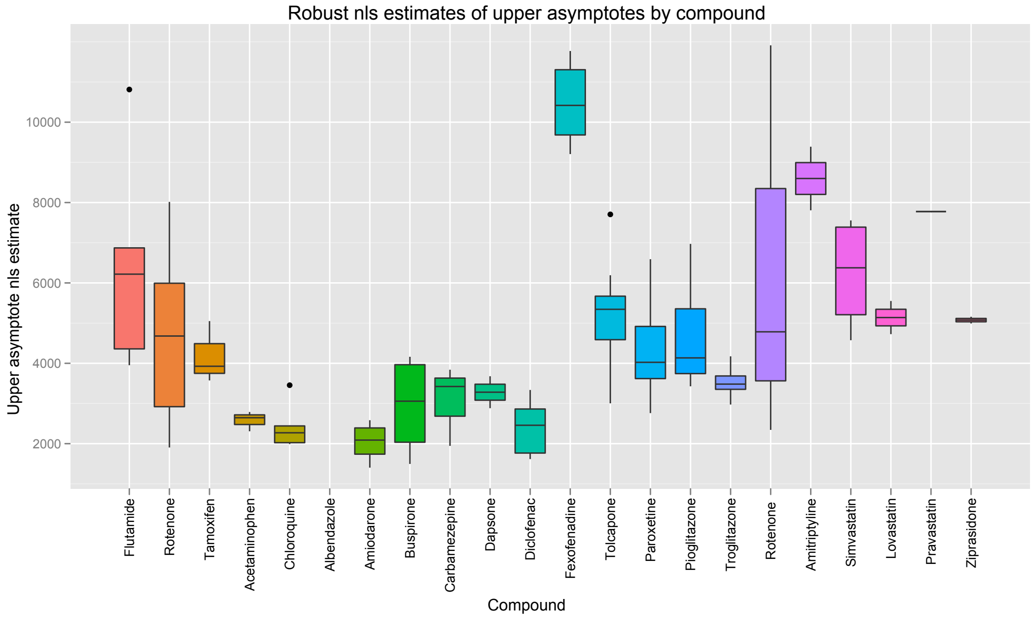


Figure 3.18: Boxplots of the individual nonlinear least squares estimates of the upper asymptotes by compound.

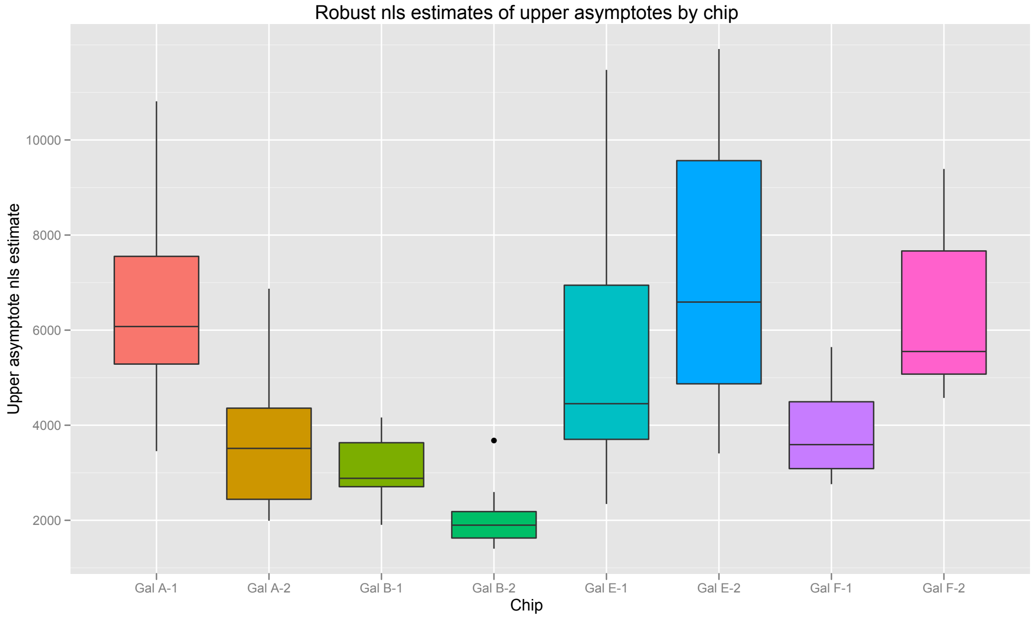


Figure 3.19: Boxplots of the individual nonlinear least squares estimates of the upper asymptotes by chip.

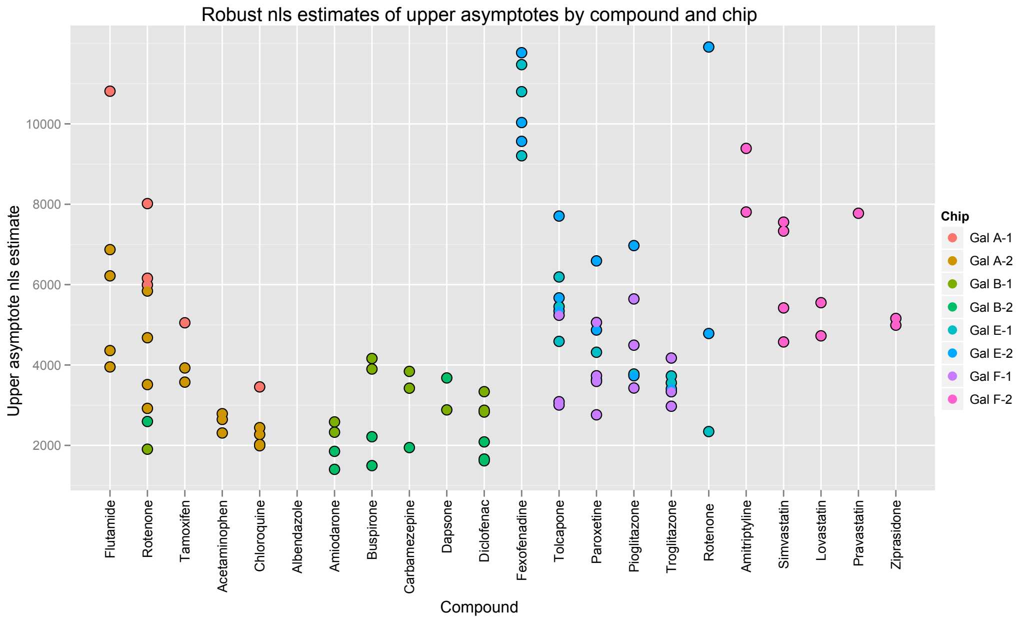


Figure 3.20: Plot of the individual nonlinear least squares estimates of the upper asymptotes by compound and chip.

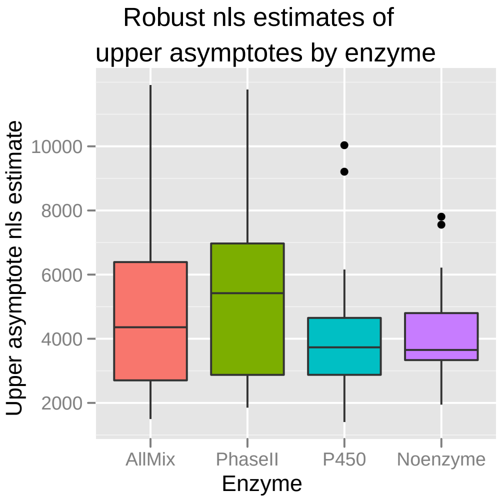


Figure 3.21: Boxplots of the individual nonlinear least squares estimates of the upper asymptotes by enzyme.

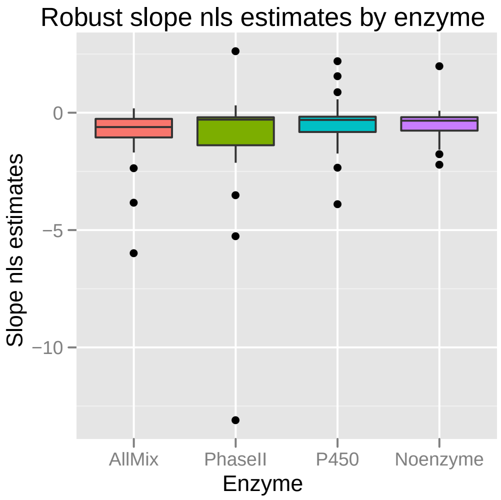


Figure 3.22: Plot of the individual nonlinear least squares estimates of the slope by enzyme.

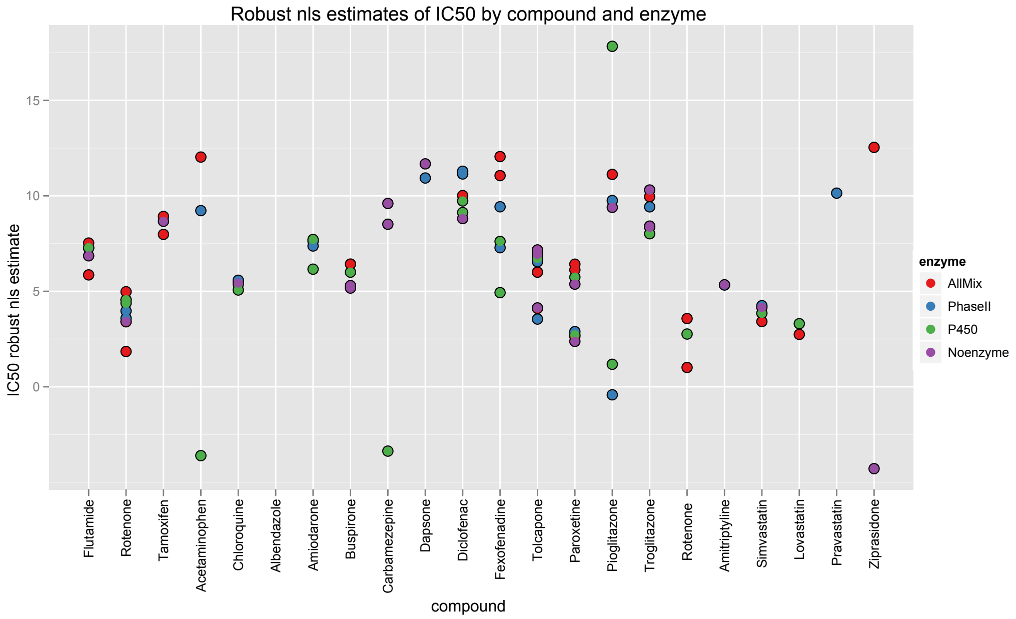


Figure 3.23: Plot of the individual nonlinear least squares estimates of the upper asymptotes by compound and enzyme.

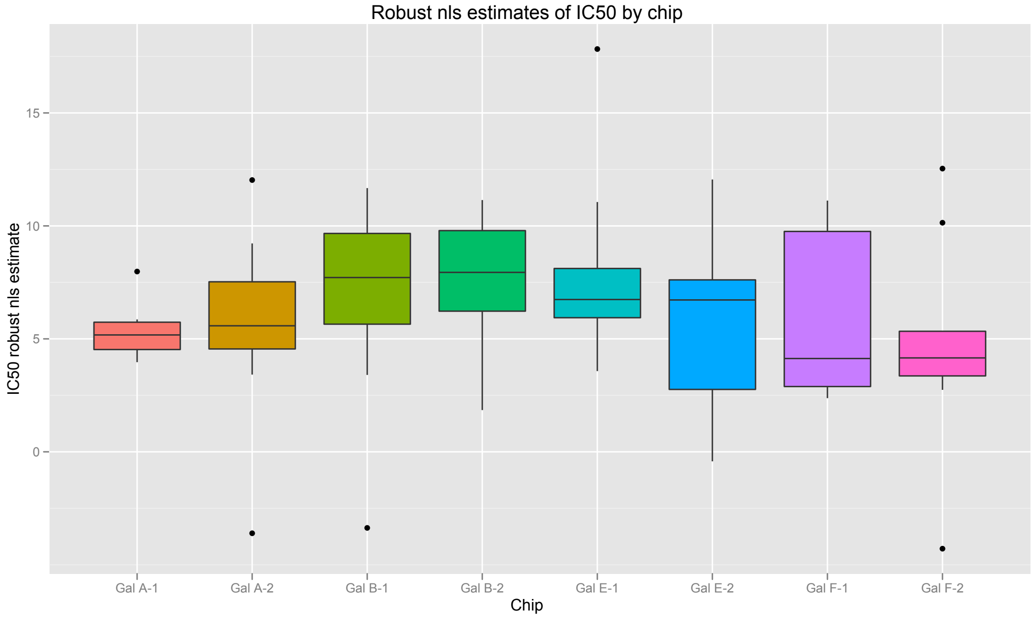


Figure 3.24: Plot of the individual nonlinear least squares estimates of the IC50 by chip.

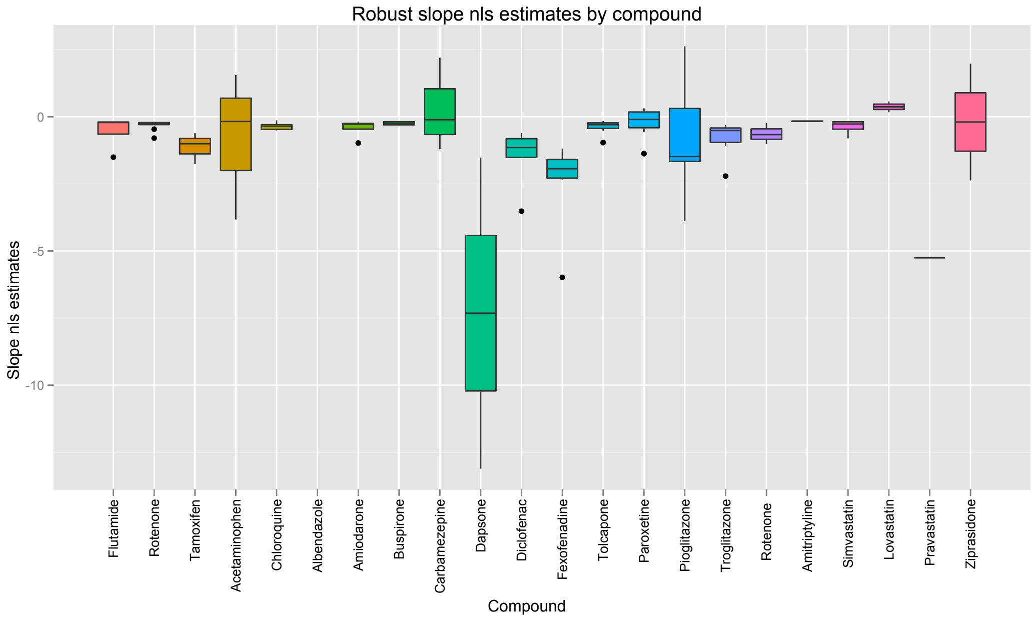


Figure 3.25: Plot of the individual nonlinear least squares estimates of the slope by compound.

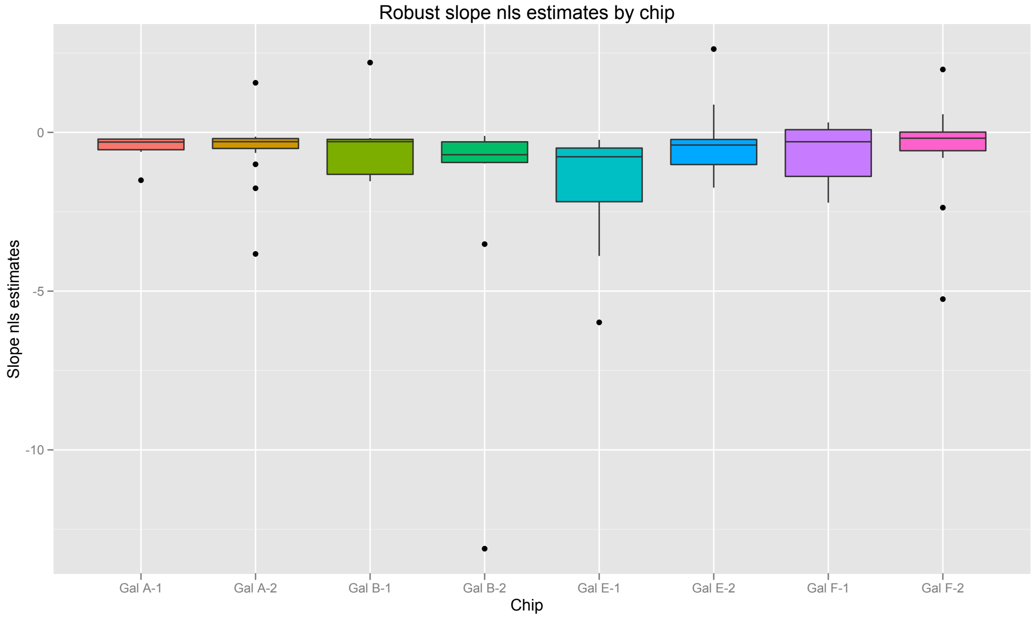


Figure 3.26: Plot of the individual nonlinear least squares estimates of the slope by chip.

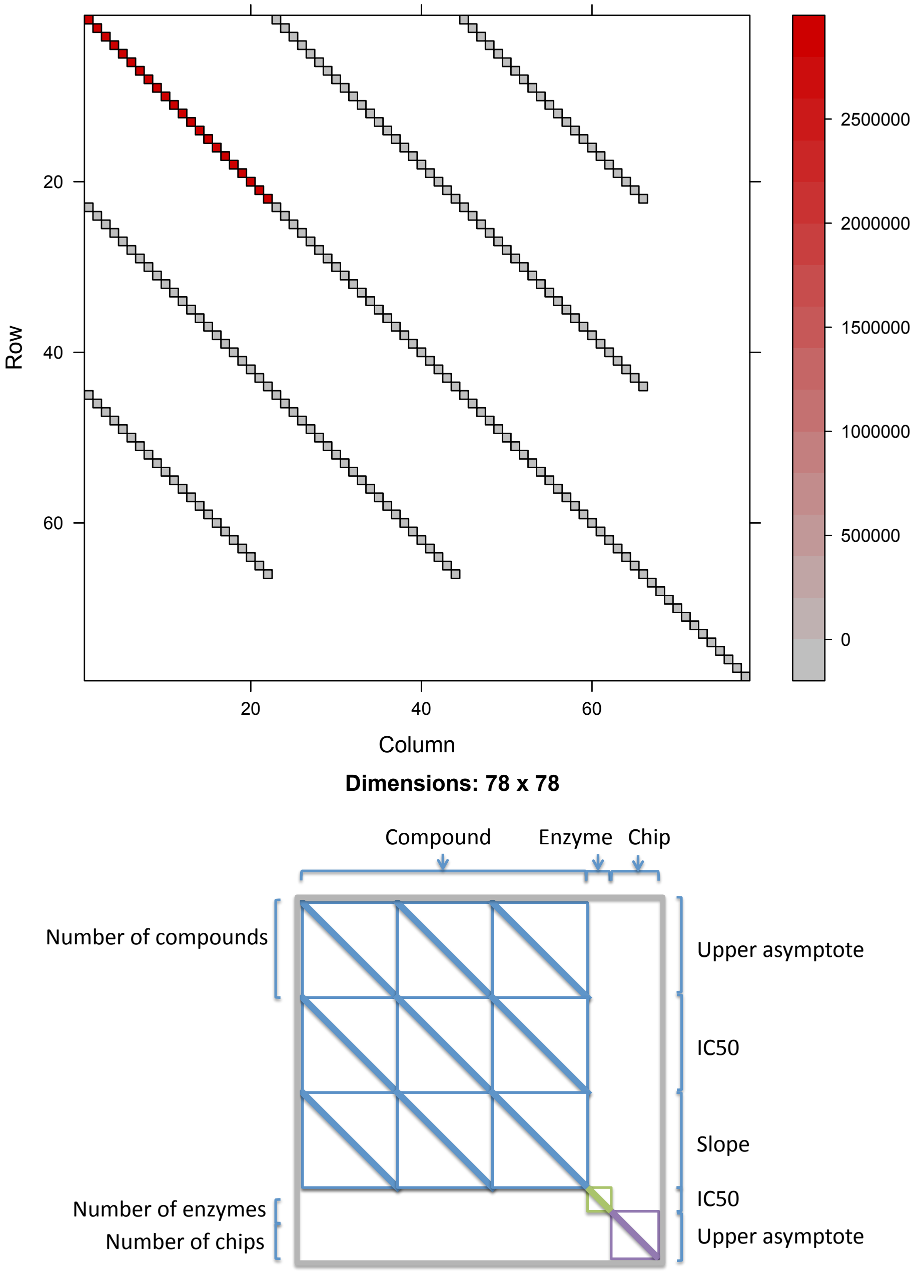


Figure 3.27: Structure of covariance matrix of random effects.

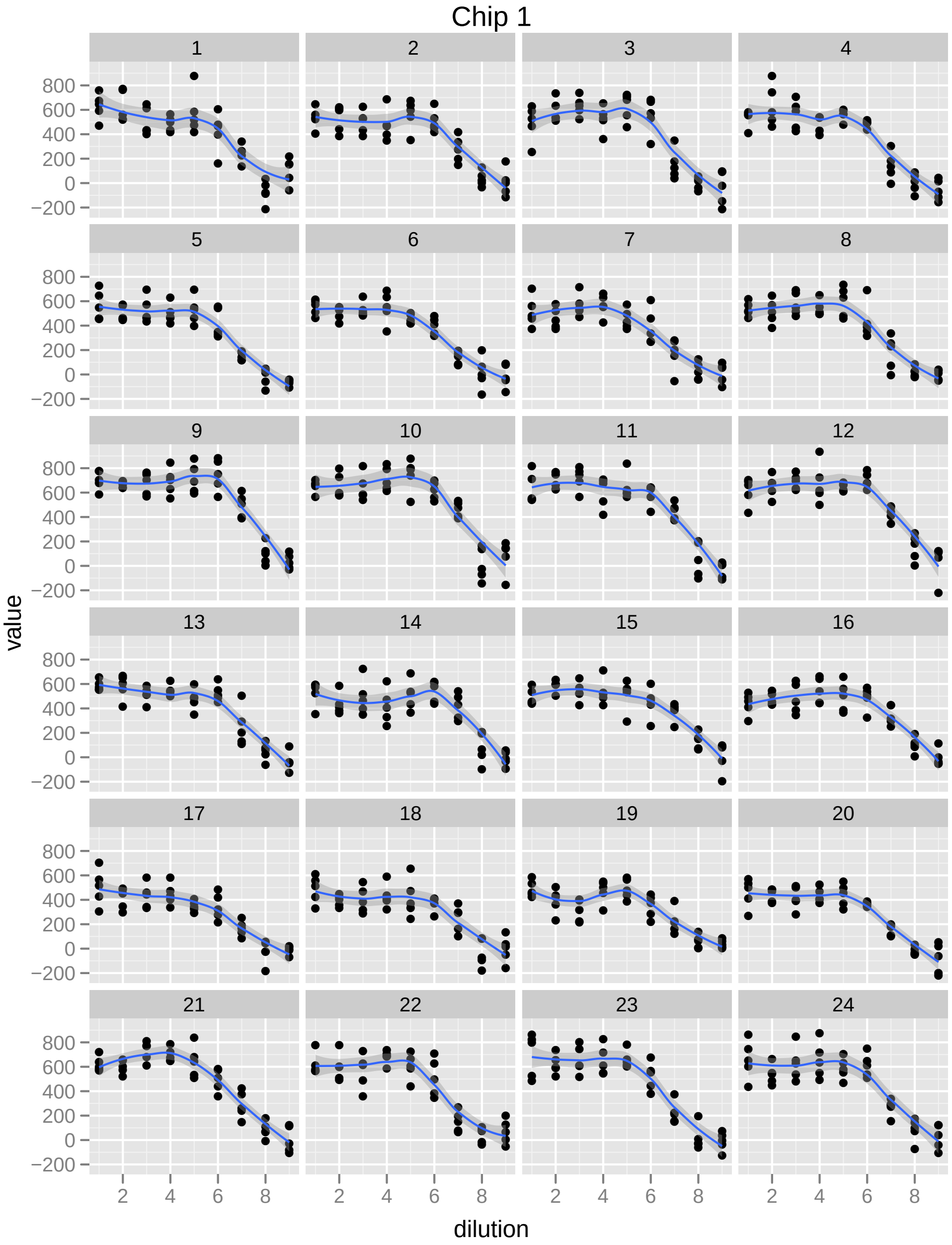


Figure 3.28: Plots of raw data from the first chip in the simulated dataset. Smoothed lines have been added to visualize the trend.

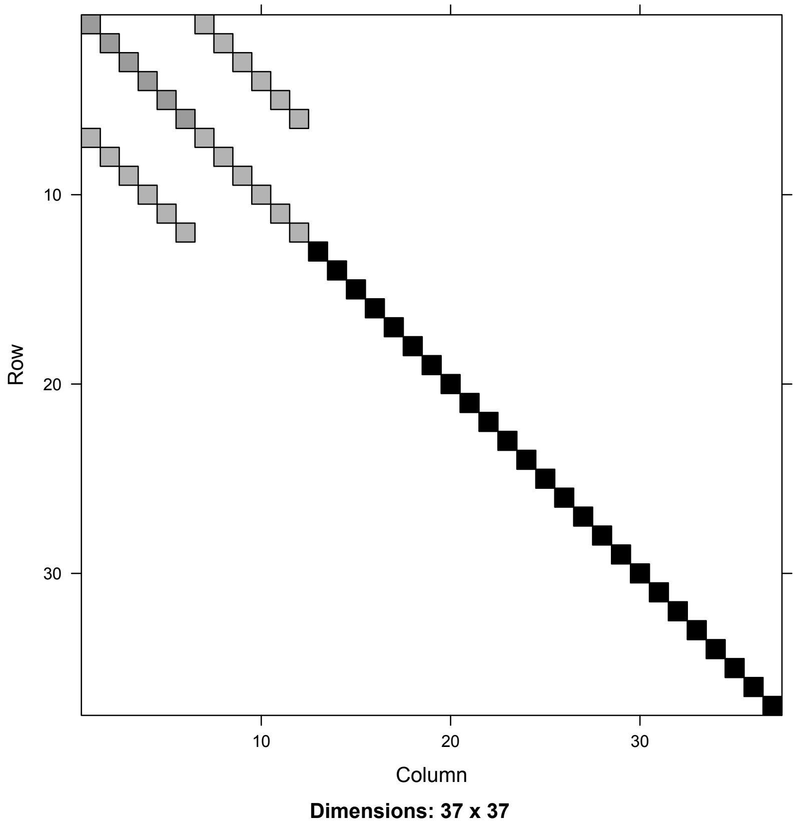


Figure 3.29: Structure of the covariance matrix for simulated data.

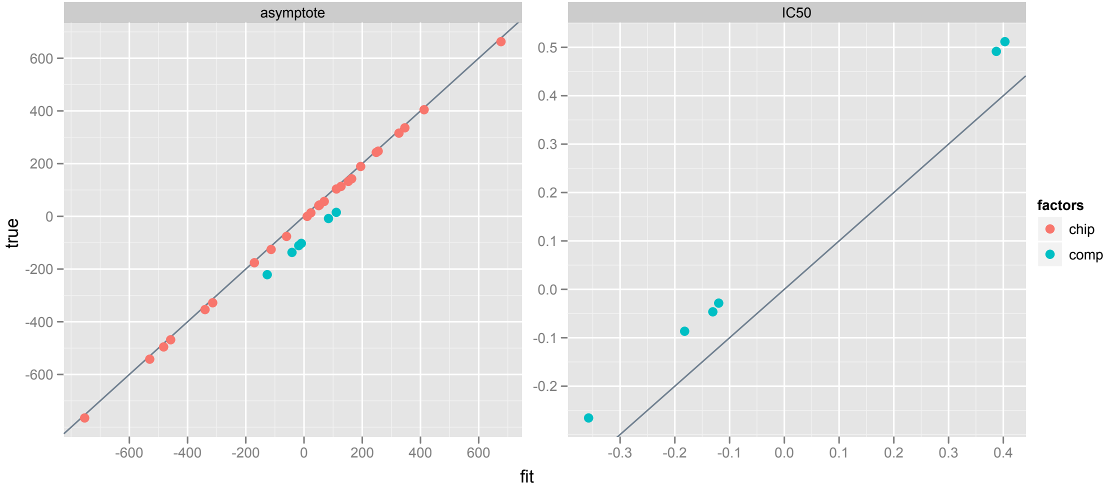


Figure 3.30: Plots of the predicted random effects versus their true values for the compound and chip factors. Lines with intercept 0 and slope 1 are provided for reference.

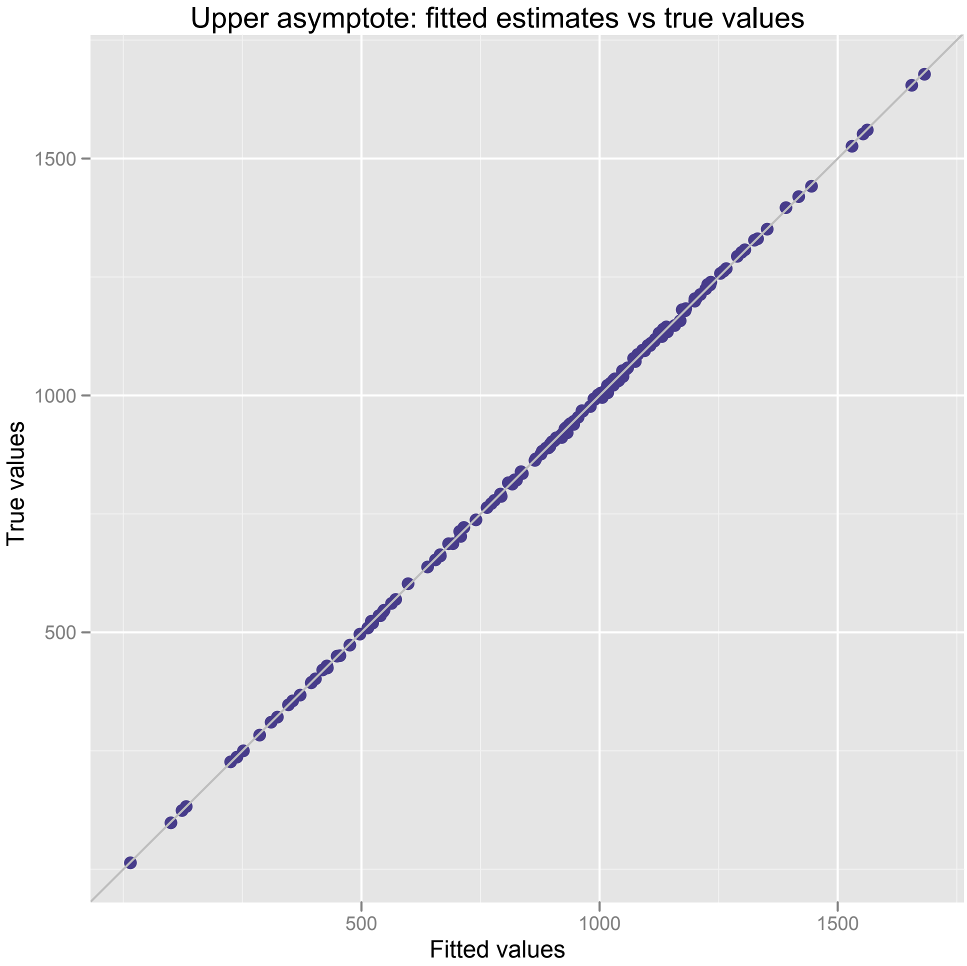


Figure 3.31: Plot of the estimated versus true values of the upper asymptotes. Lines with intercept 0 and slope 1 are provided for reference.

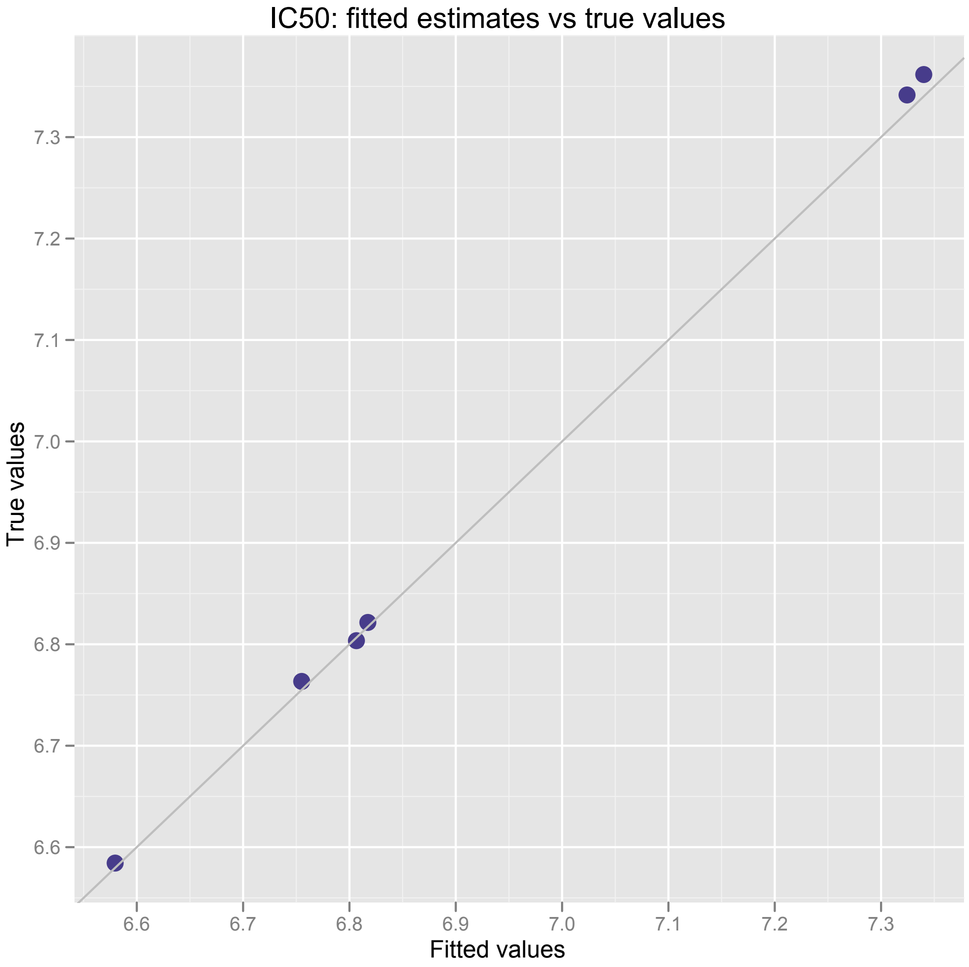


Figure 3.32: Plot of the estimated versus true values of the IC50. Lines with intercept 0 and slope 1 are provided for reference.

3.8 Fitting of Model 1 on current working dataset

Initial values generated had a sum of squares at 42,265,359,093, and this desired model with this set of initial values got stuck at 29,772,027,791, and could not converge. As a reference, the lowest sum of squares that we can obtain by fitting each dilution at the median of values is 11,655,071,270. This could be due to many factors such as the choice of initial values, the model, or the quality of the data. For now, we proceed with simplifying the model to see what happens.

3.9 Model 2: Partially crossed effects with constant slope

As we can see, just because one model seems ideal for the data setup we have does not mean that we will have any success with it. Fitting nonlinear models in general is a complicated and messy business, and if we add on top of that, a reasonably large set of data that is also not very nice, we are left with many dead ends and seemingly promising solutions that do not pan out. Any reasonable success most likely requires much compromise. In this case, we have to resort to dealing with a simpler model. We notice that many of the underlying problems result from the lab not being able determine the best range of concentrations before performing the experiment. We see a lot of wasted dilutions that overcover one or both of the asymptotes while leaving a huge gap in where the most interesting section is, the linear phase where we determine the IC50.

As such, this means that we are missing much needed information for estimating the two most important parameters in the logistic curve, the IC50 and the slope. Thus, we can see how these two are linked strongly and that such problems lead to identifiability issues; in such cases, we are clearly unable to obtain unique values for either of them.

Even though we acknowledge that being able to estimate the slopes appropriately makes much sense biologically, we have to focus on the fact that in this case the key parameter here is the IC50 and in order to help estimate it, we will have to place certain constraints and assumptions on the slope parameter. We note that the asymptote parameter does not have much influence on either the IC50 or slope.

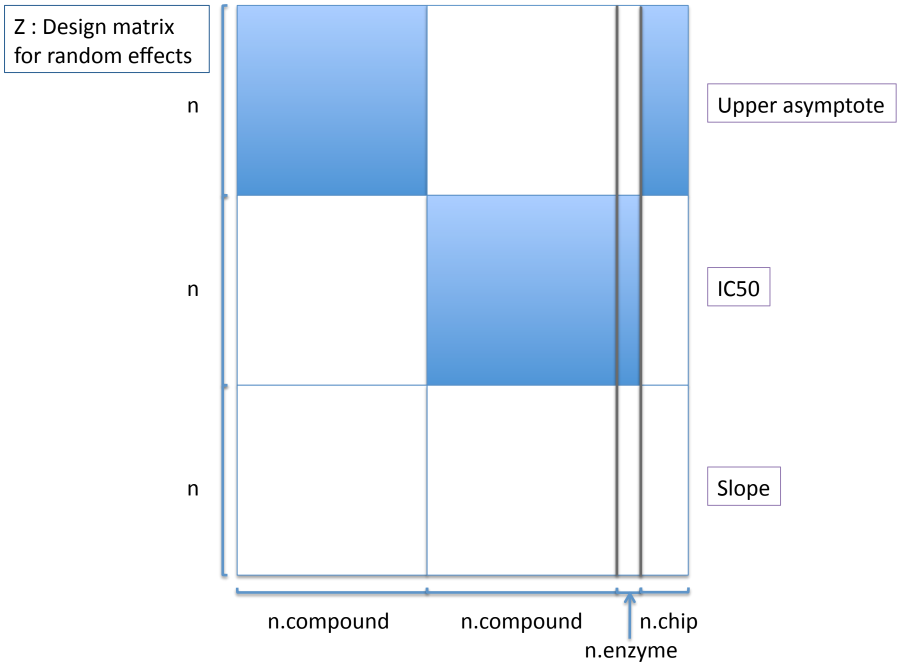


Figure 3.33: Diagram of the structure of the design matrix Z for random effects. The dimensions of the Z matrix is $pn \times q$, where p is the number of fixed effects, n is the total number of observations, and q is the number of random effects. The blue shaded boxes refer to submatrices where the ith row consists of the vector e_j, such the jth entry is 1 if the ith observation corresponds to the jth level of the specific factor and 0 otherwise. We can detect from the structure above that The example above shows that the corresponding model does not include random effects for the slope parameter, although we need to include the block of matrices for it, where all entries are zero.

3.9.1 Results

After around 190 iterations, we obtain sum of squares equal to 23,795,437,805. Visual check of all the fitted curves (also included in the appendix) show that in general, the model is fitting the data reasonably well. We check carefully to see if the constant slope is working well. Most of the units do not have sufficient data near the area where the slope parameter is expected to be estimated but if they do, there does not seem to be too much discrepancy between the fitted curve and the observations there.

Problem units

The set of units for the compound Paroxetine with the four enzyme conditions across three different chips clearly have fitting issues (Figure 3.34). While the first two chips Gal E-1 and Gal E-2 are similar, the third chip Gal F-1 has what looks like the edge bleeding problem, where the first two dilutions of each unit are at near zero intensity even though the rest have strong signals. These strong signals, however, don't coincide with the descent in the data in other units. The units on chip Gal F-1 are clearly outliers.

unit	compound	enzyme	chip	upper asymptote	IC50	slope
105	Paroxetine	AllMix	Gal_E-1.xls	3853.38	6.45	-0.33
106	Paroxetine	PhaseII	Gal_E-1.xls	3853.38	6.41	-0.33
107	Paroxetine	P450	Gal_E-1.xls	3853.38	6.17	-0.33
108	Paroxetine	Noenzyme	Gal_E-1.xls	3853.38	6.31	-0.33
129	Paroxetine	AllMix	Gal_E-2.xls	5859.66	6.45	-0.33
130	Paroxetine	PhaseII	Gal_E-2.xls	5859.66	6.41	-0.33
131	Paroxetine	P450	Gal_E-2.xls	5859.66	6.17	-0.33
132	Paroxetine	Noenzyme	Gal_E-2.xls	5859.66	6.31	-0.33
153	Paroxetine	AllMix	Gal_F-1.xls	3566.91	6.45	-0.33
154	Paroxetine	PhaseII	Gal_F-1.xls	3566.91	6.41	-0.33
155	Paroxetine	P450	Gal_F-1.xls	3566.91	6.17	-0.33
156	Paroxetine	Noenzyme	Gal_F-1.xls	3566.91	6.31	-0.33

Table 3.11: Table of fitted results for the compound Paroxetine for all enzyme levels across three chips. See corresponding plots to see that the Paroxetine data on the third file Gal_F-1.xls is very different than on the other two files.

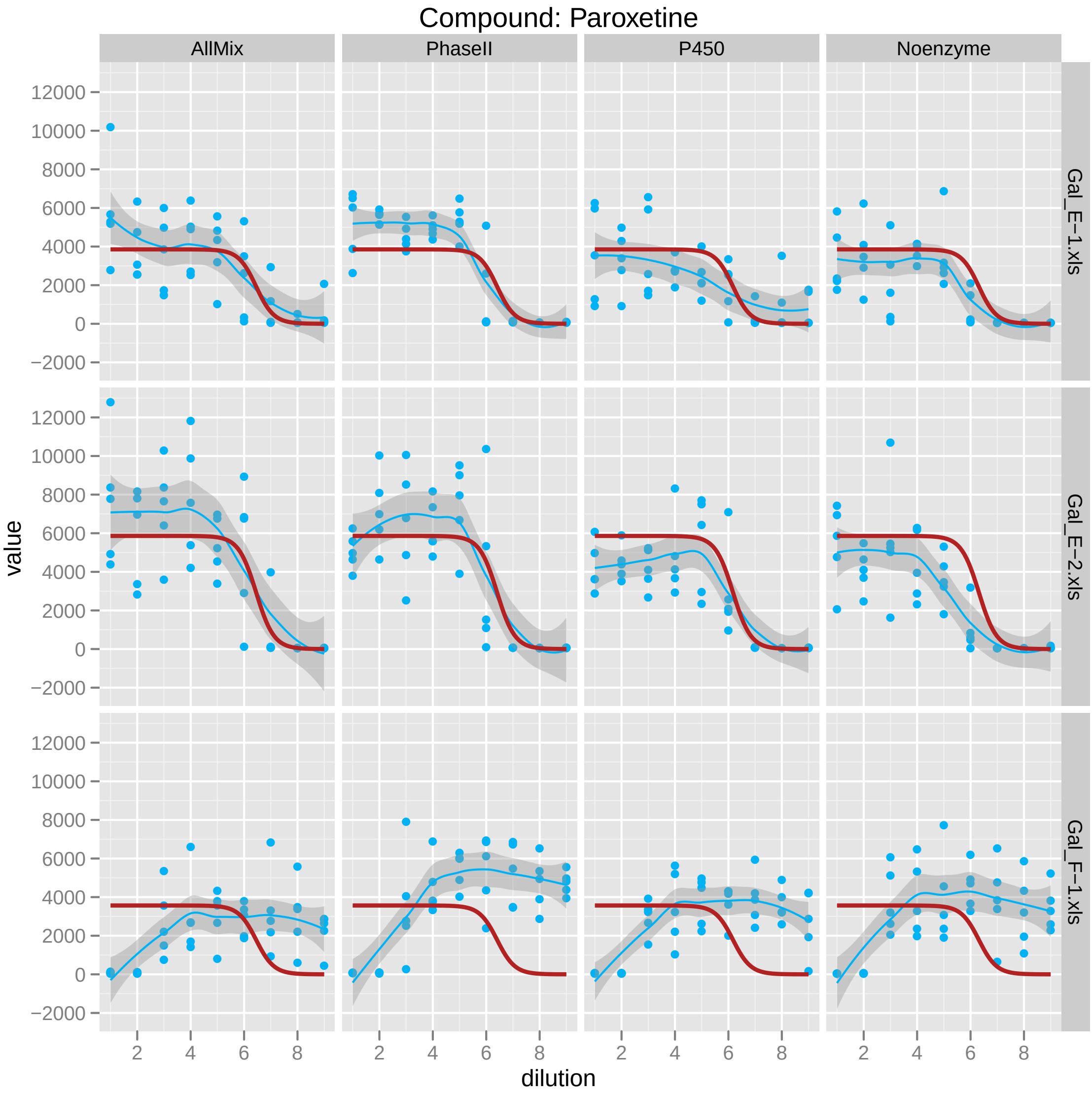


Figure 3.34: Plots of the fitted curves for the compound Paroxetine across 3 chips.

Figure 3.35 shows another example of units that clearly differ from the majority. It may seem to be an issue of different starting concentrations within this compound Tolcapone, but all the starting concentrations are found to be the same.

We will take care of the outliers in a little bit. We notice that the other fitted curves do not estimate the IC50 where it is visually expected on the data points. Because of this, we question the assumption that the levels of the enzyme effects are same across all compounds.

unit	compound	enzyme	chip	upper asymptote	IC50	slope
101	Tolcapone	AllMix	Gal_E-1.xls	3907.89	6.92	-0.33
102	Tolcapone	PhaseII	Gal_E-1.xls	3907.89	6.87	-0.33
103	Tolcapone	P450	Gal_E-1.xls	3907.89	6.64	-0.33
104	Tolcapone	Noenzyme	Gal_E-1.xls	3907.89	6.77	-0.33
125	Tolcapone	AllMix	Gal_E-2.xls	5914.16	6.92	-0.33
126	Tolcapone	PhaseII	Gal_E-2.xls	5914.16	6.87	-0.33
127	Tolcapone	P450	Gal_E-2.xls	5914.16	6.64	-0.33
128	Tolcapone	Noenzyme	Gal_E-2.xls	5914.16	6.77	-0.33
149	Tolcapone	AllMix	Gal_F-1.xls	3621.41	6.92	-0.33
150	Tolcapone	PhaseII	Gal_F-1.xls	3621.41	6.87	-0.33
151	Tolcapone	P450	Gal_F-1.xls	3621.41	6.64	-0.33
152	Tolcapone	Noenzyme	Gal_F-1.xls	3621.41	6.77	-0.33

Table 3.12: Table of fitted results for the compound Tolcapone for all enzyme levels across three chips. See corresponding plots to see that the Tolcapone data on the third file Gal F_1.xls is very different than on the other two files. All the units for this compound on that file seem to have been shifted on x-axis. This leads to a conclusion that perhaps the range of concentrations for this compound across different chips is different; however, a quick check of the data defeats this hypothesis. So we are unclear as to why the data is so.

3.10 Model 3: Nested and partially crossed random effects

We can see from the results of the crossed effects model that we are not getting the best IC50 estimates. The IC50 values for the No Enzyme column seem to be overestimated. This leads us to believe that while there is likely an enzyme effect on the value, we should not be assuming that these effects are the same for every compound. Instead, we should allow more flexibility in the enzyme random effects but not to the extreme of an effect for every single unit of data. We now suggest a model which incorporates enzyme effects that are nested in compound effects so that we can determine each compound / enzyme relationship individually.

The methodology framework discussed earlier can accommodate both nested effects and crossed effects quite easily. Basically, the only difference in the model setup is to specify the factor levels of the enzymes as distinct across all chips. The original data setup have the same four labels for every compound. This is how crossed effects are able to be estimated since these labels connect similar enzyme conditions across different compound and chips. However, we now do not want this connection; we want to be able to estimate four enzyme effects independently for each compound.

This is easily implemented by concatenating the compound level to the enzyme

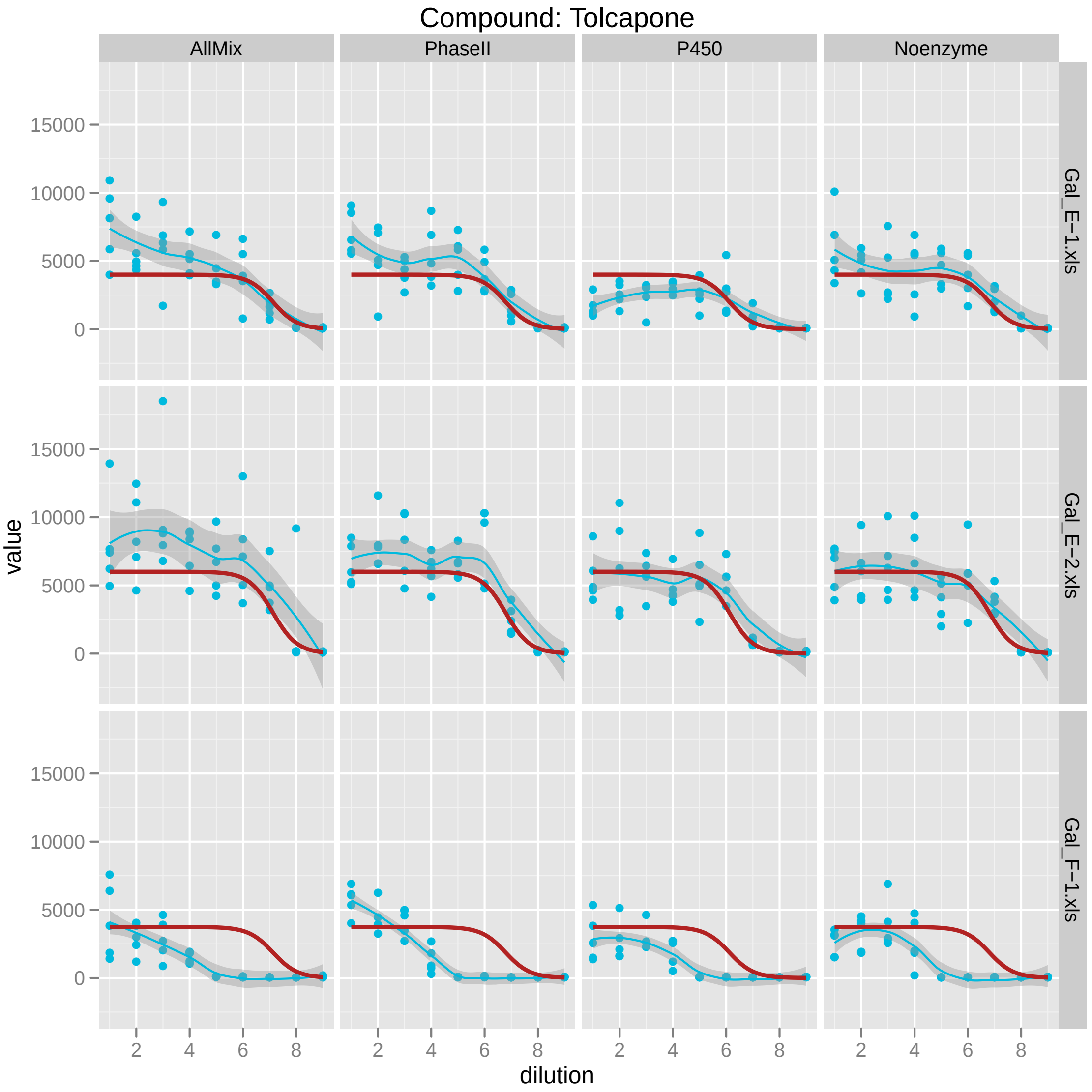


Figure 3.35: Plots of the fitted curves for the compound Tolcapone across 3 chips.

label to generate an unique enzyme condition that is specific to compounds. Any replicates of compound / enzyme combinations will have the same estimate. The only issue is that this will increase the number of random effects to be predicted by the number of enzymes multiplied by the number of compounds subtracted by one. For the current dataset, this is an additional $(4 * (22 - 1) = 84)$ effects.

Estimation of covariance components

Because of the nesting, we increase the total number of random effects to 140, thus the covariance matrix is also much larger than before (Figure 3.36). There is insufficent data to estimate a variance of the random effects for the enzyme factor within each compound; even then such estimates are highly variable. Instead we set the standard deviation of the nested enzyme effects to be the median of the the median absolute deviations of the individual robust estimates scaled by 1.4826.

	IC50 MAD * 1.4826
Flutamide	0.57
Rotenone	1.24
Tamoxifen	0.56
Acetaminophen	6.17
Chloroquine	0.18
Albendazole	NA
Amiodarone	0.37
Buspirone	0.91
Carbamezepine	2.39
Dapsone	0.81
Diclofenac	1.99
Fexofenadine	4.15
Tolcapone	1.33
Paroxetine	3.41
Pioglitazone	10.77
Troglitazone	1.61
Rotenone	1.78
Amitriptyline	0.01
Simvastatin	0.44
Lovastatin	0.62
Pravastatin	0.00
Ziprasidone	18.49

Table 3.13: Table of the median absolute deviations of the individual robust nonlinear least squares estimates of IC50 by compound scaled by 1.4826. We take the median of these numbers, 1.24, to be the standard deviation of the nested enzyme factor.

Results

This modified model is fit in the same way as the previous model. While the previous model resulted in the sum of squares of 23,795,437,805, this partially crossed and nested model is able to fit the data a bit more closely, because of the flexibility in the enzyme factor, with the sum of squares at 23,015,615,640.

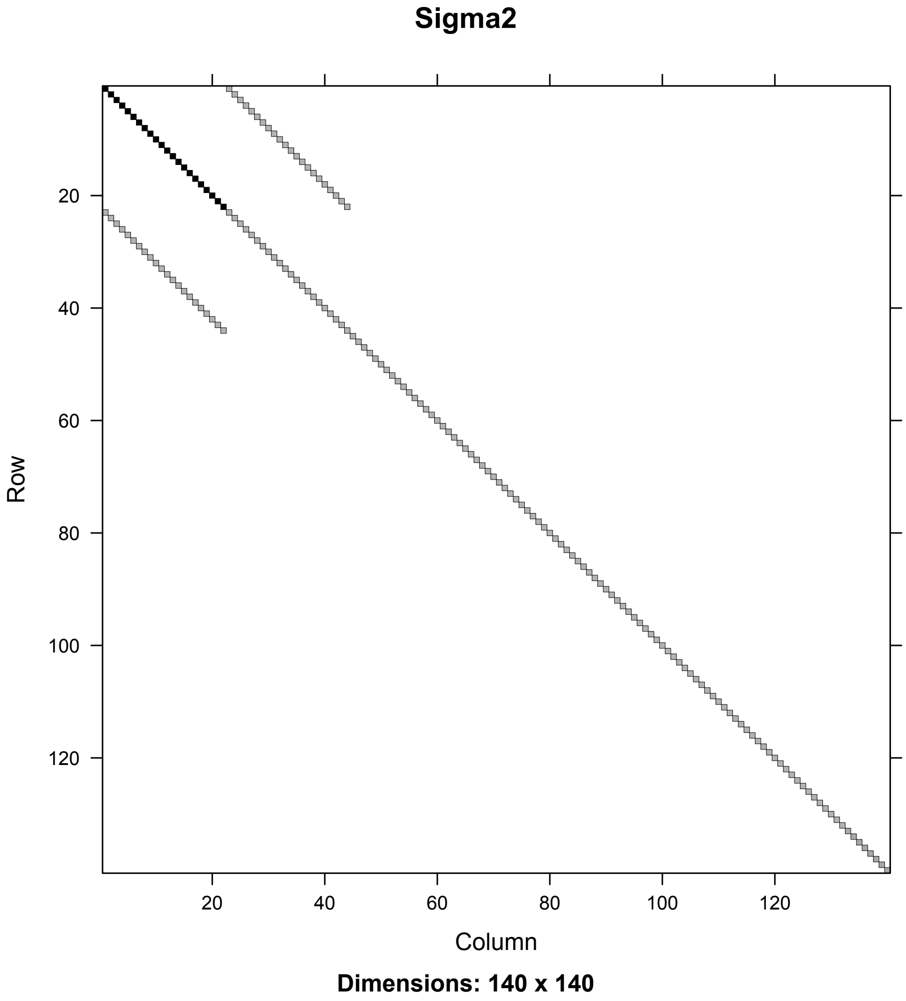


Figure 3.36: Structure of the covariance matrix of the random effects where the enzyme effects are now nested in the compound factor. With 22 compounds and 4 enzyme conditions, we will be predicting 88 enzyme effects, the middle diagonal matrix block.

3.10.1 Nested and partially crossed random effects after removing outlier units

We now take the step of eliminating outlier units altogether in the analysis. Such outlier units do not seem to provide any help in fitting any of the parameters, since the upper asymptotes already vary by file, the slope is held constant, and the IC50 value is clearly different from the units in other files. In fact, they seem to affect the estimation of the IC50s negatively as seen in a few cases. Compare Figures 3.35 and 3.38 where we can see that after we eliminate the bottom row of outliers in Figure 3.35, the fitted curves of the remaining units follow the data much closer in Figure

unit	compound	enz.comp	sfile	d	g	b
105	Paroxetine	AllMix:Paroxetine	Gal_E-1.xls	3933.77	6.66	-0.46
106	Paroxetine	PhaseII:Paroxetine	Gal_E-1.xls	3933.77	6.73	-0.46
107	Paroxetine	P450:Paroxetine	Gal_E-1.xls	3933.77	6.29	-0.46
108	Paroxetine	Noenzyme:Paroxetine	Gal_E-1.xls	3933.77	5.82	-0.46
129	Paroxetine	AllMix:Paroxetine	Gal_E-2.xls	5940.11	6.66	-0.46
130	Paroxetine	PhaseII:Paroxetine	Gal_E-2.xls	5940.11	6.73	-0.46
131	Paroxetine	P450:Paroxetine	Gal_E-2.xls	5940.11	6.29	-0.46
132	Paroxetine	Noenzyme:Paroxetine	Gal_E-2.xls	5940.11	5.82	-0.46
153	Paroxetine	AllMix:Paroxetine	Gal_F-1.xls	3682.61	6.66	-0.46
154	Paroxetine	PhaseII:Paroxetine	Gal_F-1.xls	3682.61	6.73	-0.46
155	Paroxetine	P450:Paroxetine	Gal_F-1.xls	3682.61	6.29	-0.46
156	Paroxetine	Noenzyme:Paroxetine	Gal_F-1.xls	3682.61	5.82	-0.46

Table 3.14: Estimates for Paroxetine from Model 3.

3.38 than in Figure 3.35, the IC50 estimates are shifted in a better position.

For now, we determine outliers through visual inspection. From checking all fitted curves, we take out the units that show clear differences from their set of replicates. Future work will research methods to determine outliers based on the composition of the entire dataset and perhaps a way to incorporate useful information from them while downweighting the outlying parts.

3.11 Standard errors of fixed and random effects in mixed models without concerning errors acquired from estimation of covariance

If we assume the estimated covariance matrix of the random effects are given, then the

$$\mathbb{E}\left\{\begin{bmatrix}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\mathbf{b}}-\mathbf{b}\end{bmatrix}\begin{bmatrix}\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\\ \widehat{\mathbf{b}}-\mathbf{b}\end{bmatrix}^T\right\}=\begin{bmatrix}X^TX & X^TZ\\ Z^TX & Z^TZ+\Sigma^{-1}\end{bmatrix}^{-1}\sigma_\epsilon^2 \qquad (3.69)$$

unit	compound	enz.comp	sfile	d	g	b
101	Tolcapone	AllMix:Tolcapone	Gal_E-1.xls	3997.54	7.13	-0.46
102	Tolcapone	PhaseII:Tolcapone	Gal_E-1.xls	3997.54	6.79	-0.46
103	Tolcapone	P450:Tolcapone	Gal_E-1.xls	3997.54	6.12	-0.46
104	Tolcapone	Noenzyme:Tolcapone	Gal_E-1.xls	3997.54	6.81	-0.46
125	Tolcapone	AllMix:Tolcapone	Gal_E-2.xls	6003.50	7.13	-0.46
126	Tolcapone	PhaseII:Tolcapone	Gal_E-2.xls	6003.50	6.79	-0.46
127	Tolcapone	P450:Tolcapone	Gal_E-2.xls	6003.50	6.12	-0.46
128	Tolcapone	Noenzyme:Tolcapone	Gal_E-2.xls	6003.50	6.81	-0.46
149	Tolcapone	AllMix:Tolcapone	Gal_F-1.xls	3744.14	7.13	-0.46
150	Tolcapone	PhaseII:Tolcapone	Gal_F-1.xls	3744.14	6.79	-0.46
151	Tolcapone	P450:Tolcapone	Gal_F-1.xls	3744.14	6.12	-0.46
152	Tolcapone	Noenzyme:Tolcapone	Gal_F-1.xls	3744.14	6.81	-0.46

Table 3.15: Estimates for Tolcapone from Model 3.

unit	compound	enz.comp	sfile	d	g	b
105	Paroxetine	AllMix:Paroxetine	Gal_E-1.xls	3836.69	6.43	-0.25
106	Paroxetine	PhaseII:Paroxetine	Gal_E-1.xls	3836.69	6.08	-0.25
107	Paroxetine	P450:Paroxetine	Gal_E-1.xls	3836.69	5.99	-0.25
108	Paroxetine	Noenzyme:Paroxetine	Gal_E-1.xls	3836.69	5.26	-0.25
129	Paroxetine	AllMix:Paroxetine	Gal_E-2.xls	5838.56	6.43	-0.25
130	Paroxetine	PhaseII:Paroxetine	Gal_E-2.xls	5838.56	6.08	-0.25
131	Paroxetine	P450:Paroxetine	Gal_E-2.xls	5838.56	5.99	-0.25
132	Paroxetine	Noenzyme:Paroxetine	Gal_E-2.xls	5838.56	5.26	-0.25

Table 3.16: Estimates for Paroxetine from Model 3 with outliers excluded.

where for the nonlinear case, the design matrices X and Z are of dimensions $n \times 3$ and $ntimesq$ such that each row $k = 1, \ldots, n$ is defined as follows:

$$\widehat{X}_k = \left.\frac{\partial f_k}{\partial \boldsymbol{\beta}^T}\right|_{\hat{\boldsymbol{\beta}},\hat{b_k}} \tag{3.70}$$

$$\widehat{Z}_k = \left.\frac{\partial f_k}{\partial b_k^T}\right|_{\hat{\boldsymbol{\beta}},\hat{b_k}} \tag{3.71}$$

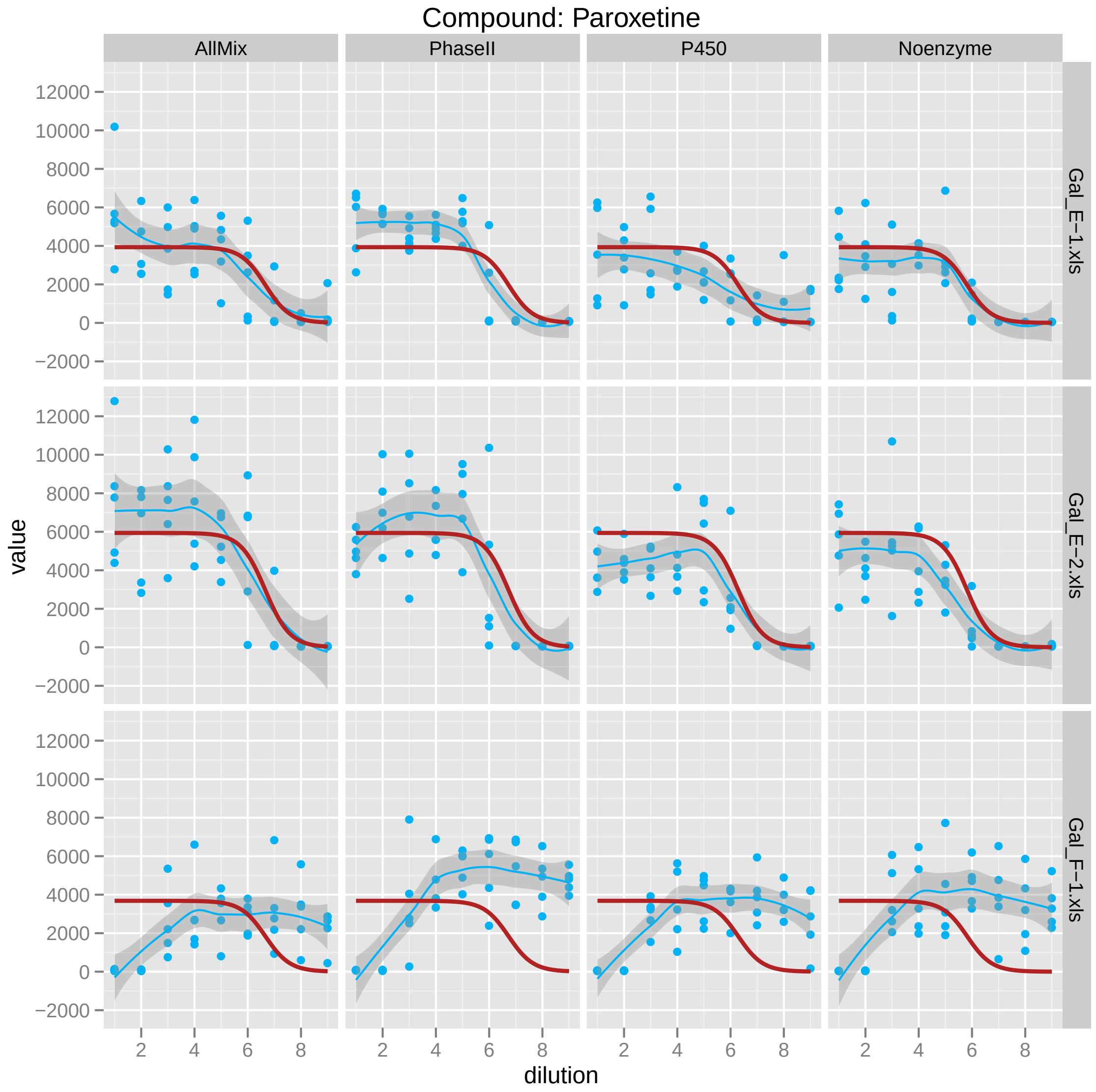


Figure 3.37: Plots of the fitted curves of paroxetine from Model 3.

3.11.1 Covariance matrix of random effects when fixed effects are not estimated

In this chapter, the fixed effects are not estimated along with the random effects. As such, the distribution of the random effects given the data has mean

$$(Z^T Z + \Sigma^{-1})^{-1} Z^T y \tag{3.72}$$

and the variance

$$(Z^T Z + \Sigma^{-1})^{-1} \sigma^2 \tag{3.73}$$

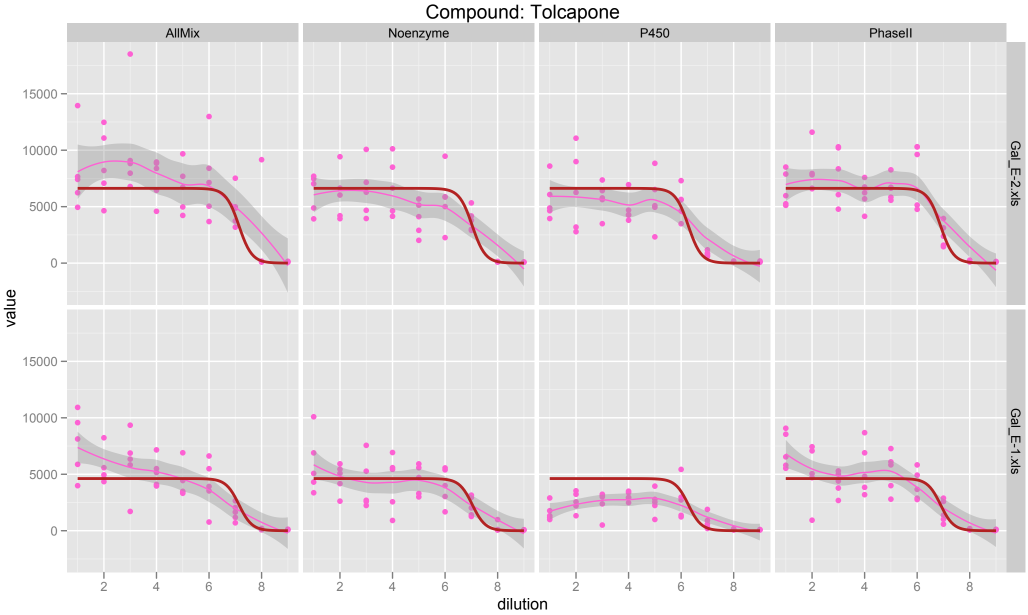


Figure 3.38: Plots of the fitted curves of tolcapone from Model 3.

where Z is a $n \times q$ design matrix, and Σ is the $q \times q$ relative covariance matrix of the random effects.

3.12 Model selection

We can utilize the likelihood ratio test to compare nested models, where the term nested in this case refers to when one model contains a proper subset of the other model's error terms. These models must have the same fixed effects terms since testing for the fixed effects specification will result in inaccurate anticonservative p-values.

Unfortunately the likelihood ratio test is not valid when comparing a model with nested random effects and another model with crossed random effects [16]. Further work is needed to figure out a reasonable way to compare these models.

3.13 Discussion

This chapter has been an exercise in methods to find the right balance in the complexity of the model and faithfulness of the data. Because data is continually and rapidly being generated in sizeable batches, we need to find a model that can fit the data in a reasonable yet general way that can sufficiently cover the inherent characteristics of such data. In addition, even when ideal in theory, a model's complexity may prevent its fitting on less than ideal data.

unit	compound	enz.comp	sfile	d	g	b
101	Tolcapone	AllMix:Tolcapone	Gal_E-1.xls	4620.06	7.14	-0.25
102	Tolcapone	PhaseII:Tolcapone	Gal_E-1.xls	4620.06	6.89	-0.25
103	Tolcapone	P450:Tolcapone	Gal_E-1.xls	4620.06	6.27	-0.25
104	Tolcapone	Noenzyme:Tolcapone	Gal_E-1.xls	4620.06	7.03	-0.25
125	Tolcapone	AllMix:Tolcapone	Gal_E-2.xls	6621.93	7.14	-0.25
126	Tolcapone	PhaseII:Tolcapone	Gal_E-2.xls	6621.93	6.89	-0.25
127	Tolcapone	P450:Tolcapone	Gal_E-2.xls	6621.93	6.27	-0.25
128	Tolcapone	Noenzyme:Tolcapone	Gal_E-2.xls	6621.93	7.03	-0.25

Table 3.17: Estimates for Tolcapone from Model 3 with outliers excluded.

The choice of a model is an evolving process. By involving the relationships found by estimating each unit of data independently, we can construct potential models that will incorporate such robust patterns for all the data simultaneously. Typical model building involves starting with the simplest model and then gradually adding more complexity. That can be a reasonable way to proceed here; however we can see in our situation that it can be more intuitive to figure out what full model to aim for first instead of adding terms haphazardly. Then it is a matter of seeing if we are able to fit that model and if not, continue with making small concessions here and there like independence of random effects, fewer random effects, etc. Through iterations of fitting different models that may involve compromises, we can hopefully settle on a general family of models that can work well on as many future datasets as possible.

Future work may involve developing methods that utilize work that has already been done in sequential estimation, where new incoming data can benefit from existing results and estimates of the fixed and random effects can be further refined each time without having to fit the old and new data together in order to gain the strength.

With using more data to develop the models, we hope to gain more insight into how the different enzyme conditions affect the activities of various compounds. And future models can be further improved with the inclusion of external data on the biological characteristics of the specific reactions of the compounds involved.

unit	compound	enz.comp	sfile	d	g	b
5	Rotenone	AllMix:Rotenone	Gal_A-1.xls	6135.11	4.85	-0.25
6	Rotenone	PhaseII:Rotenone	Gal_A-1.xls	6135.11	4.01	-0.25
7	Rotenone	P450:Rotenone	Gal_A-1.xls	6135.11	4.21	-0.25
8	Rotenone	Noenzyme:Rotenone	Gal_A-1.xls	6135.11	3.86	-0.25
29	Rotenone	AllMix:Rotenone	Gal_A-2.xls	4366.89	4.85	-0.25
30	Rotenone	PhaseII:Rotenone	Gal_A-2.xls	4366.89	4.01	-0.25
31	Rotenone	P450:Rotenone	Gal_A-2.xls	4366.89	4.21	-0.25
32	Rotenone	Noenzyme:Rotenone	Gal_A-2.xls	4366.89	3.86	-0.25
69	Rotenone	AllMix:Rotenone	Gal_B-1.xls	1935.19	4.85	-0.25
70	Rotenone	PhaseII:Rotenone	Gal_B-1.xls	1935.19	4.01	-0.25
71	Rotenone	P450:Rotenone	Gal_B-1.xls	1935.19	4.21	-0.25
72	Rotenone	Noenzyme:Rotenone	Gal_B-1.xls	1935.19	3.86	-0.25
93	Rotenone	AllMix:Rotenone	Gal_B-2.xls	965.25	4.85	-0.25
94	Rotenone	PhaseII:Rotenone	Gal_B-2.xls	965.25	4.01	-0.25
95	Rotenone	P450:Rotenone	Gal_B-2.xls	965.25	4.21	-0.25
96	Rotenone	Noenzyme:Rotenone	Gal_B-2.xls	965.25	3.86	-0.25

Table 3.18: Estimates for Rotenone from Model 3 with outliers excluded.

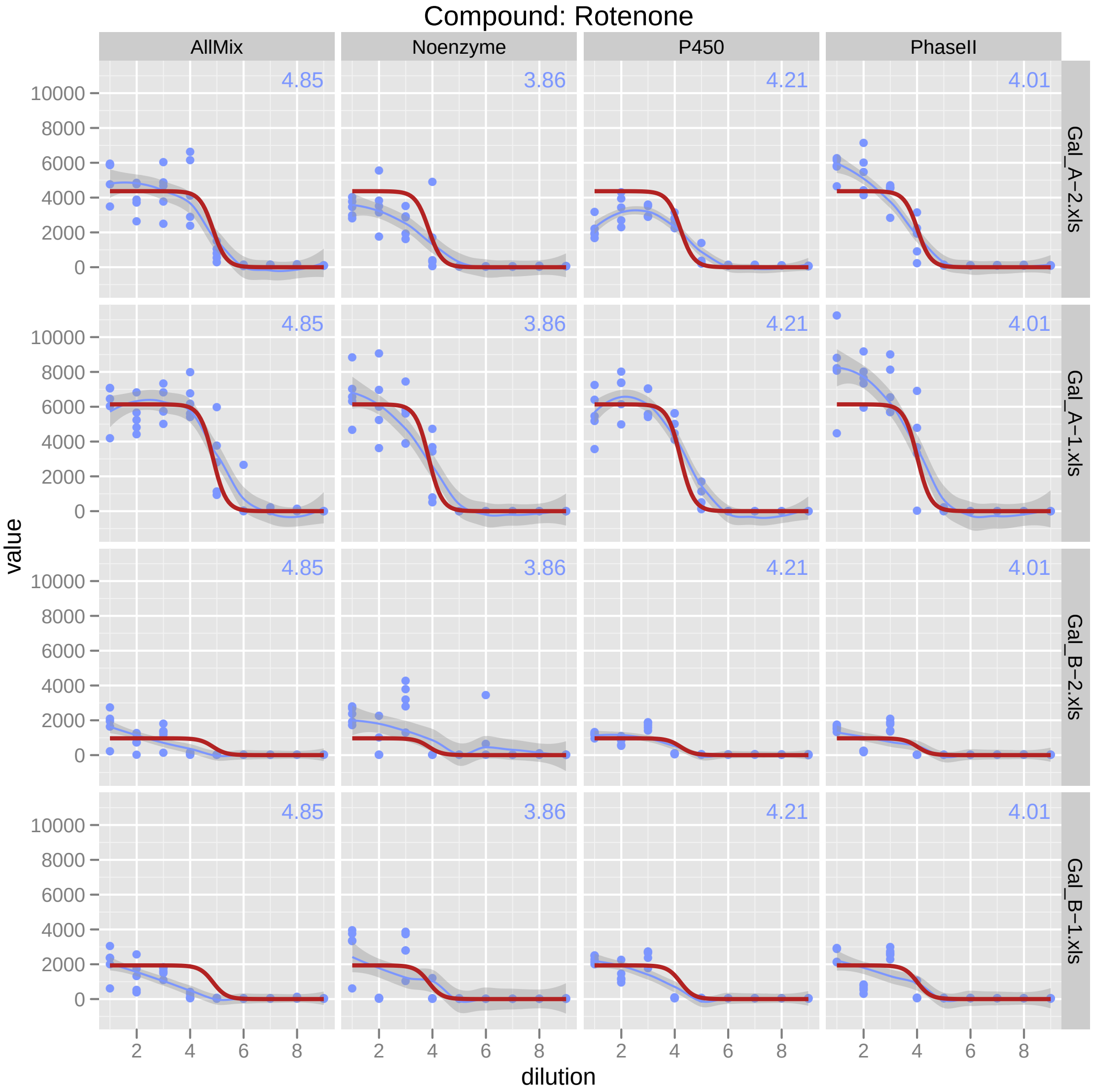


Figure 3.39: Plots of fitted curves for compound Rotenone from the nested model where outliers are removed.

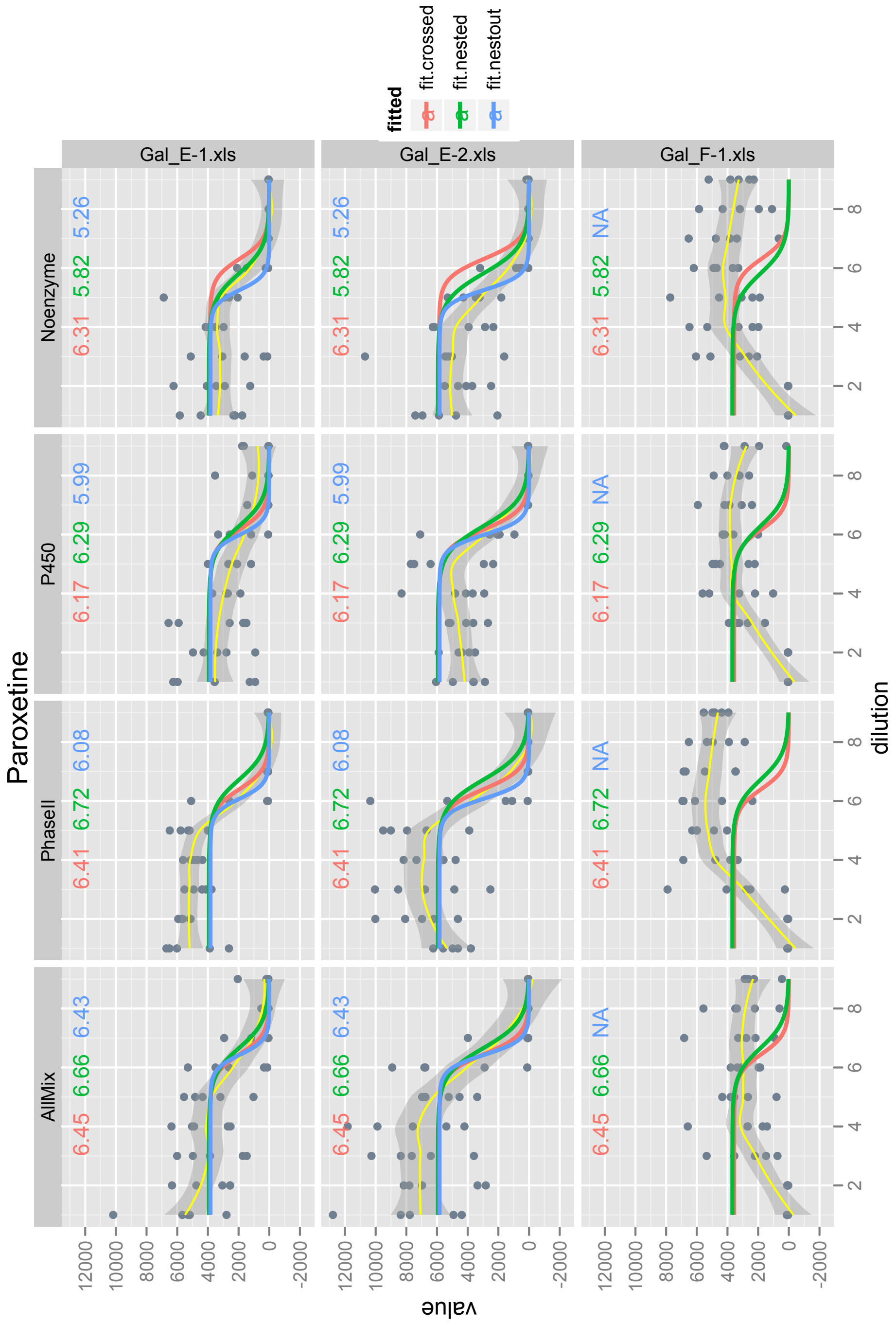


Figure 3.40: Plots to compare the fitted curves from different models.

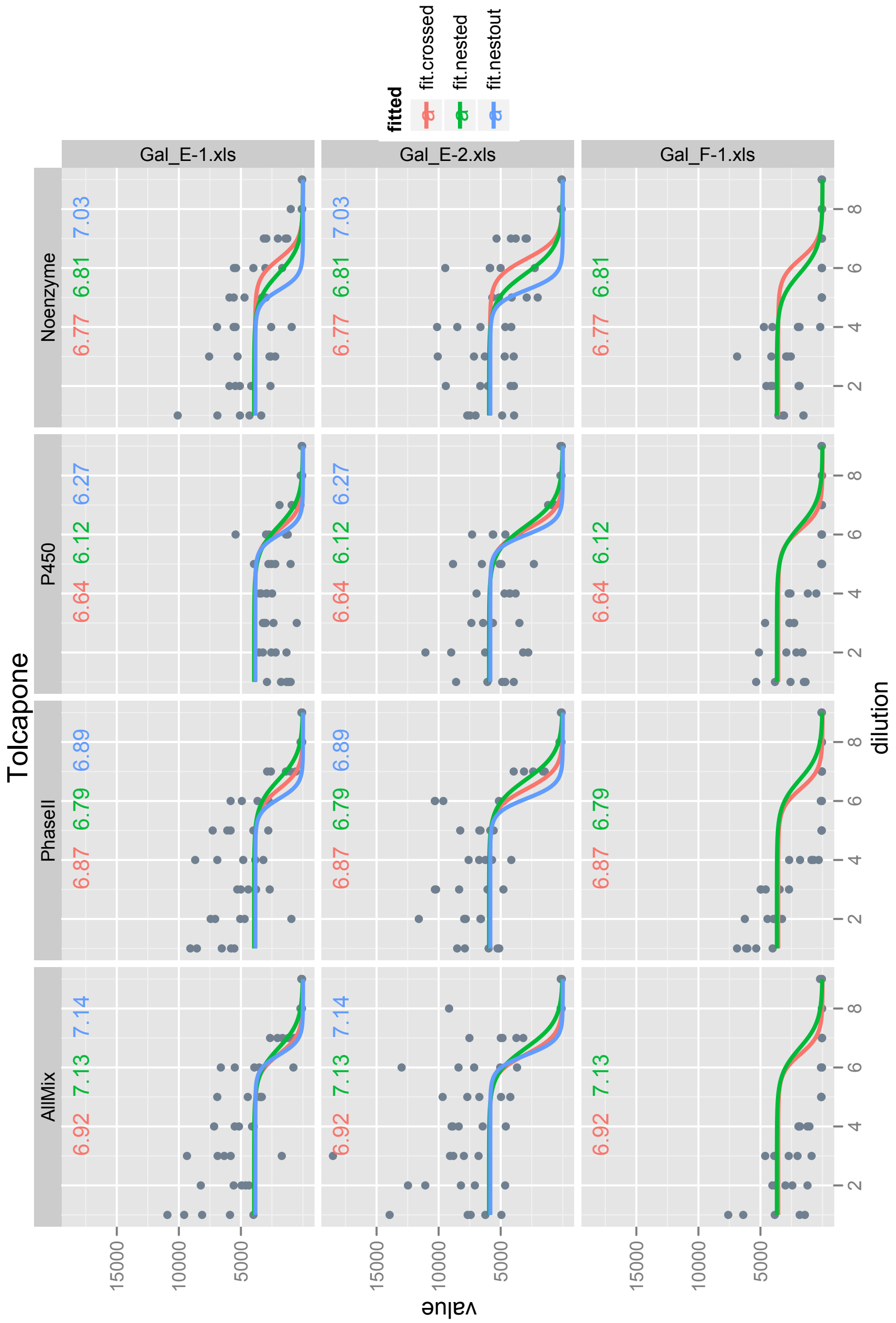


Figure 3.41: Plots to compare the fitted curves from different models.

Chapter 4

Estimation of EC50: Assessing interassay variation with control data

4.1 Introduction

When women become menopausal, many choose hormone replacement therapy (HRT) to relieve significant and unpleasant symptoms caused by the change in hormone levels. While the existing treatments like estradiol can alleviate some or all of the symptoms, they can often be ineffective or cause dangerous side effects such as increased risk for breast or endometrial cancer, since hormones like estradiol do not discriminate the tissues they are affecting. The Bionovo corporation in Emeryville, California has been researching therapies that will reduce these symptoms without the negative effects, by looking into plant extracts that can mimic estradiol's helpful effect without affecting the mammary glands or uterus.

Their product candidates to combat hot flashes and vaginal dryness are estrogen recepter beta agonists formulated from different compounds extracted from plants. They do not activate the estrogen receptor beta (ERβ), known to be implicated in both breast and uterine cancer formation.

4.2 Data

The working dataset for this chapter involves the 18 different preparations that constitute the agonist drug Minerba, designed to reduce the effects of hot flashes.

A cell line named U2OS and cultivated from the bone tissue of a fifteen-year-old human female suffering from osteosarcoma is used for these experiments. Cells are electroporated with a reporter plasmid (ERE, estrogen response element) and a receptor (ERβ) and then plated and treated with specific amounts of drug in 12 well

Dilution	100 mg/ml
1:1000	0.1000
1:2000	0.0500
1:5000	0.0200
1:7500	0.0130
1:15000	0.0070
1:40000	0.0025
1:100000	0.0010
1:200000	0.0005
1:500000	0.0002
1:1000000	0.0001

Table 4.1: Table of dilutions for dataset.

plates. After a period of 24 hour incubation, they are transferred to a 96 well plate to be read in a luminometer that detect the luciferase reagents.

Negative and positive controls are also included onto these plates where the overall goal is to be able to utilize the negative and positive control values to normalize the experimental data. Each plate contains 96 wells, where 90 are used for 3 experiments that involve 10 dilutions with 3 replicates each, and the other 6 are split between 4 negative and 2 positive controls. Since we are dealing with a total of 6 plates, we have 540 observations total for 18 experiments and 36 observations for controls. All the control values are included in Table 4.2 for reference.

The positive controls are estradiol (E2) while the negative controls are simply wells of cell samples alone.

The mapping of the wells is semirandom where every triplicate of dilutions in each unit is physically next to each other on the plate and where the majority of the controls are on the boundary of the plates and especially in the corners where it is riskier for wells to be affected by some systematic factor of the experimental process.

One difference in this dataset from the other two that we have looked at is the nonconstant intervals between the dilutions on the log scale (Table 4.1). This does not cause any change in the usual methods, but we have to be careful to use the correct x values during any model fitting and transformation back to absolute concentration units.

We can visualize all the data on each plate through Figure 4.5 where all the dilution series data is plotted together by common plate, and the negative and positive controls are included on the left and right hand side of each plot respectively. These values can be visually summarized through boxplots in Figure 4.1 and the logged data in Figure 4.2, where the first three boxplots in each of the 6 plots contain the intensity values for the three different experiments on each plate, while the last 2 boxplots are for the negative and positive controls. Note that the boxplots for the positive control values actually only contain 2 values. Even though all 18 experiments are different, we do see a strange pattern where the majority of the observations for the second experiment in five of the plates reside in a smaller range than the other two experiments on the same plate. Perhaps this is an artifact of the location of the wells where these measurements are taken, if the wells for each experiment are next to each other and if there is spatial heterogeneity on the plate as a whole, caused by the scanner or some other systematic effect.

In general, we do see that the control values, which ideally should be the same by type across the plates, differ in roughly the same relationship as the experimental data. This suggests that we may be able to normalize the data such that the distribution for each plate will be roughly similar and at around the same mean. This can also be easily seen in the boxplots of the raw dilution series data arranged by experiment and plate separately (Figures 4.3 and 4.4).

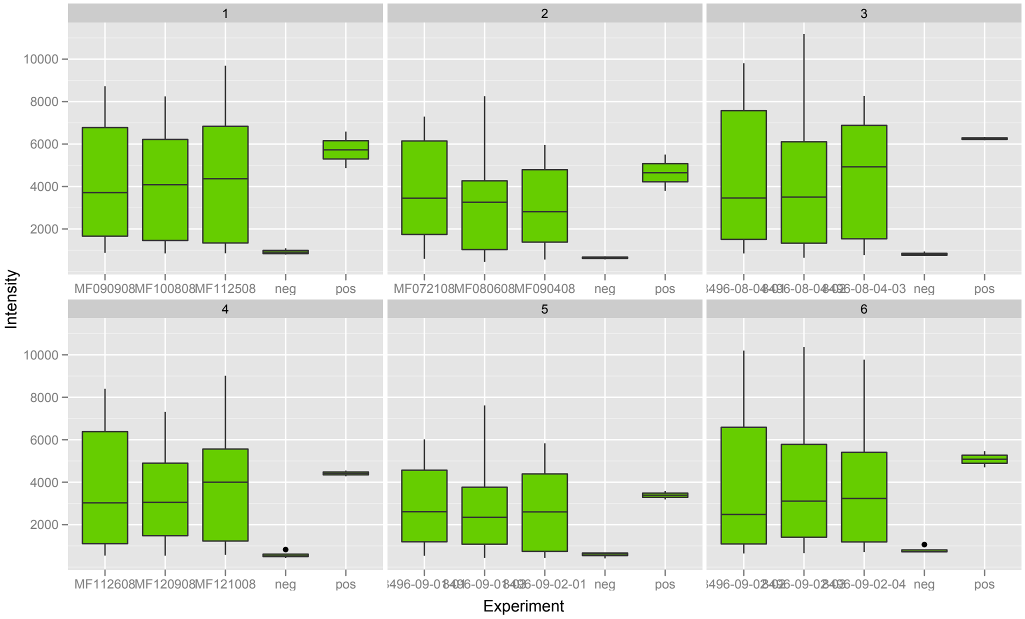


Figure 4.1: Boxplots of all raw data including controls separated into each plot by the plate, and then by experiment or control. Note that the boxplots for the positive controls actually only contain 2 values.

4.3 Analysis of individual nonlinear least square estimates

We include some basic analysis of the individual nonlinear least squares estimates of the four parameters, ϕ_0 (lower asymptote), ϕ_1 (upper asymptote), ϕ_2 (EC50),

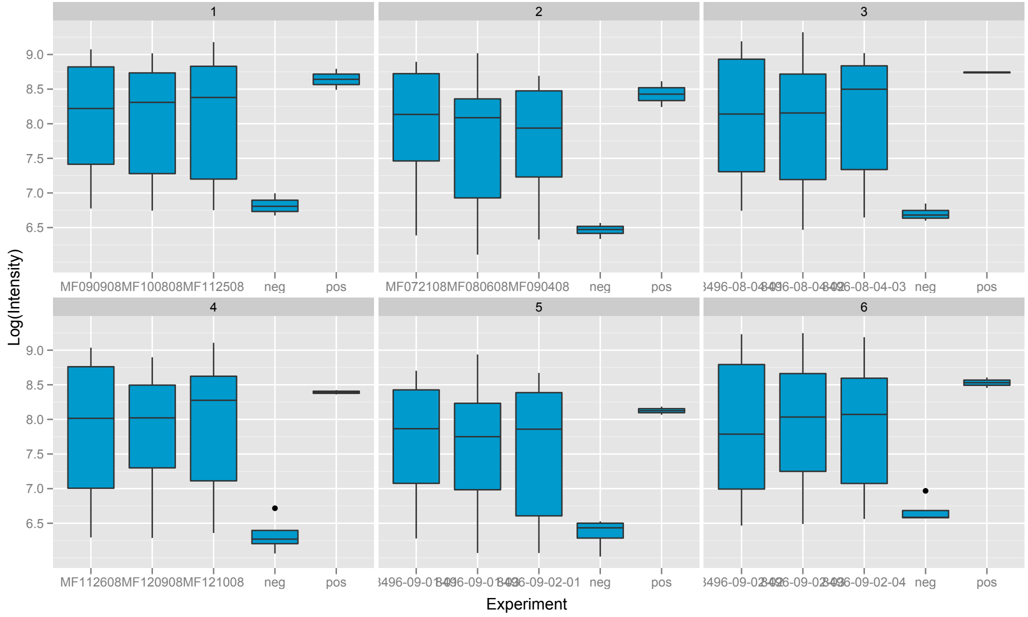


Figure 4.2: Boxplots of all logged raw data including controls separated into each plot by the plate, and then by experiment or control. Note that the boxplots for the positive values actually only contain 2 values and the negative ones contain 4.

ϕ_3 (slope), Table 4.3. We also estimate these parameters robustly using the Tukey biweight. Unlike data from the two previous chapters, each unit was successfully fit here. The medians of the estimates of these four parameters are included in Table 4.4, where we can see that there is not much difference between the ordinary and robust estimates in general.

An overview of where these estimates compare in regards to the control data is included in Figure 4.6. It is expected that the variance of the upper asymptotes is higher due to the higher values. Because of this, the plot of the logged values (Figure 4.7) may be more useful in evaluating potential relationships.

An interesting thing to see in the plot comparing the EC50 and slope estimates (Figure 4.10) is that the larger the EC50 estimates, the steeper the slope. This is most likely due to the fact that many of the units do not have enough data to estimate an upper asymptote sufficiently and thus, individual nonlinear least squares will give a larger than expected estimate for the upper asymptote that translates to a steeper slope for the curve to climb to this asymptote and shifting the EC50 estimate to the right (larger).

Ratios between asymptote estimates and control values

We can now revisit the issue of utilizing control values for normalization by looking at relationship of the lower and upper asymptote estimates (from individual nonlinear

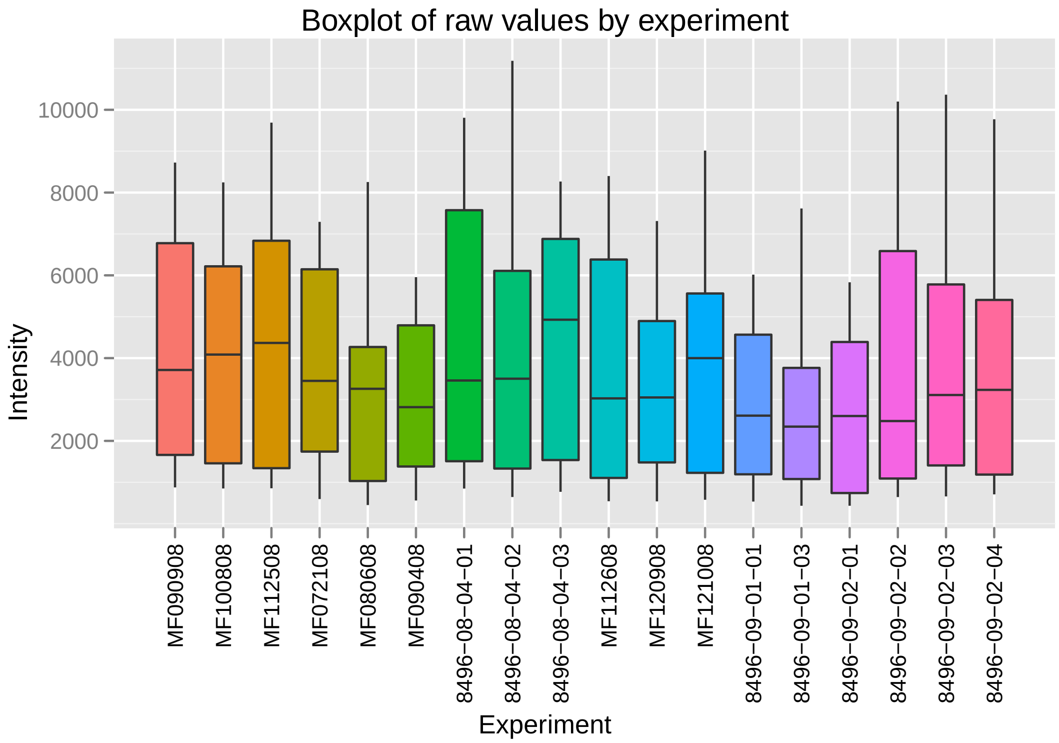


Figure 4.3: Boxplots of raw dilution series data by experiment.

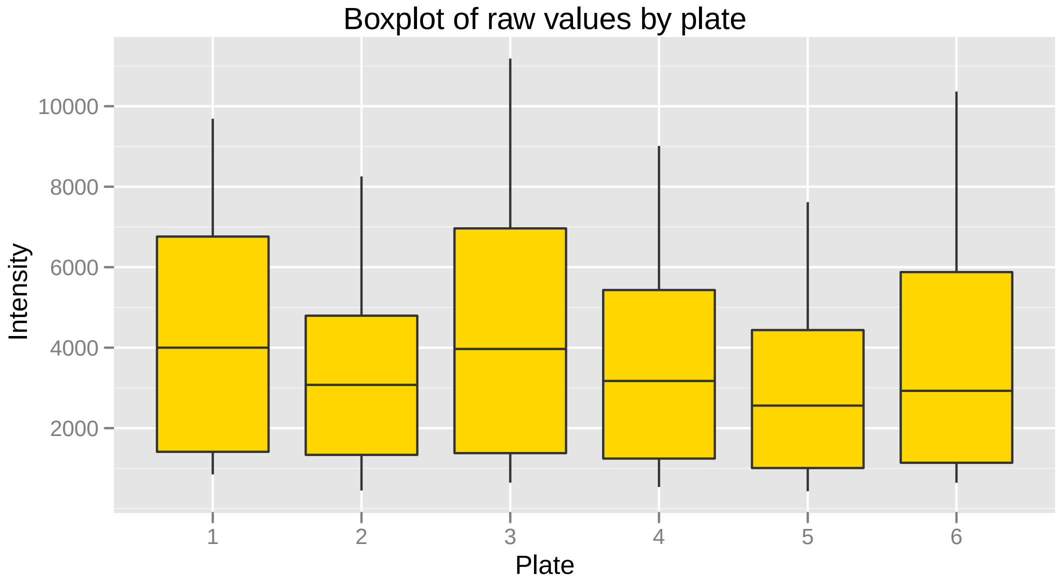


Figure 4.4: Boxplot of raw dilution series data by plate.

	Negative				Positive	
Plate	Rep 1	Rep 2	Rep 3	Rep 4	Rep 1	Rep 2
1	855	792	956	1091	6587	4867
2	565	666	711	628	5505	3795
3	771	941	824	735	6330	6181
4	519	826	430	539	4549	4278
5	587	411	683	660	3585	3188
6	727	713	723	1062	5462	4700

Table 4.2: Table of control data

least squares) and the negative and positive controls. The ratios of the medians of these values are included in Table 4.5 where the positive controls are roughly 6 times the negative controls (Figure 4.13) and the lower asymptotes are roughly the same values as the negative controls (Figure 4.14; best fit line if intercept is set to zero has slope 0.95.). The relationships between the upper and lower asymptotes (Figure 4.15) is a bit more varied, perhaps due to the nature of estimating each independently, similarly with the ratios between the upper asymptotes and the positive controls (Figure 4.16).

4.3.1 Simple linear and log linear models for control values

Exploratory analysis of control values

To begin, we analyze only the control data to see how being on different plates affect the values of the positive and negative controls. Individual nonlinear least squares estimates may sometimes be a bit extreme depending on whether or not a set of points has enough data to cover the asymptote areas sufficiently, so although we can definitely include them here, we should take a look at the controls on their own since those are actual observations.

Figure 4.17 provides a way to visualize the logged control values all at once. Since the control values are not supplied as pairs of negative with positive values, we are unable to produce regular scatterplots that show all the data. Instead, one thing to do is to plot the medians of the negative values in each plate against the medians of the positive values in the corresponding plates. Although it can show a general trend, such a plot doesn't show the spread of the points about the medians. Figure 4.17 show all the control values in a way that we can see the general trend in a scatterplot and also see where all the points lie. We see that overall, the positive and negative control values correspond pretty closely.

Linear model

For a simple linear model of the control values where we can adjust for plate effects,

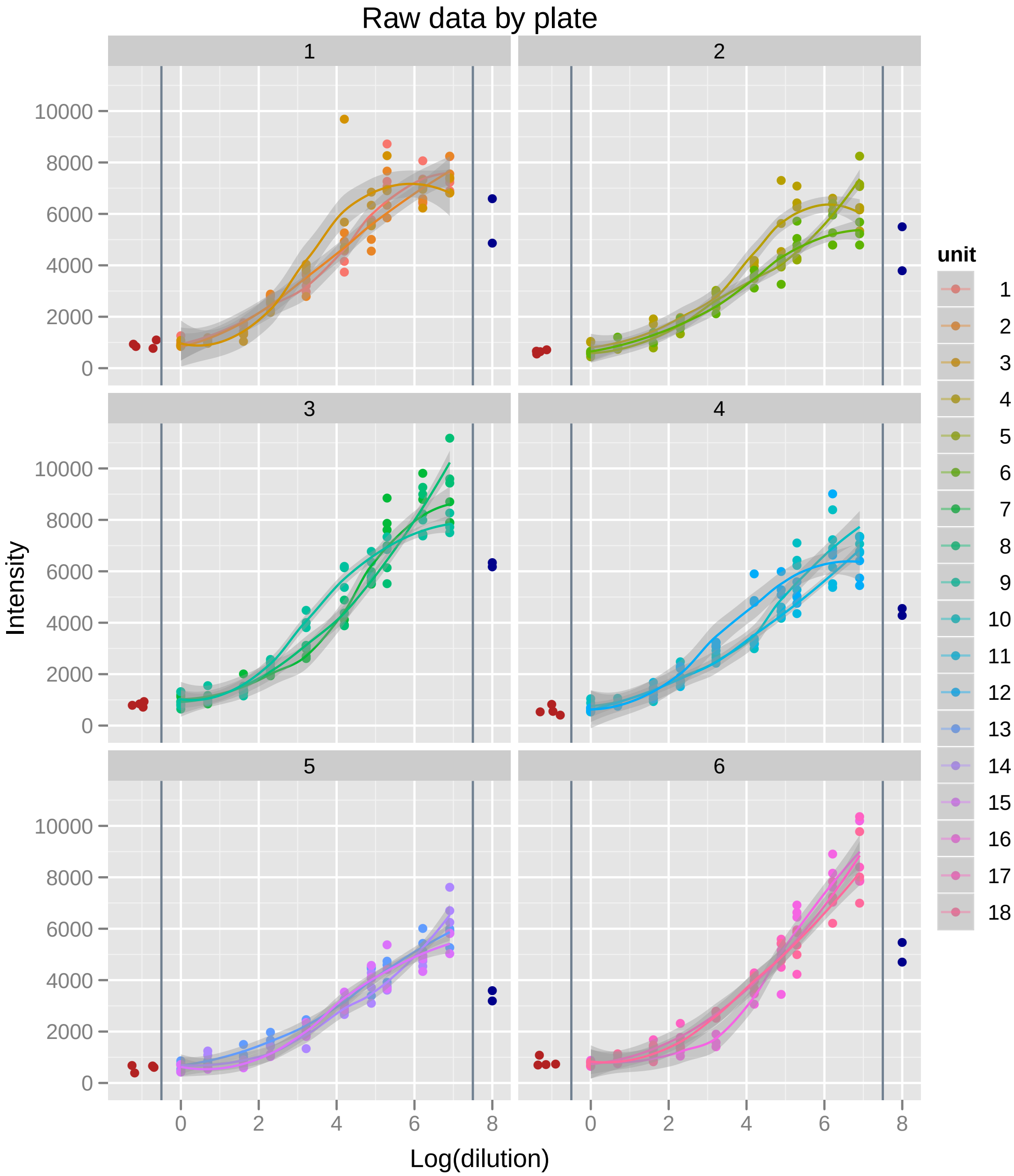


Figure 4.5: Plots of original dataset by plate, including negative controls (left, red) and positive controls (right, blue)

we assume that the effects are additive. We let t_{cjk} refer to the kth control observation on plate j, indexed by whether it is negative ($c = 1$) or positive ($c = 2$). Each observation results from the true negative or positive control (τ_1, τ_2) being affected

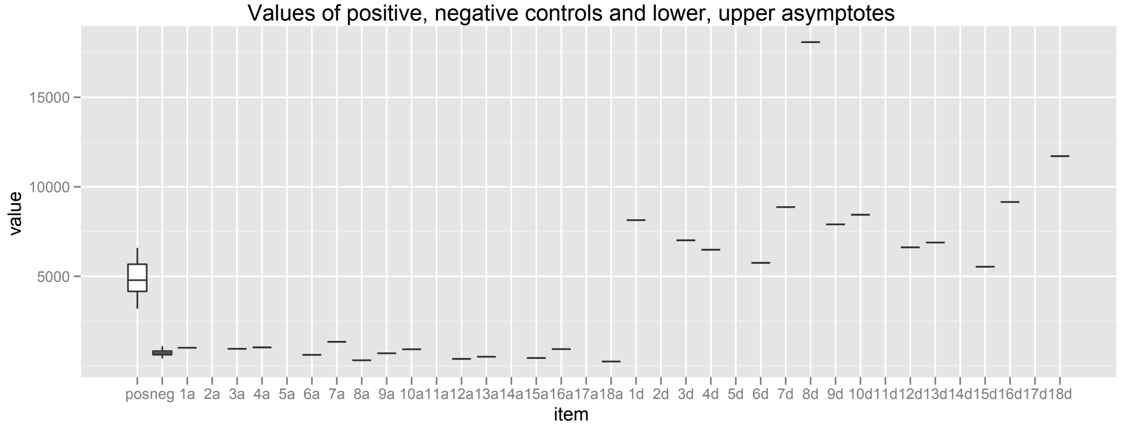


Figure 4.6: Boxplots of the positive, negative controls along with individual estimates of the lower and upper asymptotes.

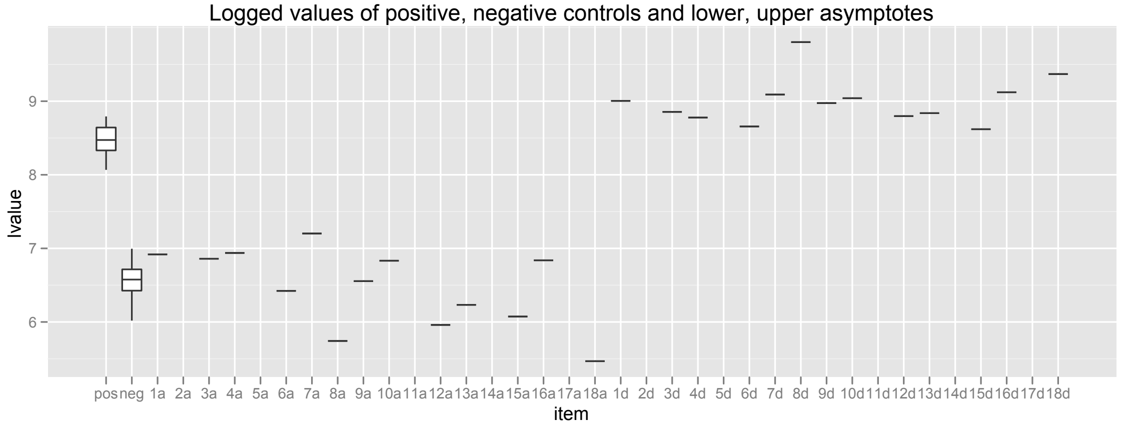


Figure 4.7: Boxplots of the logged positive, negative controls along with logged individual estimates of the lower and upper asymptotes.

additively by a plate effect m_j and by random error ϵ_{cjk}, where $\epsilon_{cjk} \sim \mathcal{N}(0, \sigma_\tau^2)$.

$$t_{cjk} = \tau_i + m_j + \epsilon_{cjk} \tag{4.1}$$

We also have to set the constraint that $\sum_{j=1}^{6} m_j = 0$ in order to obtain estimates.

Using lm() in R, we obtain the following estimates in Table 4.7 (for reference). We ascertain whether the model and fitted values are reasonable with Figure 4.18, which provides a diagnostic check, while the upper left hand plot in Figure 4.20 shows the fitted values against the observations. Both sets of plots suggest that this additive model isn't adequate. This leads to a multiplicative model that will be fitted as a log linear model in the next section.

Log linear model

We make a minor change to the model above by defining the plate effects to be

		Ordinary				Robust			
unit	plate	ϕ_0	ϕ_1	ϕ_2	ϕ_3	ϕ_0	ϕ_1	ϕ_2	ϕ_3
1	1	955.66	8174.09	4.02	1.05	869.54	8282.43	4.07	1.18
2	1	733.94	8337.59	4.07	1.33	730.13	8145.97	3.99	1.31
3	1	939.74	7011.89	3.12	0.56	872.81	7043.64	3.22	0.69
4	2	834.69	6539.21	3.66	0.79	745.62	6649.72	3.73	0.92
5	2	402.95	14612.75	7.04	2.02	445.91	11498.72	6.13	1.77
6	2	633.18	5738.47	3.87	1.08	621.51	5704.51	3.83	1.08
7	3	1119.24	9004.37	4.39	0.77	942.61	9172.16	4.48	1.00
8	3	694.63	15230.45	5.93	1.59	714.02	13500.33	5.53	1.49
9	3	769.29	7867.41	3.42	0.92	789.68	7768.83	3.38	0.87
10	4	714.79	8751.61	4.73	1.13	638.00	8987.46	4.90	1.27
11	4	459.17	9286.84	5.34	1.71	467.45	8884.84	5.15	1.61
12	4	505.29	6568.14	3.36	0.94	519.41	6256.53	3.28	0.90
13	5	564.63	6794.75	4.60	1.35	562.42	6745.87	4.55	1.32
14	5	423.73	18776.38	8.29	2.03	449.94	14528.69	7.50	1.90
15	5	507.84	5488.04	4.04	0.91	523.98	5343.66	4.00	0.86
16	6	893.30	9185.37	4.87	0.70	900.06	8562.97	4.69	0.64
17	6	616.18	13511.32	6.03	1.62	674.28	10299.56	5.12	1.35
18	6	628.93	10263.56	5.22	1.39	669.10	9123.99	4.84	1.25

Table 4.3: Table of individual nonlinear least squares estimates (lower and upper asymptotes, EC50, and slope). Estimates were also obtained robustly using the Tukey biweight.

multiplicative effects. To fit this, we log both sides of the model to obtain:

$$\log(t_{cjk}) = \log(\tau_c) + \log(m_j) + \log(\epsilon_{cjk}) \tag{4.2}$$

where the random error is now normally distributed on a log scale, $\log(\epsilon_{cjk}) \sim \mathcal{N}(0, \sigma_\tau^2)$, and $\sum_{j=1}^{6} \log(m_j) = 0$.

Results from this are better than from the previous model (Table 4.8), where we can see in good diagnostic plots in Figure 4.19, and the fitted line in the lower left plot in Figure 4.19. In the same figure in the right column, we see that using a mixed effects model on the same data where the plate effect is random results in extremely close values.

We can now continue with building a larger model to fit the experimental data where the control data will be used to normalize the results. We assume that data from each plate is scaled by a multiplicative factor that will be determined by the control values.

	ϕ_0 (lower asym)	ϕ_1 (upper asym)	ϕ_2 (EC50)	ϕ_3 (slope)
Ordinary	663.91	8544.60	4.50	1.11
Robust	671.69	8422.70	4.51	1.21

Table 4.4: Medians of the nonlinear least squares estimates of the four parameters across all 18 experiments. We also include robust version of the 18 sets of estimates.

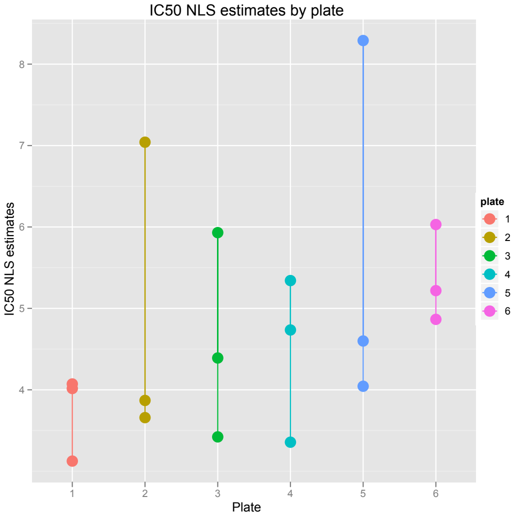


Figure 4.8: Plot of the EC50 nonlinear least squares estimates by plate

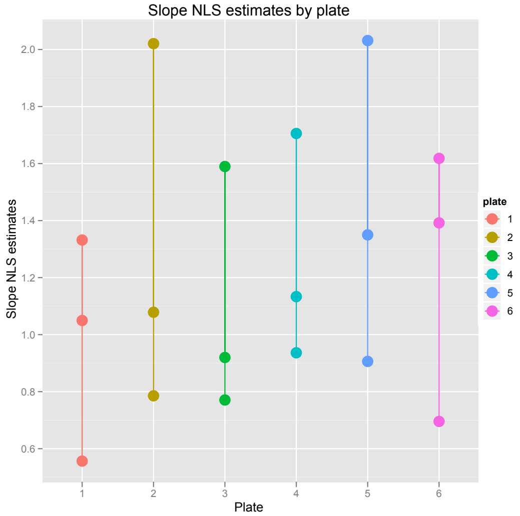


Figure 4.9: Plot of the slope nonlinear least squares estimates by plate

4.4 Mixed effects model

To implement this flexibly, we once again rely on using random effects in our model. For the dilution series data, the observations y_{ij} are the intensities from experiment i on the plate j, and they are modeled with a four parameter logistic curve where the lower and upper asymptotes of each unit, $\phi_{0,ij}$ and $\phi_{1,ij}$, are the products of the mean asymptote values and their corresponding plate factor m_j. To accommodate the differences in each experiment, we include a random effect $b_{2,i}$ for the key parameter $\phi_{2,ij}$, the EC50 value. The slope parameter $\phi_{3,ij}$ remains constant across all units unless there is evidence that there are significant effects from either the experimental or plate factor.

$$y_{ij} = f(x, \boldsymbol{\phi}_{ij}) = \phi_{0,ij} + \frac{\phi_{1,ij} - \phi_{0,ij}}{1 + \exp\left(-\frac{x - \phi_{2,ij}}{\phi_{3,ij}}\right)} + \epsilon_{ij} \tag{4.3}$$

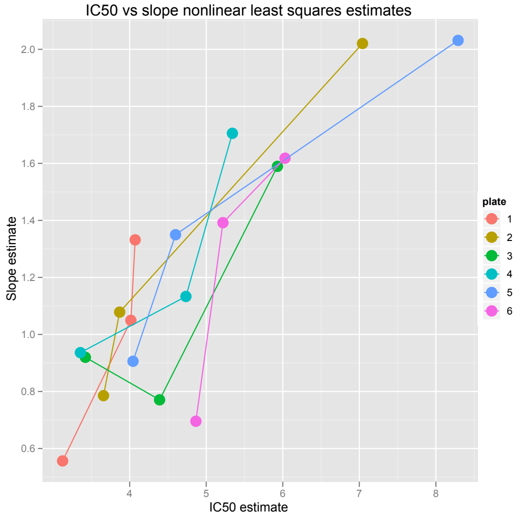


Figure 4.10: Plot of individual nonlinear least squares estimates of the slope against the IC50, color coded by which plate they belong on.

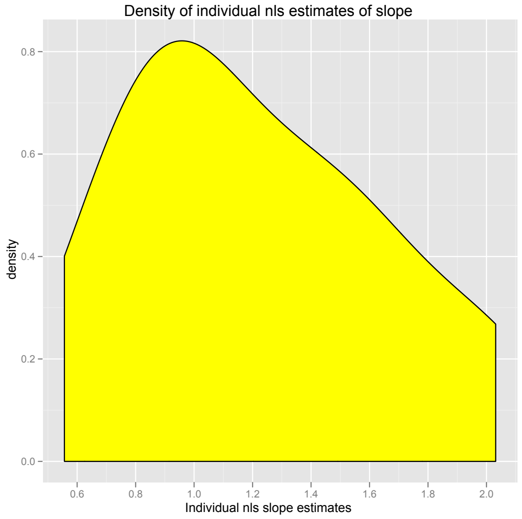


Figure 4.11: Density of individual nonlinear least squares estimates of slope

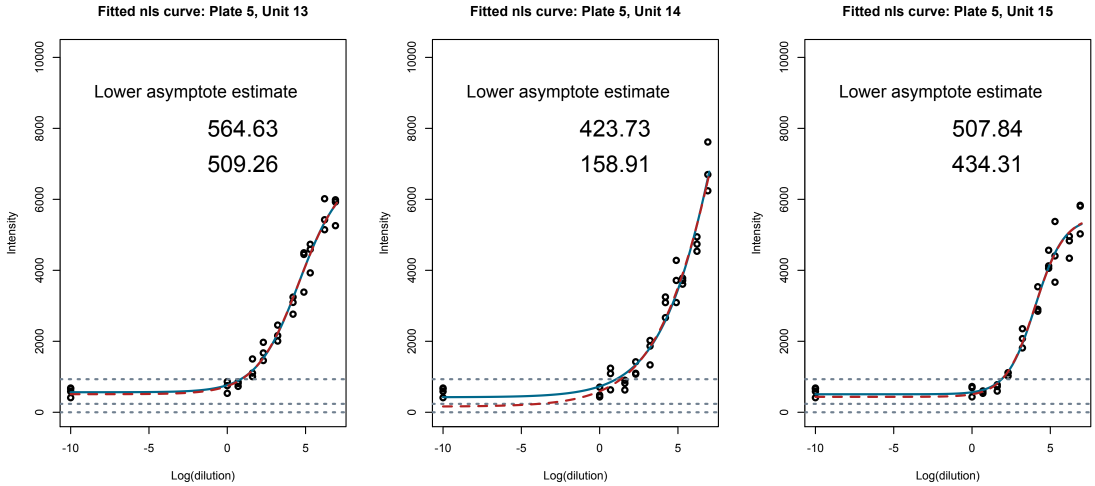


Figure 4.12: Plots of the data of Plate 5 where two different fitted curves are compared. The red dotted line is the original best fit curve through the dilution series data only. The blue solid line is the best fit curve through the data and anchored by the negative controls on the left side. Both sets of lower asymptote estimates are included on each unit's plot where the first is the estimate influenced by the negative controls and the second is the original estimate.

Plate	1	2	3	4	5	6
Positive / negative controls	6.32	7.19	7.84	8.34	5.43	7.01
Upper / lower asymptotes	8.66	10.24	11.53	16.70	13.56	18.62
Upper / positive	1.42	1.39	1.42	1.91	2.03	2.31
Lower / negative	1.04	0.98	0.96	0.96	0.81	0.87

Table 4.5: Table of ratios of median asymptote estimates from nonlinear least squares and median control values by plate.

	Within SS	Between SS	Total SS
Postive controls	3,358,198 (0.25)	10,283,575 (0.75)	13,641,773
Negative controls	308,053 (0.43)	409,610 (0.57)	717,664

Table 4.6: Table of within, between, and total sums of squares for negative and positive control values by plate. The proportions of each sum are included in the parentheses.

where

$$\boldsymbol{\phi}_{ij} = \begin{bmatrix} \phi_{0,ij} \\ \phi_{1,ij} \\ \phi_{2,ij} \\ \phi_{3,ij} \end{bmatrix}$$

$$\phi_{0,ij} = \beta_0 m_j; \qquad \phi_{1,ij} = \beta_1 m_j; \qquad \phi_{2,ij} = \beta_2 + b_i; \qquad \phi_{3,ij} = \beta_3$$

Because $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon^2)$, the observations are assumed to be normally distributed. For the control data, the observations are a $c \times 1$ vector $\mathbf{t}$ where its log is normally distributed with mean as the sum of the logged true values τ and plate effects $\log(\mathbf{m})$[1] and variance as σ_δ^2.

$$\mathbf{y}|\mathbf{b}, \mathbf{m} \sim \mathcal{N}(f(\mathbf{x}, \boldsymbol{\phi}(\boldsymbol{\beta}, \mathbf{b}, \mathbf{m})), \sigma_\epsilon^2 I_n) \tag{4.4}$$

$$\log(\mathbf{t})|\mathbf{m} \sim \mathcal{N}(X \log(\boldsymbol{\tau}) + Z \log(\mathbf{m}), \sigma_\delta^2 I_c) \tag{4.5}$$

where the design matrices X and Z are of dimensions $n \times 2$ and $n \times 6$.

The random effects $\mathbf{b}$ and $\log(\mathbf{m})$ are independent and also follow the normal distribution with mean zero and variances $\sigma_\epsilon^2 \Sigma_1$ and $\sigma_\delta^2 \Sigma_2$.

$$\mathbf{b} \sim \mathcal{N}(0, \sigma_\epsilon^2 \Sigma_1) \tag{4.6}$$

$$\log(\mathbf{m}) \sim \mathcal{N}(0, \sigma_\delta^2 \Sigma_2) \tag{4.7}$$

[1] Note that the log of a vector $\mathbf{x}$ is defined as the vector of the logged elements of $\mathbf{x}$.

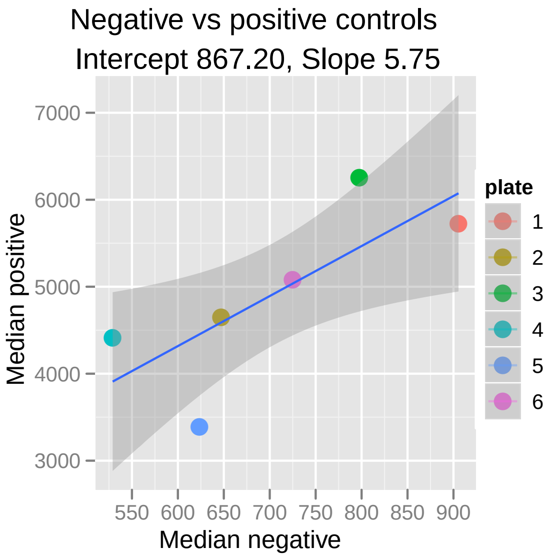


Figure 4.13: Plot of the median negative vs median positive control values.

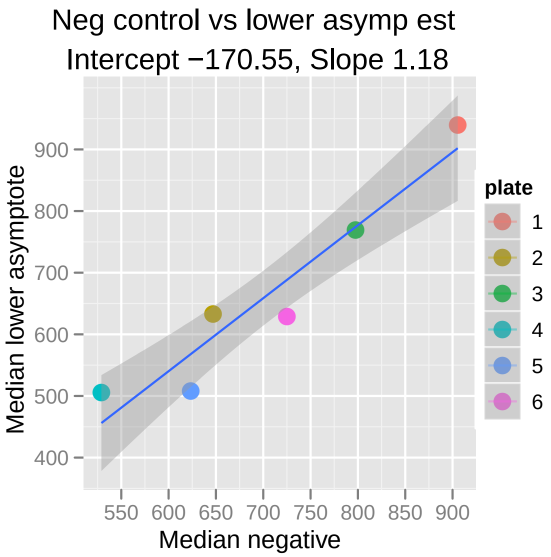


Figure 4.14: Plot of the median negative controls vs median lower asymptote estimates.

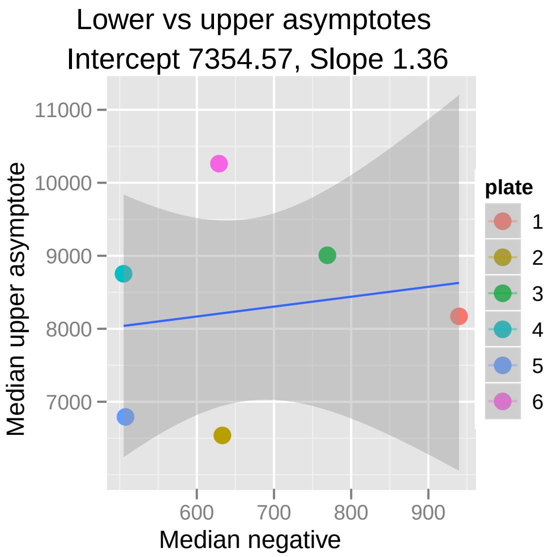


Figure 4.15: Plot of median lower asymptotes vs median upper asymptotes

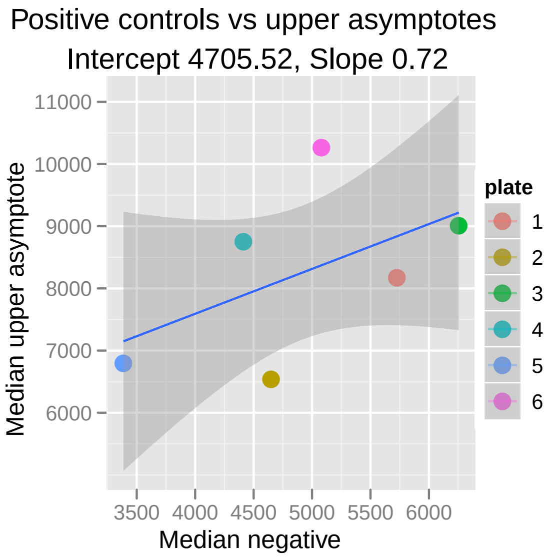


Figure 4.16: Plot of median positive controls vs median upper asymptotes

The relative covariance matrices Σ_1 and Σ_2 are diagonal matrices of dimensions q_1

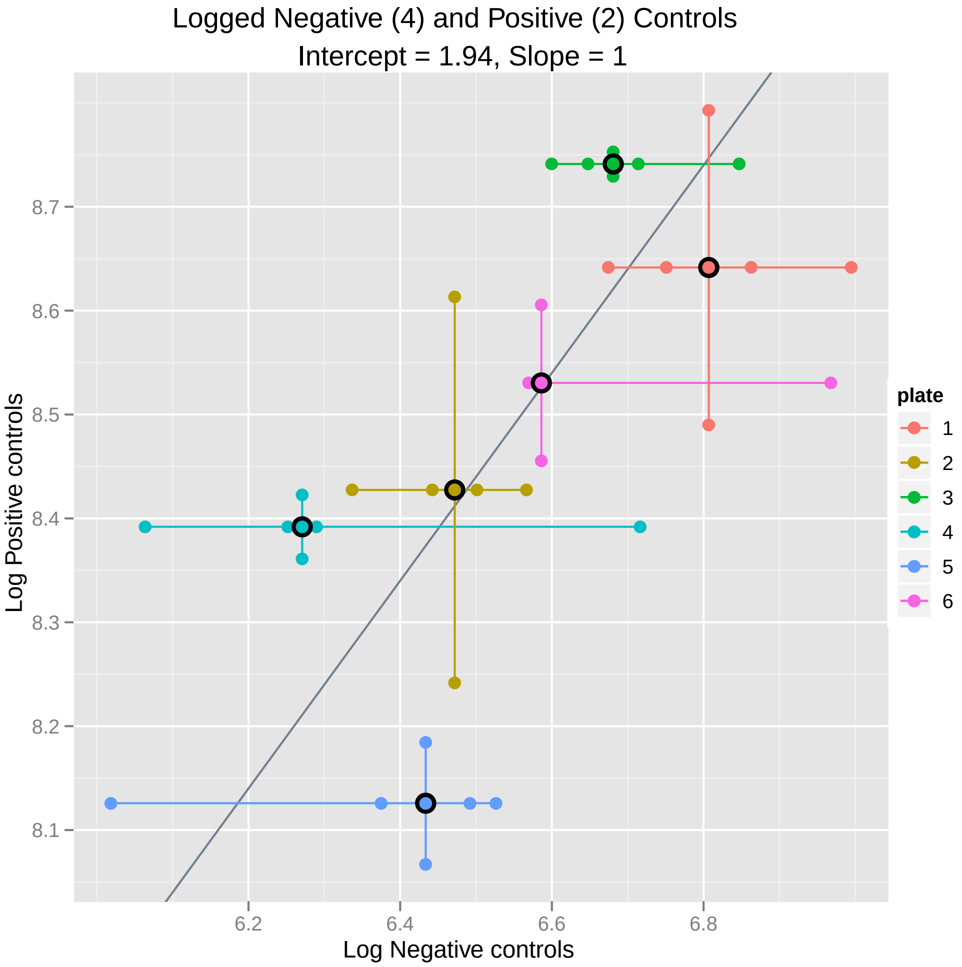


Figure 4.17: Plot of all logged negative and positive control data. Since there are four replicates of negative controls per plate and two replicates of positive controls per plate, we cannot use a typical scatterplot to visualize the relationship between the two. Instead, for each negative control value, we pair it with the median of all the positive controls in the same plate; so we see four negative values for each plate on horizontal lines located at the median of the positive values of the corresponding plate, and vice versa for the positive control values. The median values for both sets of controls are distinguished by points surrounded by black borders. The best fit line through these medians is included.

and q_2, defined as

$$\Sigma_1 = \frac{\sigma_b^2}{\sigma_\epsilon^2} I_{q_1} \tag{4.8}$$

$$\Sigma_2 = \frac{\sigma_m^2}{\sigma_\delta^2} I_{q_2} \tag{4.9}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1126.90	237.21	4.75	0.00
control:positive	4193.29	197.37	21.25	0.00
plate2	-546.33	322.30	-1.70	0.10
plate3	105.67	322.30	0.33	0.75
plate4	-667.83	322.30	-2.07	0.05
plate5	-1005.67	322.30	-3.12	0.00
plate6	-293.50	322.30	-0.91	0.37

Table 4.7: Estimates from a simple linear model on control values by plate

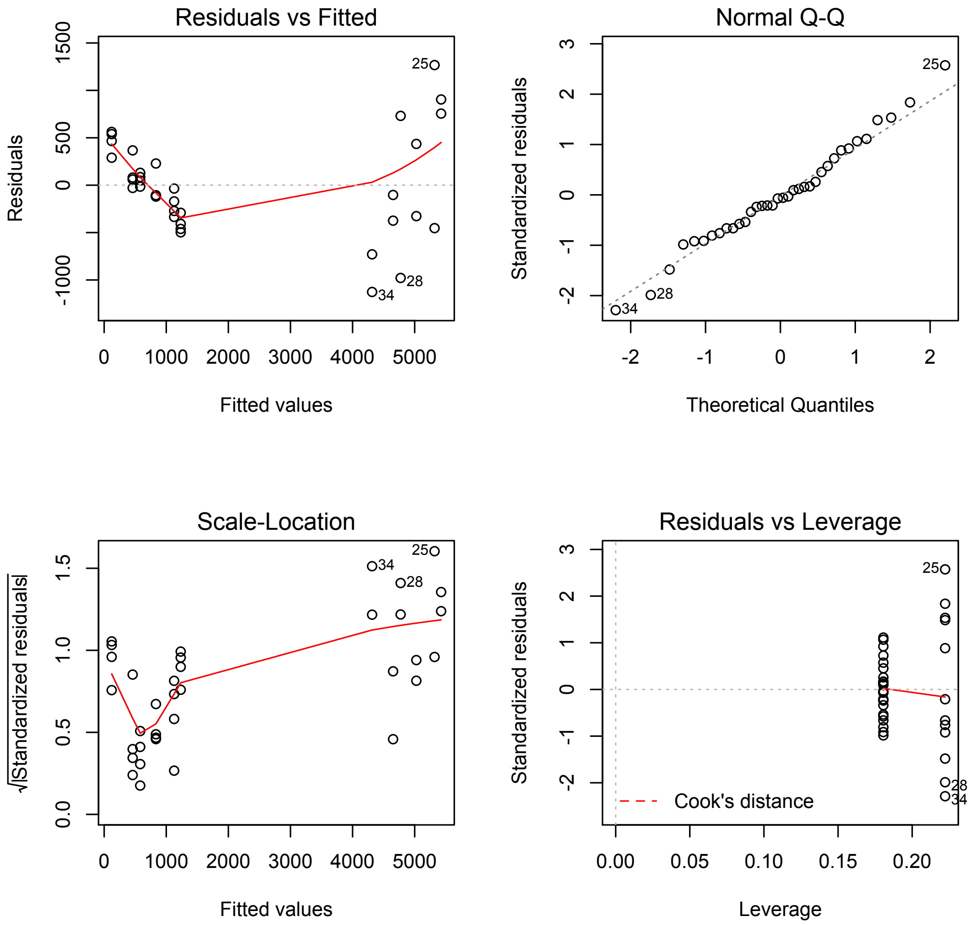


Figure 4.18: Regression diagnostic plots for simple linear model. The residual plot in the upper left corner show heteroscedasticity in the points, the variance increasing with the value.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.79	0.07	93.01	0.00
control:positive	1.92	0.06	31.60	0.00
plate2	-0.31	0.10	-3.13	0.00
plate3	-0.05	0.10	-0.46	0.65
plate4	-0.41	0.10	-4.14	0.00
plate5	-0.48	0.10	-4.88	0.00
plate6	-0.13	0.10	-1.34	0.19

Table 4.8: Estimates from a log linear model on control values by plate

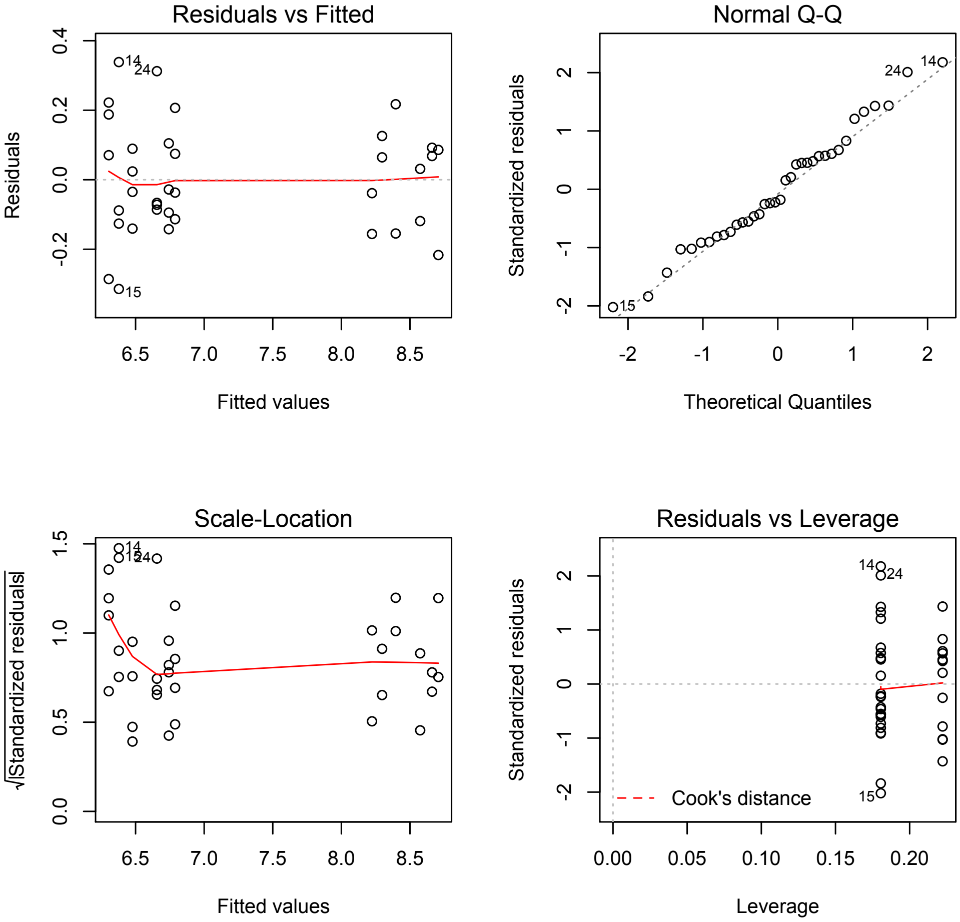


Figure 4.19: Regression diagnostic plots for simple log linear model. Compared to the ones for the simple linear model, these plots suggest that a log linear model is the more appropriate choice.

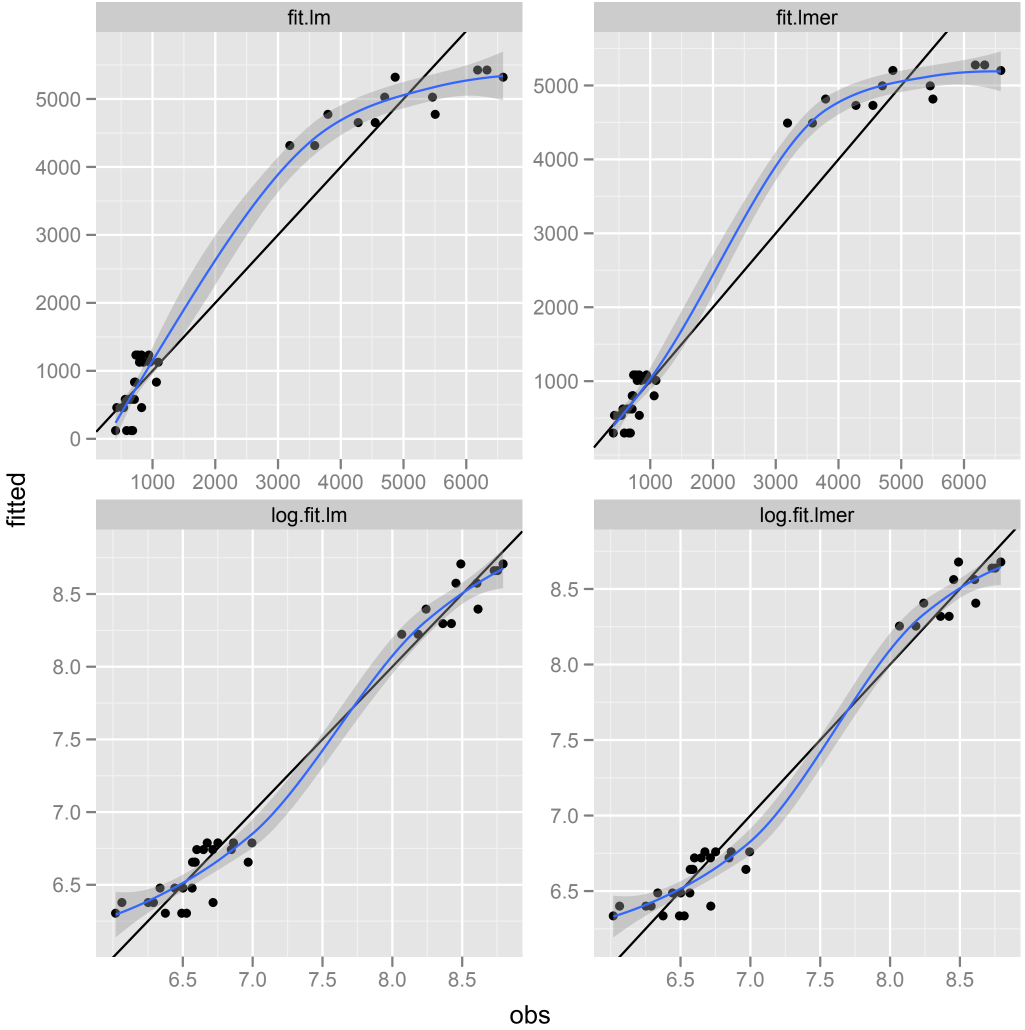


Figure 4.20: Observed versus fitted plots for the simple linear and log linear model and the linear and log linear random effects model. Smoothed lines through the points are provided for visual ease.

The negative log-likehood is proportional to

$$l(\boldsymbol{\beta}, \tau, \Sigma_1, \Sigma_2, \sigma_\epsilon, \sigma_\delta | \mathbf{y}, \log(\mathbf{t})) =$$

$$\sum_{i=1}^{18}\sum_{j\in\mathcal{P}_i}\sum_{r=1}^{10}\sum_{s=1}^{3}\left(\frac{(y_{ij}(r,s) - m_j f(x(r,s), \boldsymbol{\beta}, b_i))}{\sigma_\epsilon}\right)^2 + \left(\frac{\mathbf{b}^T\Sigma_1^{-1}\mathbf{b}}{\sigma_\epsilon^2}\right)$$

$$+\sum_{l=1}^{2}\sum_{j=1}^{6}\sum_{k=1}^{n_l}\left(\frac{\log(t_{ljk}) - (\log(m_j) + \log(\tau_l))}{\sigma_\delta}\right)^2 + \left(\frac{[\log(\mathbf{m})]^T\Sigma_2^{-1}[\log(\mathbf{m})]}{\sigma_\delta^2}\right) \tag{4.10}$$

where $\mathcal{P}_i$ = set of plates which contain experiment i.

4.5 Penalized least squares

We want to be able to fit all the data at once, such that the control values can help to normalize the data across all the plates. We attempt to minimize the following sums of squares [4.11] in order to estimate 29 parameters/random effects:

- 3 parameters for the logistic curve: upper asymptote (β_1), EC50 (β_2), slope (β_3)
- 2 parameters for the negative (τ_1) and positive control (τ_2), where the negative control is also the lower asymptote ($\tau_1 = \beta_0$)
- 18 random effects for the EC50, $b_i, i = 1, \ldots, 18$
- 6 random effects for the plates, in log scale ($\log(m_j)$: $j = 1, \ldots, 6$)

$$\sum_{i=1}^{18}\sum_{j\in\mathcal{P}_i}\sum_{r=1}^{10}\sum_{s=1}^{3}\left(\frac{(y_{ij}(r,s)-m_j f(x(r,s),\boldsymbol{\beta},b_i))}{\sigma_\epsilon}\right)^2+\left(\frac{\mathbf{b}^T\Sigma_1^{-1}\mathbf{b}}{\sigma_\epsilon^2}\right)$$

$$+\sum_{l=1}^{2}\sum_{j=1}^{6}\sum_{k=1}^{n_l}\left(\frac{\log(t_{ljk})-(\log(m_j)+\log(\tau_l))}{\sigma_\delta}\right)^2+\left(\frac{[\log(\mathbf{m})]^T\Sigma_2^{-1}[\log(\mathbf{m})]}{\sigma_\delta^2}\right) \qquad (4.11)$$

where $\mathcal{P}_i$ = set of plates which contain experiment i.

Estimation of variance components

Four variance components are involved in this model, where two (σ_b and σ_m) are the standard deviations of random effects and the other two (σ_ϵ and σ_δ) are the standard deviations of errors from the two parts of the model.

σ_b is calculated as the median absolute deviation of the deviations between individual EC50 estimates and the overall mean of the individual estimates. σ_ϵ is obtained from taking the square root of the mean of the individual estimates of the error variances from the R function nls(). σ_m and σ_δ are the estimated standard deviations of the plate random effect and residuals, respectively, from the linear mixed effects model 4.2 that was fitted by the R function lmer(). See Table 4.9 for all these values.

Initial values

Initial values are easily obtained from the same values that were used to estimate the variance components in the previous section. Initial values for b_i are the deviations of the individual EC50 estimates from their overall means. Initial values for $\log(\mathbf{m})$ are the predicted plate effects from the linear mixed effects model.

σ_b	1.36
σ_ϵ	493.85
σ_m	0.19
σ_δ	0.17

Table 4.9: Table of estimated variances

Since all the units can be fit individually quite well, the initial values for the fixed effects, β_1, β_2, and β_3, are the medians of the individual estimates for upper asymptote, EC50, and slope. The initial values for the controls, τ_1 and τ_2, are estimated in the linear mixed effects model as the fixed effects. Recall that the lower asympote parameter β_0 is set equal to the negative control τ_1.

Robustness

Because the complexity of this model involves balancing information of two different datasets together, including robustness in a common way through iteratively reweighted least squares does not work as well here. Instead of punishing true outliers, any discrepancies in the curve fitting from the observations gets continually weighted down, and then the model is content with producing curves that stray from the data points.

Instead, we take advantage of the fact that we already have independently fitted curves. We apply the Tukey bisquare function to the residuals from these curves to obtain weights (Figure 4.21) that are used in the full model fitting. Figure 4.22 of the unit 3 data shows an example of where fitting robustly is helpful. The two highest intensity values in this set are quite far from the rest of the points near the curve and have the weights set at 0 and 0.73, respectively, where they are not allowed to pull the fitted curve away from the other points.

4.5.1 Results

The parameter estimates that resulted are in Table 4.11. The lower and upper asymptote estimates are about a multiple of 11 away from each other, equivalent to 2.4 difference on the log scale, and follow a relationship between plates through the predicted plate effects on the log scale (Table 4.10). The predicted values in general are close to and of the same sign as the initial values, except for plate 3, where the pair is roughly the same distance in opposite directions from zero (Figure 4.23).

Fitting the mixed effects model on the dataset results in less extreme estimates for the upper asymptotes through utilizing the negative and positive controls on each plate (Table 4.11). Because the plots of the raw data show that those units with significantly higher individual upper asymptote estimates are such due to lack of data points in that area, the mixed effects model can take into account the information

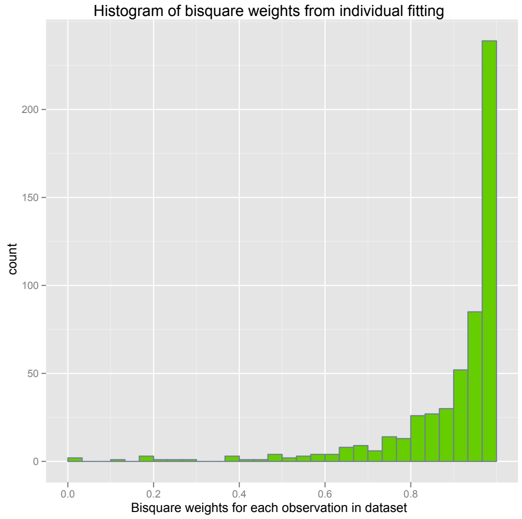
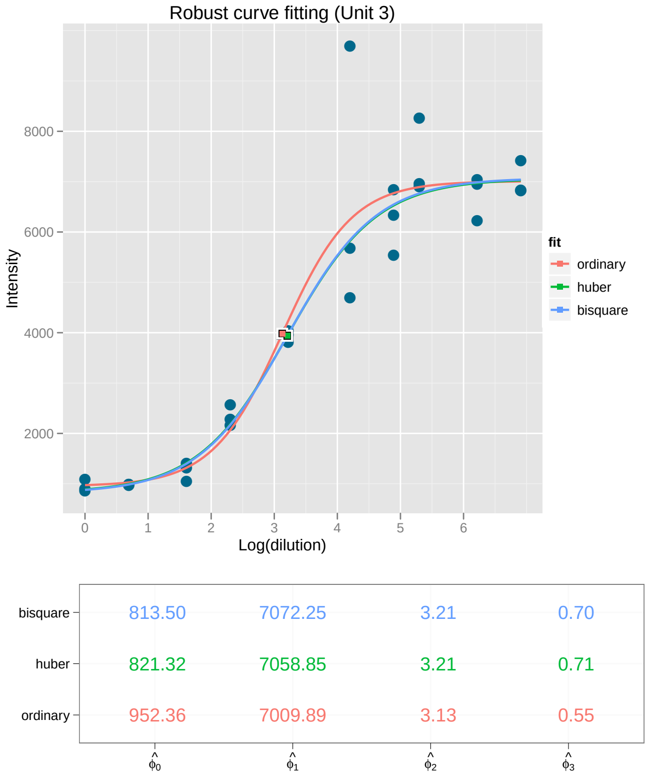


Figure 4.21: Histogram of weights generated from individual fitting of the data through the Tukey bisquare function



	$\hat{\phi}_0$	$\hat{\phi}_1$	$\hat{\phi}_2$	$\hat{\phi}_3$
bisquare	813.50	7072.25	3.21	0.70
huber	821.32	7058.85	3.21	0.71
ordinary	952.36	7009.89	3.13	0.55

Figure 4.22: Plot of robust fitting of points. Two robust methods, Huber weights and Tukey biweights, are compared with the ordinary fit. The highest point in the plot is given a zero weight.

Plate	Predicted	Initial
1	0.04	0.12
2	-0.13	-0.21
3	0.17	-0.20
4	-0.10	-0.01
5	-0.25	-0.06
6	0.15	0.36

Table 4.10: Table of predicted and initial plate effects from mixed effects model.

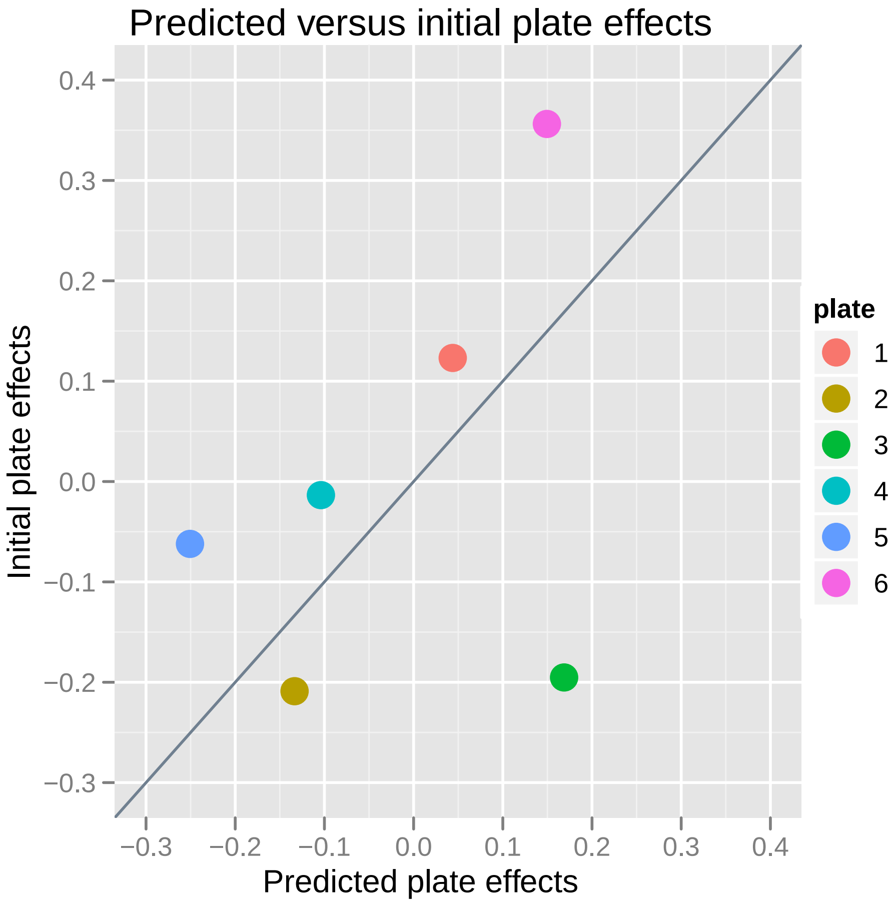


Figure 4.23: Plot of the predicted plate effects from the mixed effects model versus the initial values of plate effects used. Gray line with intercept 0 and slope 1 is included for reference.

from other units on the same plate and effectively adjust the estimates according to the relationships of the controls across the plates. Figure 4.29 shows that the mixed effects estimates are similar to robust means of the individual estimates. It is also easy to see that there exist the outliers in the individual estimates, which ideally should have values closer to others within each plate. Even from the plots of the raw data that included the control data (Figures 4.7, 4.5), we can tell that the maximum values of the dilution series observations corresponded to the magnitudes of the positive controls across all the plates.

unit	plate	ϕ_0 (lower asym)	ϕ_1 (upper asym)	ϕ_2 (EC50)	ϕ_3 (slope)
1	1	756.21	8339.94	4.01	1.10
2	1	756.21	8339.94	4.12	1.10
3	1	756.21	8339.94	3.62	1.10
4	2	633.33	6984.69	3.72	1.10
5	2	633.33	6984.69	4.46	1.10
6	2	633.33	6984.69	4.59	1.10
7	3	856.80	9449.23	4.38	1.10
8	3	856.80	9449.23	4.42	1.10
9	3	856.80	9449.23	4.13	1.10
10	4	652.34	7194.33	4.11	1.10
11	4	652.34	7194.33	4.52	1.10
12	4	652.34	7194.33	3.76	1.10
13	5	563.28	6212.17	4.28	1.10
14	5	563.28	6212.17	4.63	1.10
15	5	563.28	6212.17	4.44	1.10
16	6	840.42	9268.58	4.84	1.10
17	6	840.42	9268.58	4.83	1.10
18	6	840.42	9268.58	4.93	1.10

Table 4.11: Table of estimates from fitting the mixed model with penalized least squares.

The residuals from the mixed model are included in Figure 4.26. While at first glance, there seem to be units that do not fit well under this model. However, with comparison of the residuals from the individual fits (Figure 4.27), we can tell for several units, for instance, unit 3, 7, and 10, that it is not a matter of difference between the two models but of the logistic curve being unable to address the points that are either outliers or perhaps non-monotonic, although that may be very unlikely given that these jumps only occur within one dilution.

EC50 estimates: mixed effects versus individual fits

We now compare the estimates of EC50 we obtain from the mixed effects model with the individual estimates with Figure 4.30. There is considerable shrinkage in the mixed effects estimates; the figure clearly shows a much larger range of values in

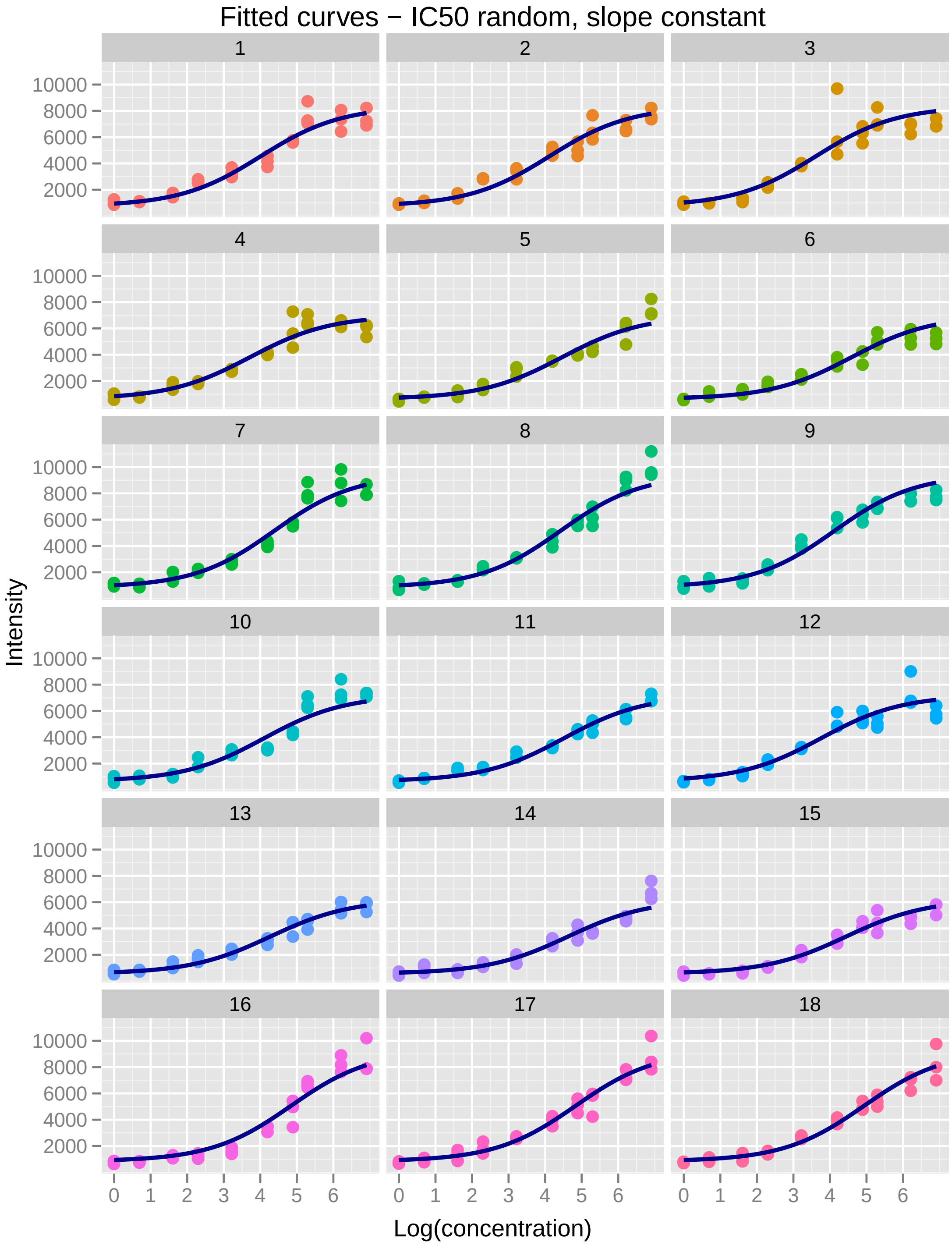


Figure 4.24: Fitted curves from mixed model where the asymptotes are adjusted by the control data and the slope is held constant.

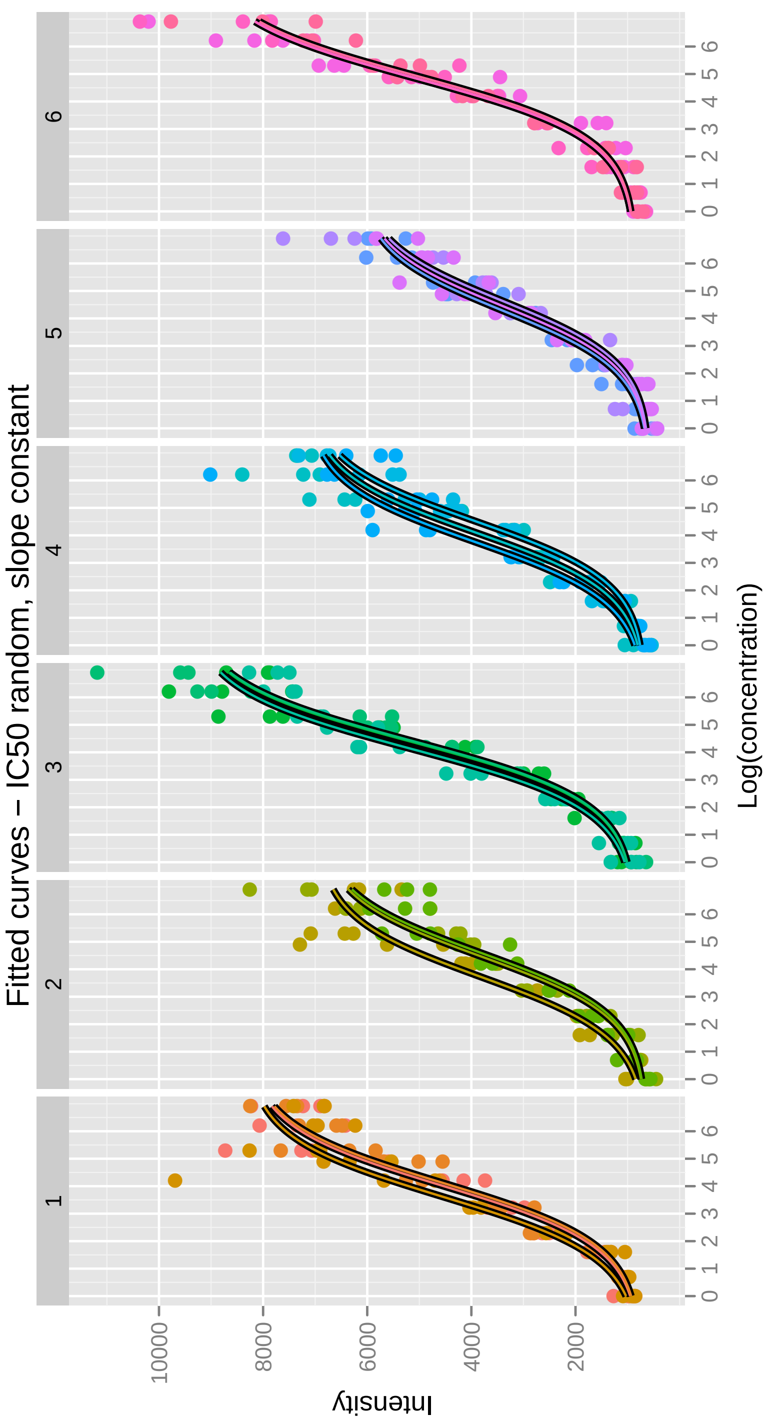


Figure 4.25: Fitted curves from mixed model where the asymptotes are adjusted by the control data and the slope is held constant.

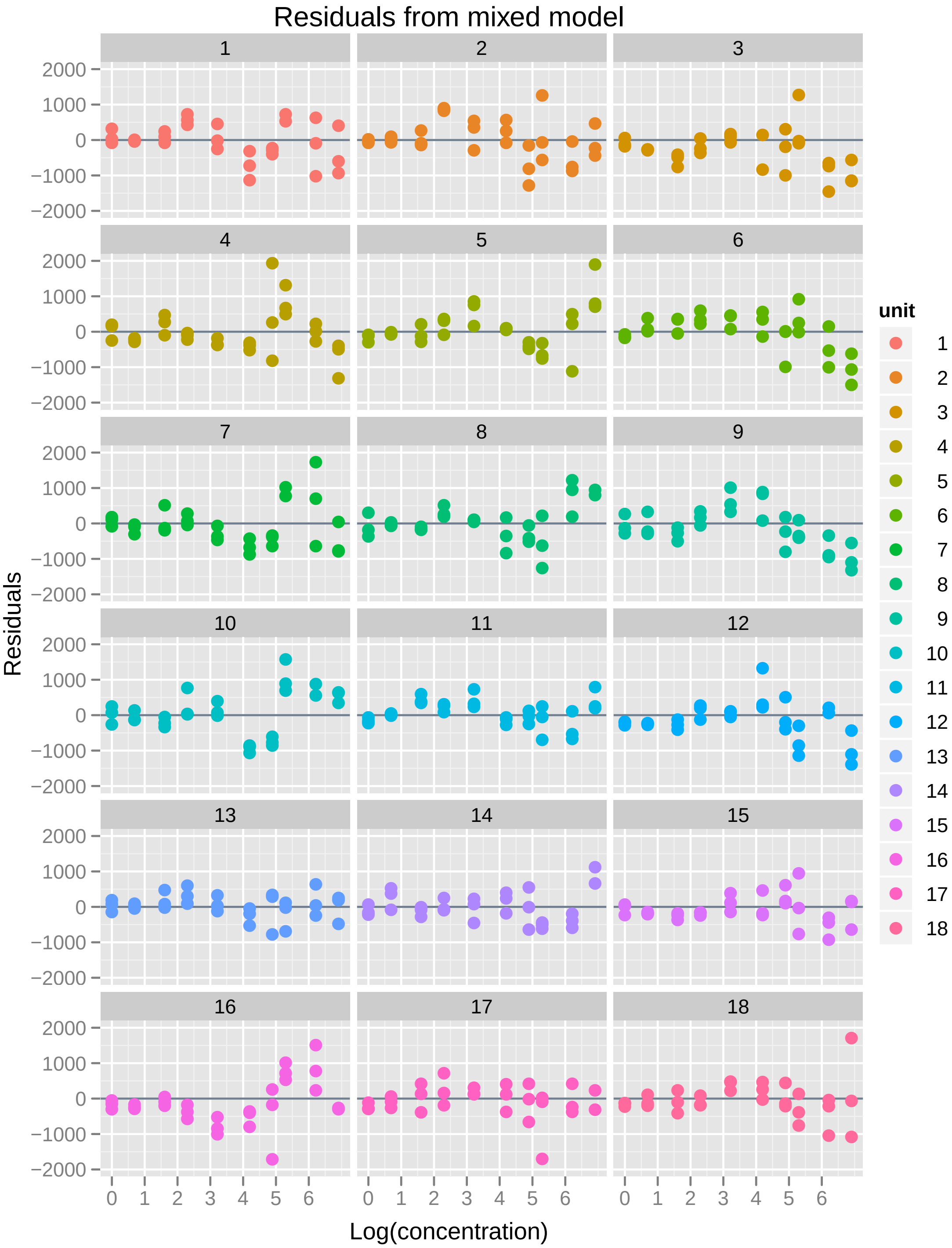


Figure 4.26: Residuals from mixed model.

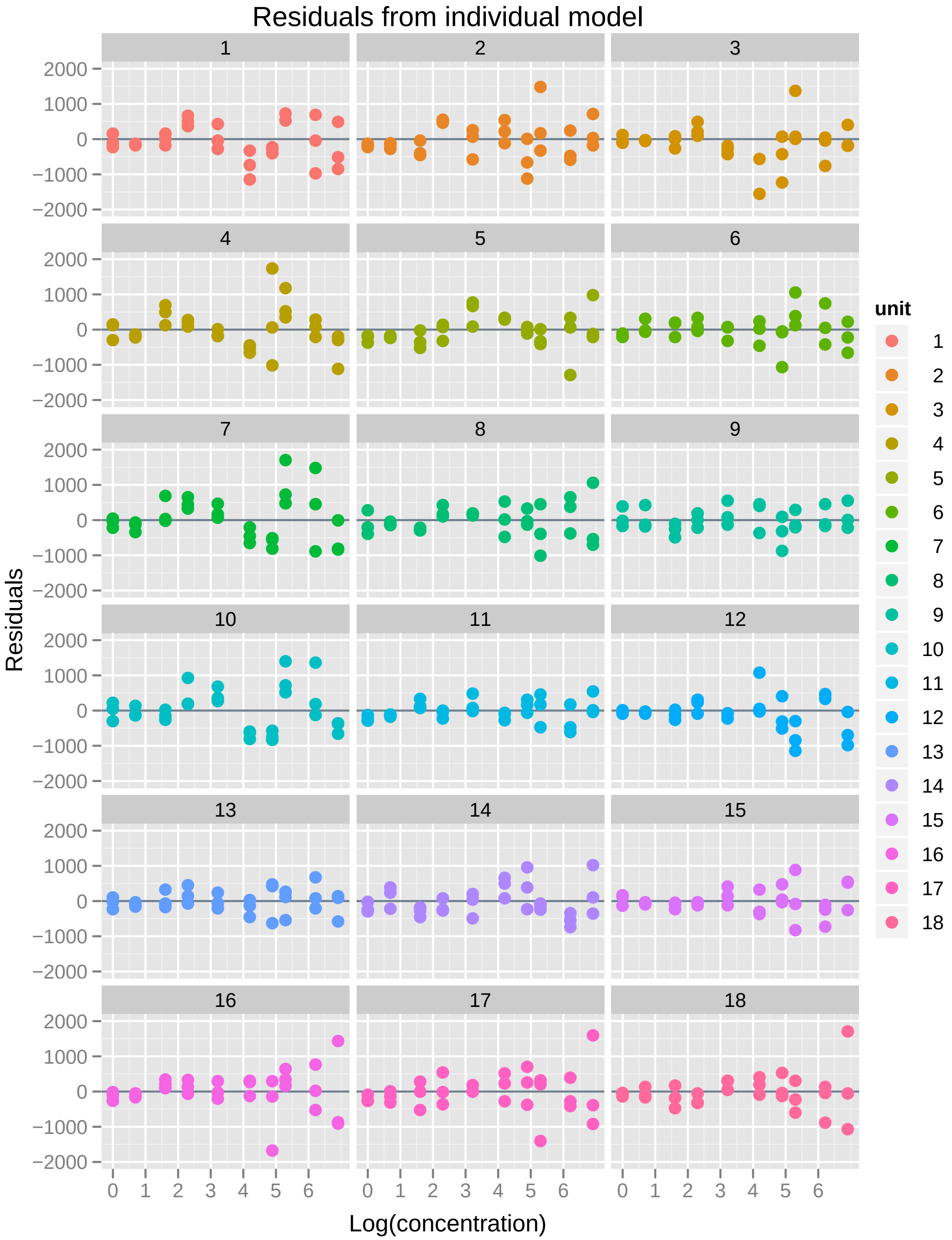


Figure 4.27: Residuals from individual model.

the individual estimates than the mixed effects estimates. The high extreme values of the individual estimates are from the units with the outlying upper asymptote values. Because there is insufficient evidence of where the true upper asymptotes should lie, artificially higher values result in larger EC50 values. The mixed effects model effectively shares the information across units within the same plate to prevent such extreme values.

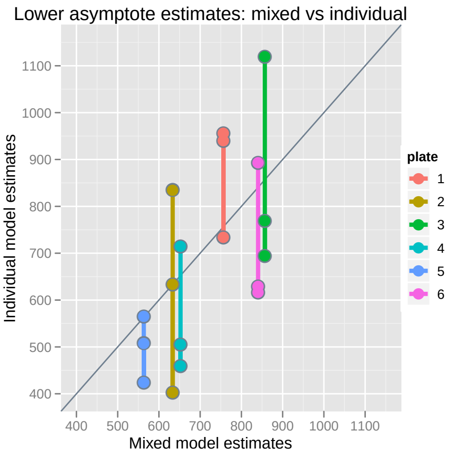


Figure 4.28: Mixed versus individual estimates of lower asymptotes

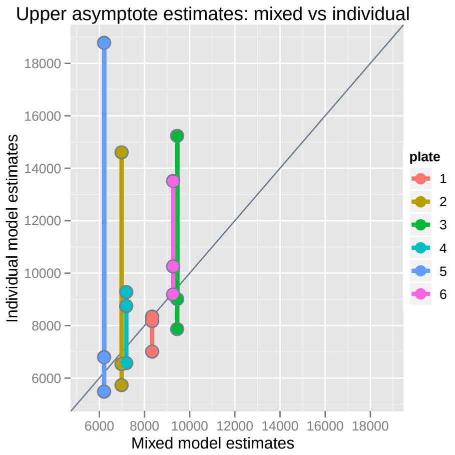


Figure 4.29: Mixed versus individual estimates of upper asymptotes

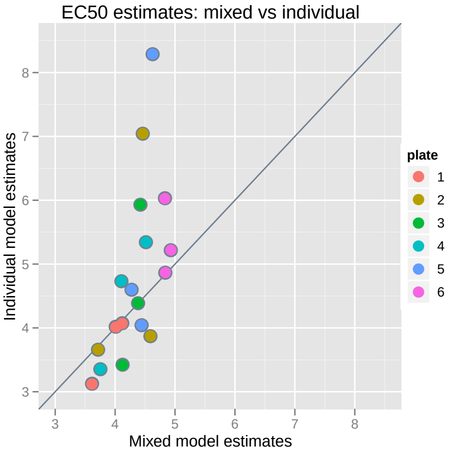


Figure 4.30: Mixed versus individual estimates of EC50

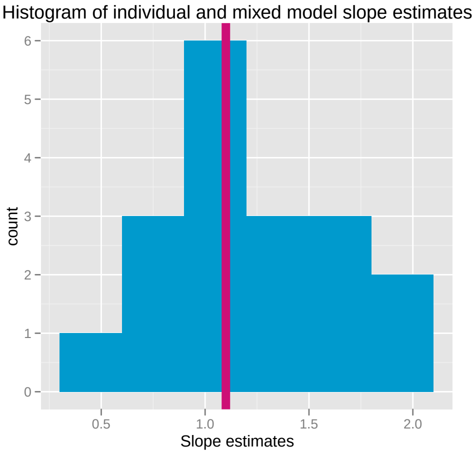


Figure 4.31: Histogram of individual estimates of slope, the mixed effects estimate is included as the pink line.

Normalization of data

Once the plate effects have been predicted from the mixed effects model, we can see how this effectively adjusts the raw observations to be relatively comparable across all the plates. Figure 4.32 is of boxplots of the raw data grouped by the plates. As before, notice that there is clear differences between the plates that is mirrored by the control values. Figure 4.33 shows the boxplots obtained after scaling with the prediced plate effects. The medians of the values by plate are adjusted to be much more similar now; however we do notice that the sixth plate has been adjusted in the incorrect direction. We find that this is because the plate's relationship between the dilution series data and control data is not as in line with the relationships of the other five plates. First of all, the ratio between the medians of the positive and negative controls (7.01) for the sixth plate is fine.

Plate	1	2	3	4	5	6
Ratio	6.32	7.19	7.84	8.34	5.43	7.01

However, when we look at the ratios of the medians of the positive controls and the dilution series data, we see that the value for the sixth plate is much higher than the rest, resulting in the discrepancy that we see in the scaled boxplots.

Plate	1	2	3	4	5	6
Ratio	1.43	1.51	1.58	1.39	1.32	1.74

Standard errors of EC50

The standard errors of the EC50 estimates can be estimated from the covariance matrix of the fixed and random effects given the data [2].

$$\mathbb{E}\left\{ \begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\mathbf{b}} - \mathbf{b} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \widehat{\mathbf{b}} - \mathbf{b} \end{bmatrix}^T \right\} = \begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \Sigma_1^{-1} \end{bmatrix}^{-1} \sigma_\epsilon^2 \tag{4.12}$$

where we plug in the estimates of the matrices X and Z defined as

$$\widehat{X}_k = \left. \frac{\partial f_k}{\partial \boldsymbol{\beta}^T} \right|_{\hat{\boldsymbol{\beta}}, \hat{b_k}} \tag{4.13}$$

$$\widehat{Z}_k = \left. \frac{\partial f_k}{\partial b_k^T} \right|_{\hat{\boldsymbol{\beta}}, \hat{b_k}} \tag{4.14}$$

In our case, the matrices are of dimensions $n \times 4$ and $n \times 18$, for the 4 fixed effects and 18 random effects of the dilution series data. At each observation $k = 1, \ldots, n$,

[2] Note that this covariance matrix does not take into account the estimation error of the covariance components of the model. There is no commonly agreed upon best analytical method to include this. The recommended method by Douglas Bates is Markov chain Monte Carlo sampling of the posterior distribution of the parameters (in a Bayesian model, the fixed and random effects are considered parameters).

the partial derivative of the four-parameter logistic curve f with respect to the EC50 parameter is evaluated at the observation's specific set of estimated $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{b}}$ values.

Table 4.12 includes the standard errors of the EC50 estimates for all 18 units, along with the corresponding standard errors for the EC50 estimates obtained from the individual fitting. Figure 4.34 shows that all errors involved in the mixed model are much lower than the individual errors, especially for units where the individual fitting could not estimate the EC50 as well as it could with only unit data. Units 5 and 14 have errors of 1.52 and 2.82, well above the median individual error 0.25. This can be explained by these units' lack of information to estimate the upper asymptote. Since the EC50 is determined as the concentration at halfway between the lower and upper asymptote, unstable estimates of the upper asymptotes can lead to uncertain EC50 estimates, reflected in estimates of standard error. The mixed effects model uses the control information and other units' data to better estimate all four parameters, so that the standard errors are the same or smaller than individual errors.

Unit	Plate	Mixed standard errors	Individual standard errors
1	1	0.10	0.23
2	1	0.10	0.27
3	1	0.11	0.17
4	2	0.13	0.17
5	2	0.12	1.52
6	2	0.12	0.20
7	3	0.09	0.14
8	3	0.09	0.53
9	3	0.09	0.10
10	4	0.12	0.30
11	4	0.12	0.42
12	4	0.12	0.23
13	5	0.14	0.27
14	5	0.14	2.82
15	5	0.14	0.15
16	6	0.09	0.12
17	6	0.09	0.81
18	6	0.09	0.42

Table 4.12: Standard errors from the mixed model and individual model

4.6 Simulation

To see how well the estimation procedure is working, we simulate a similar dataset to test it on. We start first by assuming that all four parameters of the logistic curve is the same for all experiments, which we can refer to as units since there are no replicates. For each unit i,

$$y_i = \beta_0 + \frac{\beta_1 - \beta_0}{1 + \exp\left(-\frac{x-\beta_2}{\beta_3}\right)} \tag{4.15}$$

Then recall that, depending on which plate j the unit belongs to, the lower and upper asymptotes are scaled by a multiplicative factor m_j. In addition, we allow ϕ_2 to vary by experiment.

$$y_{ij} = \phi_{0,ij} + \frac{\phi_{1,ij} - \phi_{0,ij}}{1 + \exp\left(-\frac{x-\phi_{2,ij}}{\phi_{3,ij}}\right)} \tag{4.16}$$

where

$$\phi_{0,ij} = \beta_0 m_j \tag{4.17}$$

$$\phi_{1,ij} = \beta_1 m_j \tag{4.18}$$

$$\phi_{2,ij} = \beta_2 + b_i \tag{4.19}$$

$$\phi_{3,ij} = \beta_3 \tag{4.20}$$

4.6.1 Simulating lower asymptotes

The lower asymptotes are simulated first. We set $\log \beta_0$ equal to 6.53, the log of the mean of the lower asymptote estimates from individual nonlinear least squares. We generate plate effects by sampling 6 numbers from a normal distribution with mean 0 and standard deviation $\sigma_m^2 = 0.19$ (Table 4.13).

Plate	log(m)
1	0.14
2	-0.23
3	-0.22
4	-0.04
5	-0.09
6	0.37

Table 4.13: Table of simulated plate effects with standard deviation $\sigma_m^2 = 0.19$.

By adding these plate effects and additional error that is normally distributed with mean 0 and standard deviation $\sigma_\delta = 0.17$ to $\log(\beta_0)$, we can simulate a set of lower asymptotes $\phi_{0,ij}$ (Table 4.14).

4.6.2 Simulating upper asymptotes

We assume the upper asymptotes are a constant multiple of their corresponding lower asymptote values. To set a reasonable number for this, we look at the mean and standard deviation of the ratios of the upper to lower asymptote estimates: 2.6 and 0.23. Since we assume that there are a few extreme values in the asymptote estimates, by looking at the plots of the raw values, we adjust the value to 2.45; otherwise we obtain upper asymptote values that are much higher than what we see in our data. We then simulate ratios with this mean and standard deviation.

Plate	1	2	3	4	5	6
Upper / lower asymptote nls	2.19	2.58	2.55	2.67	2.83	2.77

These simulated ratios are then added to the simulated lower asymptotes to obtain the simulated upper asymptotes.

4.6.3 Simulating EC50s

The empirical estimate of the standard deviation of EC50 values from individual fitting is $\sigma_b = 1.71$. We simulate 18 different EC50 values from a normal distribution with mean 0 and standard deviation σ_b.

Unit	Plate	$\log(\phi_0)$	$\log(\phi_1)$	ϕ_0	ϕ_1	ϕ_2	ϕ_3
1	1	6.66	8.60	776.77	5405.35	4.35	1.10
2	1	6.66	8.60	776.77	5405.35	3.84	1.10
3	1	6.66	8.60	776.77	5405.35	4.40	1.10
4	2	6.29	8.23	536.85	3735.78	3.84	1.10
5	2	6.29	8.23	536.85	3735.78	5.36	1.10
6	2	6.29	8.23	536.85	3735.78	5.46	1.10
7	3	6.29	8.23	541.18	3765.94	5.35	1.10
8	3	6.29	8.23	541.18	3765.94	5.30	1.10
9	3	6.29	8.23	541.18	3765.94	4.82	1.10
10	4	6.47	8.41	647.01	4502.39	2.49	1.10
11	4	6.47	8.41	647.01	4502.39	6.92	1.10
12	4	6.47	8.41	647.01	4502.39	2.28	1.10
13	5	6.42	8.36	616.07	4287.07	4.96	1.10
14	5	6.42	8.36	616.07	4287.07	2.41	1.10
15	5	6.42	8.36	616.07	4287.07	4.79	1.10
16	6	6.88	8.82	974.57	6781.80	4.95	1.10
17	6	6.88	8.82	974.57	6781.80	5.17	1.10
18	6	6.88	8.82	974.57	6781.80	6.01	1.10

Table 4.14: Table of simulated parameters

4.6.4 Simulating control values

We simulate control values in the same way as the asymptote values; we first assign the same lower asymptote value to the negative control: $\log(\tau_1) = 6.5$. Then we add the same simulated plate effects and a random error that we sample from the normal distribution with mean 0 and standard deviation that we estimate from the real control data, 0.19. (Here we use the within sum of squares divided by 6 * (4 - 1), for 6 plates and 4 replicates.)

The positive controls are also a multiple from the negative controls, where it is a bit less than for the upper asymptote, as we can see from plots. The mean of the log ratios of the positive to negative controls by plate (included below) turn out to be 1.94, so we will add that to the logged negative control value to generate the simulated positive control value $\log(\tau_2)$.

Plate	1	2	3	4	5	6
Log(Positive / negative)	1.82	1.97	2.04	2.06	1.77	1.85

We include all these values in Table 4.15 as reference; we'll be able to visualize them better in plots.

4.6.5 Visualizing the simulated data

We provide plots of the simulated data, Figure 4.35. We can see then that the simulated data does not stray too far from the original data.

4.6.6 Results

Results indicate that the fitting method is working sufficiently well. Figure 4.40 provides a quick visual check as to whether all the fitted curves are reasonably applied to the observations.

Comparison with individual estimates

Individual fits of the curves are conducted first, where a small proportion of the units are unable to be fitted. Various adjustments to the sampling parameters also resulted in a similar situation. Since these simulated sets look relatively similar to the real dataset, we can hopefully proceed without this affecting the rest of the analysis. The parameter estimates are listed in Table 4.16 while the fitted curves are displayed in Figure 4.36.

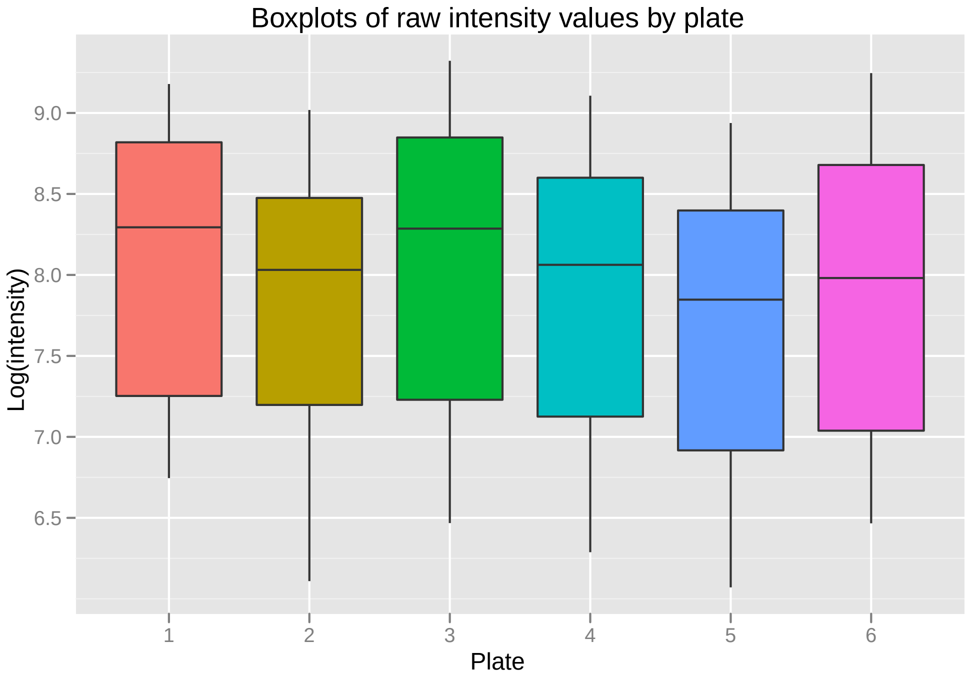


Figure 4.32: Boxplot of the logged intensity values by plate

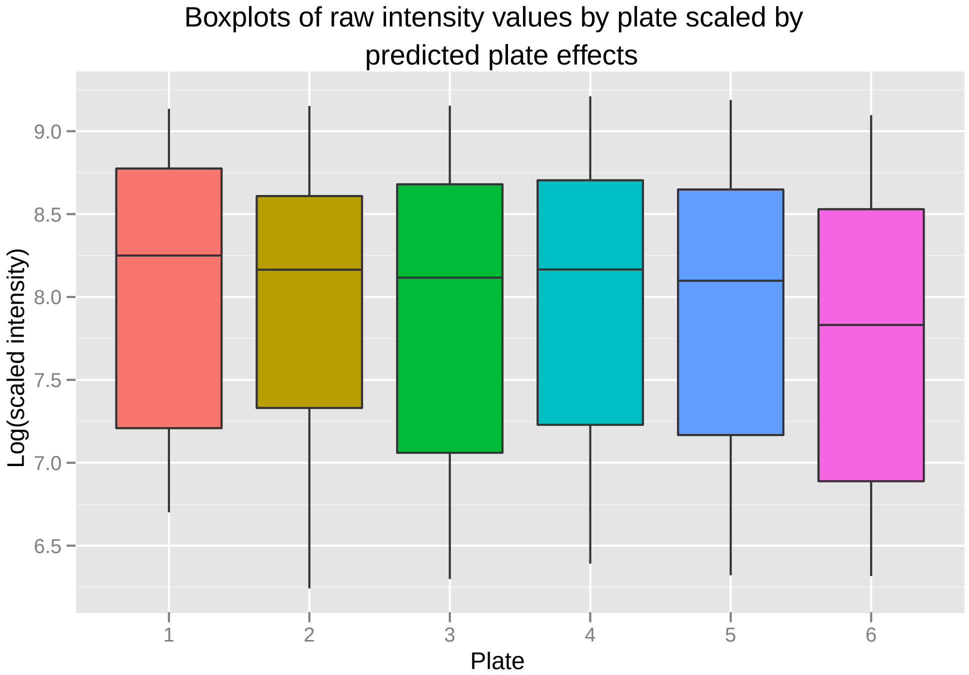


Figure 4.33: Boxplot of the logged intensity values by experiment after scaling with predicted plate effects from the mixed model.

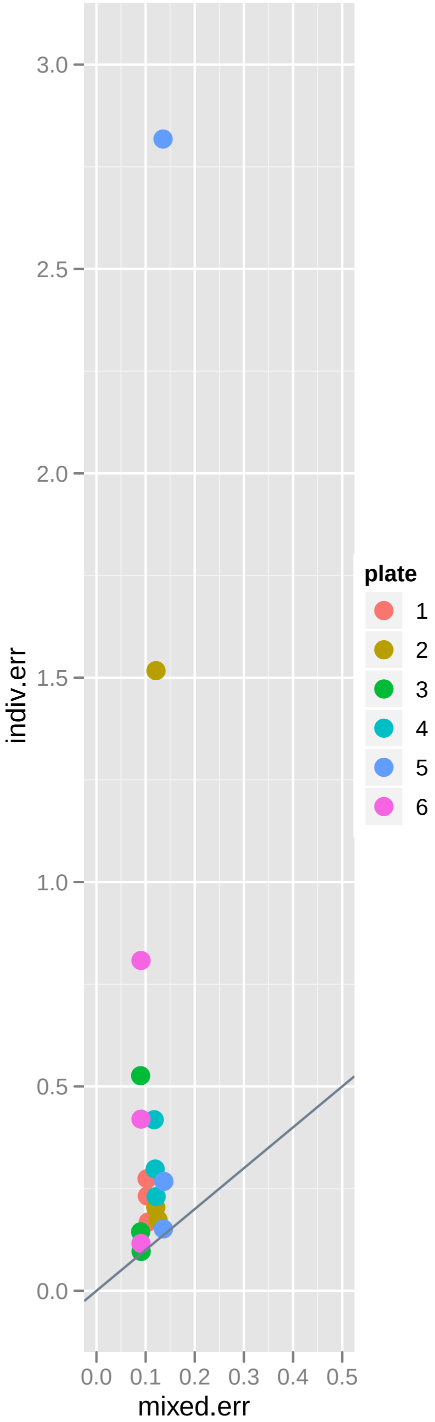


Figure 4.34: Plot of standard errors of IC50 from mixed versus individual models.

Plate	Type of control	Log(intensity)	Intensity
1	neg	6.45	632.67
1	neg	6.70	814.28
1	neg	6.84	934.02
1	neg	6.26	521.33
2	neg	6.36	577.47
2	neg	6.37	585.08
2	neg	6.19	486.87
2	neg	6.19	489.21
3	neg	6.20	491.66
3	neg	6.14	465.19
3	neg	6.21	499.01
3	neg	6.12	456.70
4	neg	6.34	567.02
4	neg	6.48	654.14
4	neg	6.64	761.64
4	neg	6.45	634.99
5	neg	6.34	564.81
5	neg	6.27	527.66
5	neg	6.28	534.34
5	neg	6.83	928.95
6	neg	6.90	997.04
6	neg	6.80	896.57
6	neg	6.81	904.25
6	neg	6.96	1053.77
1	pos	8.82	6769.32
1	pos	8.68	5893.03
2	pos	8.19	3608.59
2	pos	8.27	3910.64
3	pos	8.45	4693.83
3	pos	8.53	5077.66
4	pos	8.39	4421.33
4	pos	8.55	5152.02
5	pos	8.51	4963.88
5	pos	8.31	4046.94
6	pos	9.11	9052.19
6	pos	8.70	5982.30

Table 4.15: Table of simulated control values

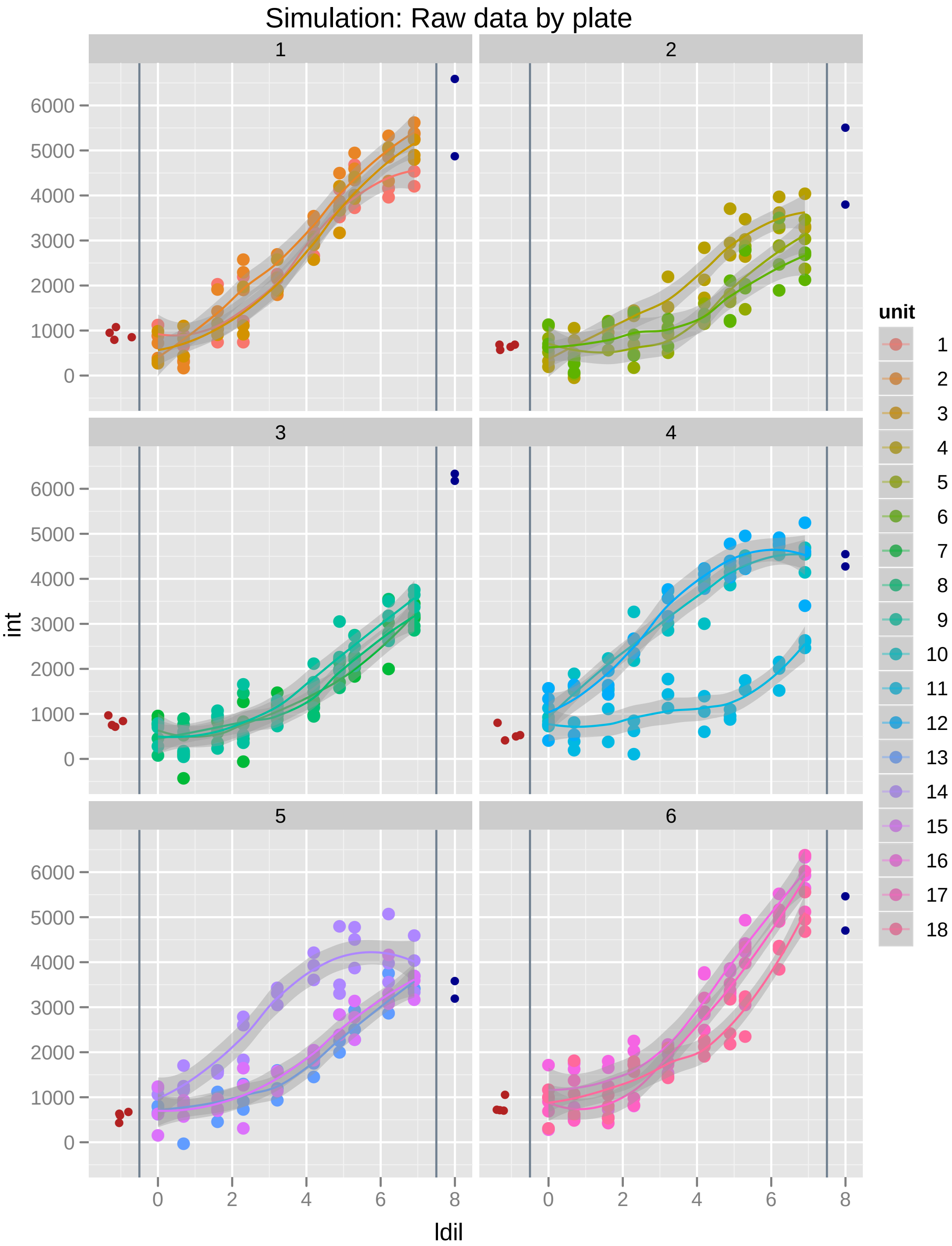


Figure 4.35: Plots of the simulated dataset by plate, including simulated negative controls (left, red) and positive controls (right, blue).

Plate	Unit	ϕ_0 (lower asym)	ϕ_1 (upper asym)	ϕ_2 (EC50)	ϕ_3 (slope)
1	1	840.21	4650.19	3.88	0.88
1	2	NA	NA	NA	NA
1	3	451.01	5727.20	4.29	1.25
2	4	159.53	4141.11	3.77	1.52
2	5	604.83	3128.86	4.87	0.58
2	6	746.27	2734.23	4.84	0.69
3	7	354.84	6075.12	6.93	1.81
3	8	645.70	3194.61	4.86	0.63
3	9	365.02	4252.58	4.97	1.31
4	10	NA	NA	NA	NA
4	11	NA	NA	NA	NA
4	12	921.04	4629.32	2.58	0.82
5	13	753.15	3958.80	4.98	1.03
5	14	965.50	4163.59	2.51	0.72
5	15	661.01	3942.20	4.60	1.09
6	16	1081.59	6842.14	4.93	1.17
6	17	518.27	7815.27	5.49	1.47
6	18	NA	NA	NA	NA

Table 4.16: Table of individual estimates on simulated data

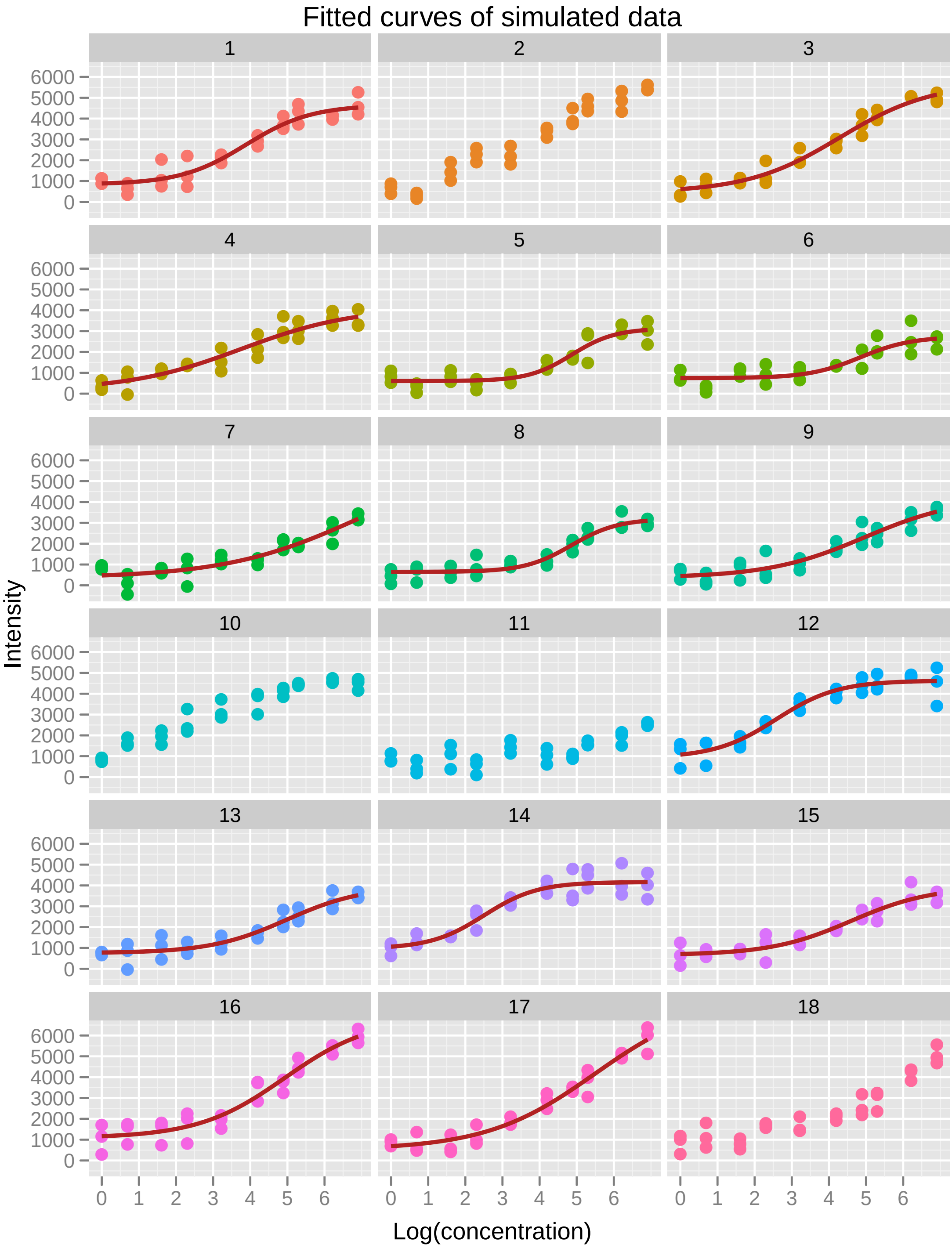


Figure 4.36: Fitted curves of individual model on simulated data.

Even though the true value of the slope parameter is 1.1, individual nonlinear least squares picks up on the slightest patterns in the data, and the resulting range of these estimates is quite large. Figure 4.37 is a histogram of the estimates, whose standard deviation is 0.37. We can compare this histogram of the simulated results with the histogram of the individual estimates of slope for the real dataset (Figure 4.31) to see that the ranges are quite similar. This gives support to the model assumption that the slopes be constrained equal for the real data. Because of the separate involvement of all the data units and the noise, it is quite typical for individual estimation to result in values more extreme than the true values; it does the best it can to fit as close as it can to the data points it is allotted. But with the mixed effects model, the overall patterns can be found from using all the data simultaneously, and the estimates are less likely to be affected by the noise in the observations.

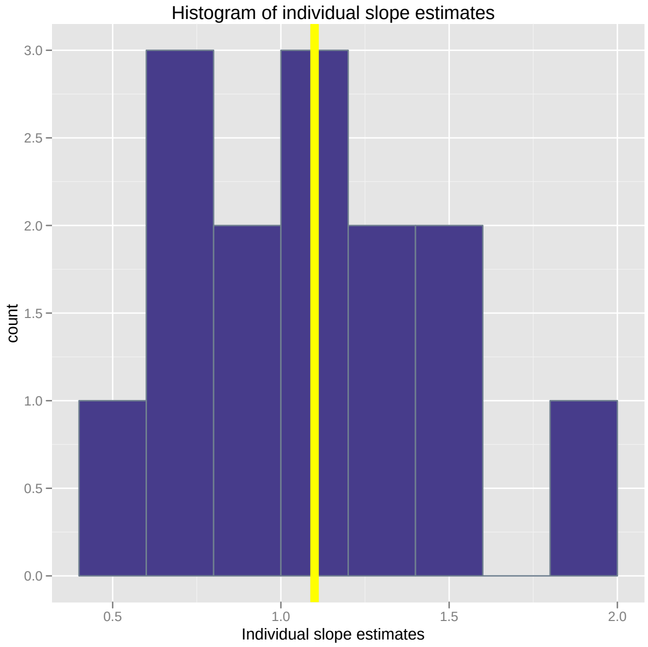


Figure 4.37: Histogram of the individual slope estimates

The estimated covariance components compare well with the true values in Table 4.17. This shows that the methods to estimate these covariance parameters are doing quite reasonably. Future analysis can help ascertain under what conditions these simple methods may not do as well.

Parameter	True	Estimated
σ_ϵ	400.00	394.41
σ_b	1.35	1.13
σ_m	0.19	0.22
σ_δ	0.17	0.17

Table 4.17: Table of true and estimated covariance components of simulation

Figure 4.38 is composed of three plots comparing the estimates for the lower and upper asymptotes and EC50 with their true values. Keeping in mind that the scales are different, we notice that both sets of asymptote estimates correspond well with their true values, with errors maxed at around 100, which is small relative to the magnitudes of the upper asymptote values, but not to the magnitudes of the lower asymptotes in the high hundreds. However, this does not seem to have affected the most important parameter, the EC50, where the mean and standard deviation of the errors is 0.02 and 0.23, respectively, resulting in unbiased and close estimates. Compared to the mean and standard deviation of the errors for the individual estimates at 0.004 and 0.54, the mixed effects model does much better at estimating the EC50s in this simulation, balancing between a mean estimate and the more extreme individual estimates.

The predicted plate effects are plotted against the true plate effects in Figure 4.39. Although the predicted effects do not follow the gray reference line with intercept 0 and slope 1, they do fall on a straight line, which means that the relationships between the plates are correctly predicted, but the scale factors are off by a constant value. This actually does not affect the end result as this discrepancy is just due to the differences in the control estimates from the true values.

	True	Estimated
Negative (τ_1)	665.1	578.6
Positive (τ_2)	4628.6	5020.6

Figure 4.40 of the fitted curves and Figure 4.41 of the residuals from these curves on the simulated dataset show that the fitting methods work relatively well. Despite the lack of individual fits on some of the units, fitting the mixed effects model resulted in reasonable fitted curves for all units in the dataset.

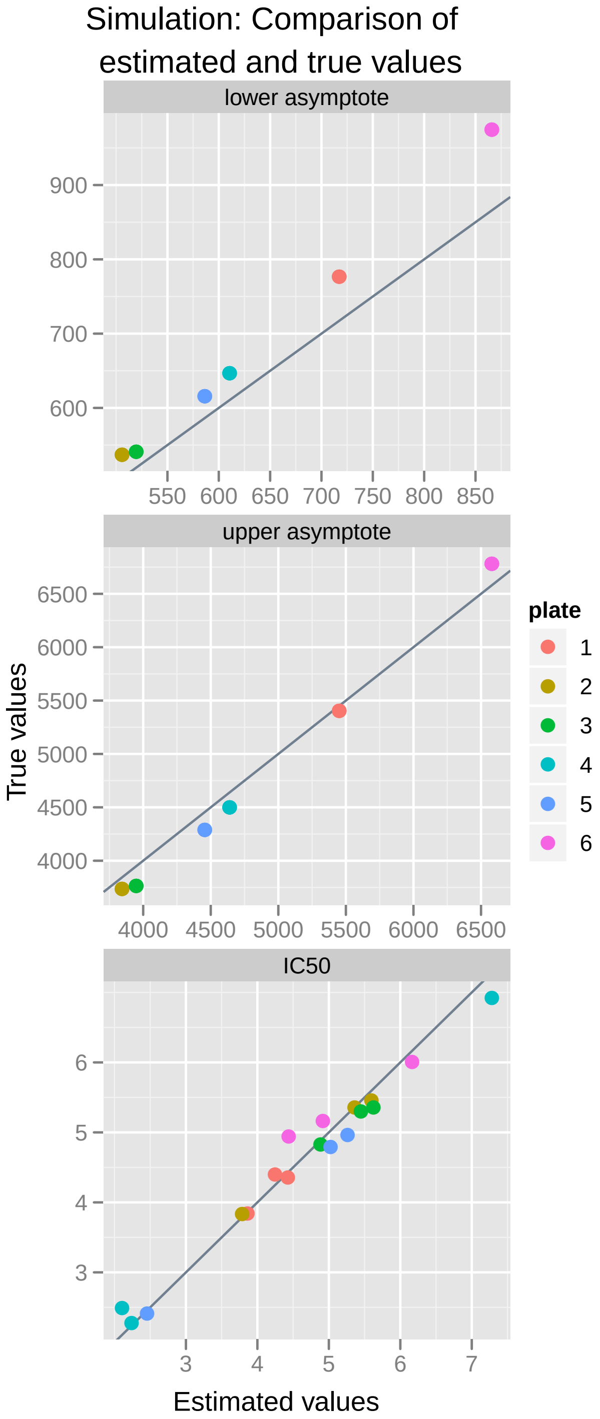


Figure 4.38: Plots of estimated versus true values of parameters for simulated data.

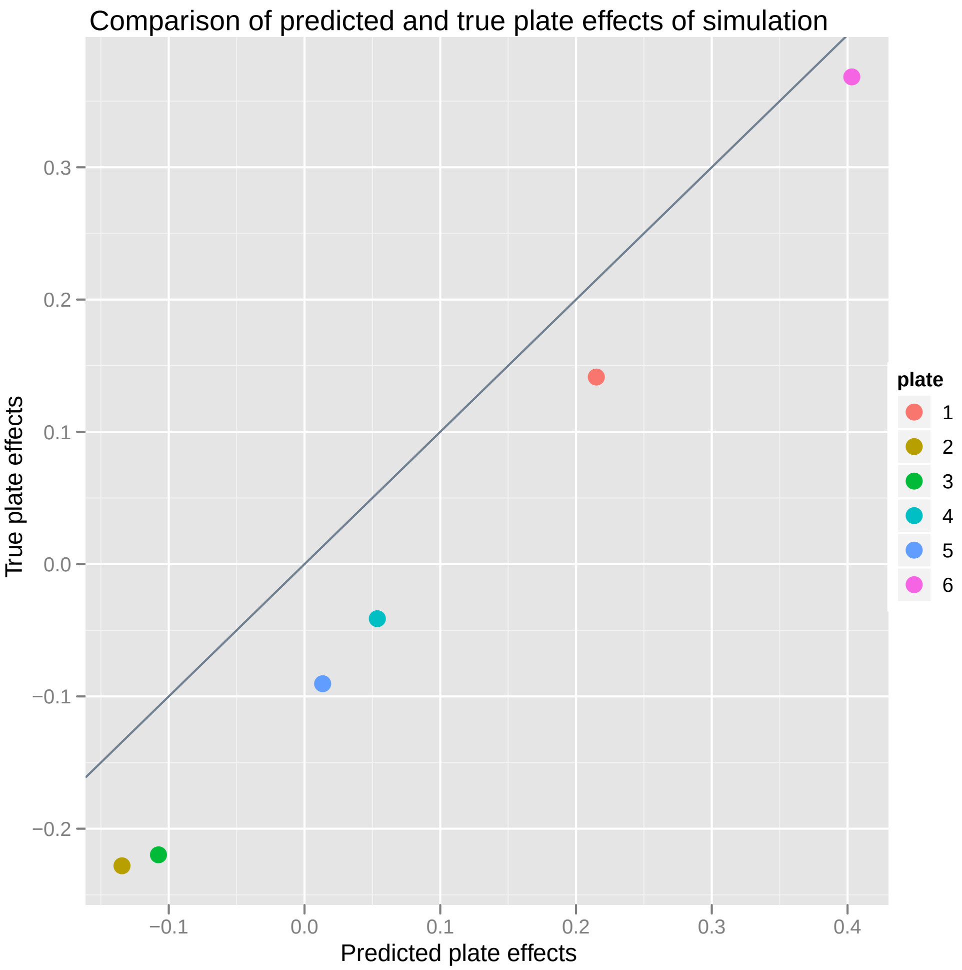


Figure 4.39: Plot of predicted plate effects from mixed model versus the true plate effects of the simulated data.

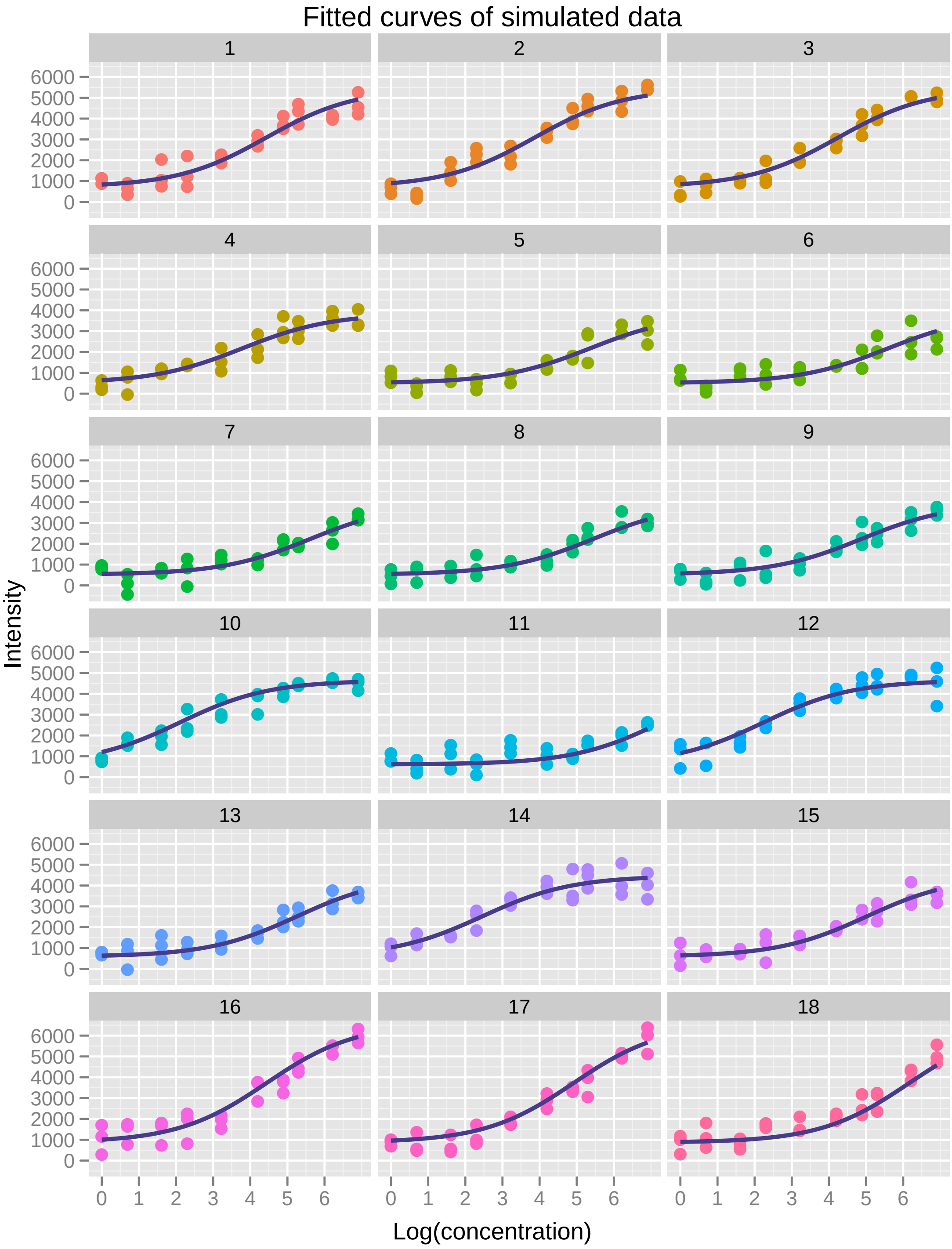


Figure 4.40: Fitted curves of mixed model on simulated data.

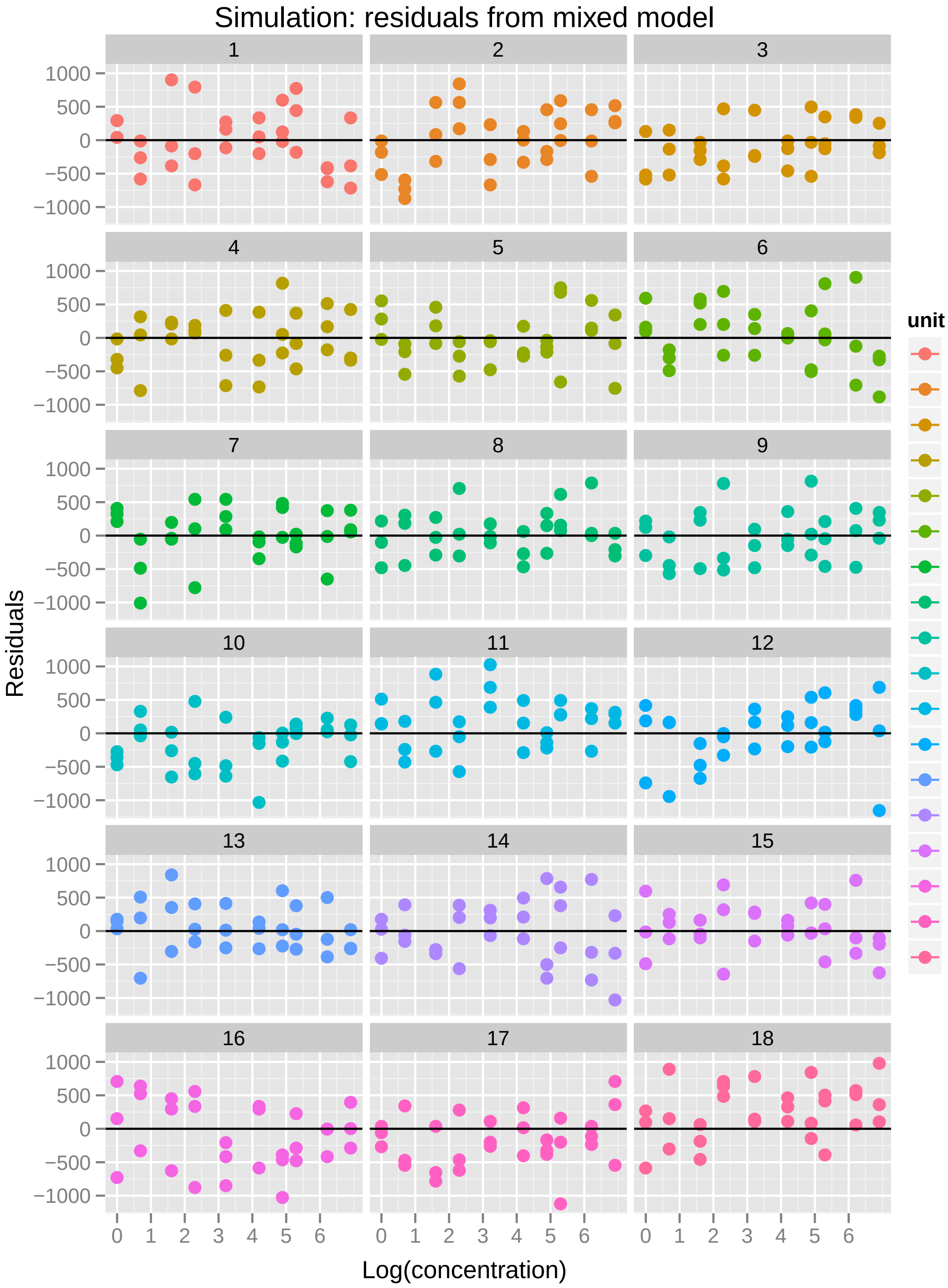


Figure 4.41: Residuals from the mixed model on simulated data.

4.7 Discussion

We saw from Figure 4.33 that the model is able to adjust the raw values such that their observations are not as affected by the plate effects as they were originally (Figure 4.32). In general, this depends on moderately strong relationships between the control and experimental values within and between each plate. This then allows shrinkage estimates of the EC50 that are more reasonable and less dependent on uncertain asymptote estimates than estimates of the EC50 from individual fitting of the units.

The fitting of these types of models rely on good initial values. Using random sampled values for the initial random effects, the early runs of this model took as many as 600 iterations in optim(), and sometimes with nonconvergence. Further work is needed to investigate more practical implementations of the fitting method.

Testing on larger and different data sets would be very helpful here; we used a relatively small dataset with only 540 observations on 18 experiments on only 6 plates with no replicates by experiment. Having experiments replicated across multiple plates will help with the construction of better and better models. In particular, the assumption that the slopes are the same may be relaxed. With smaller datasets, it may not be possible to fit the ideal model.

Another important issue in the fitting of any models on this dataset is the odd occurrence of significantly higher upper asymptotes of the second experiments on each plate. Since the data for the upper asymptote are on the other two experiments for each plate show much similarity and follow the relationship with the control data, we cannot simply assume that each set of upper asymptotes should vary independently. But because of this, fitting reasonable curves on all data involve a balance between extreme values of the predicted random effects to accommodate these experiments and loosening the relationship of the controls across plates. If this issue cannot be explained by biological or experimental artifacts, further work will be needed to explore how this balance can be achieved such that these methods can be useful for a larger set of similar data.

Chapter 5

Conclusion

As we can see from the examples in this thesis, fitting nonlinear curves to dose response data and obtaining useful information is tricky and different from case to case. It requires immense involvement with the data itself and plenty of trial and error. We want a model that is specific enough for the data at hand but general enough to accommodate future data. We also need to balance between the complexity of the model and fidelity to the data.

5.1 General guidelines

We discuss some general tips on proceeding with analyzing dose-response data. Many of the methods can be commonly applied to a majority of such datasets. The key is to try to fully understand the unique characteristics of each dataset and to work to tailor such techniques to optimally address them.

Raw data

Direct output from luminometers and other such machines that can detect the light intensities from samples is frequently a flat file of these measurements listed in a single column. Most likely there may be additional columns that indicate the specific location of each measurement on the read chip or plate. Making sure to map these values to the correct factors is extremely important. All the scripts written to process raw data in this dissertation include many checks to confirm that the resulting data frame to be used for analysis is correct. Any misassignments of things such as compound level or even a shift in the numbering of replicates will ruin all future attempts at exploring and understanding the data to be modelled. Because randomization of the spot or well layout is preferred, functions are written to provide simple visual checks for the experimenter as well.

Such a raw format should typically be the preferred and requested format. Data obtained in a processed form may be subjected to errors in human handling or frequently, incompleteness, as the experimenter may feel that certain observations are bad and chose to exclude them. Other times, data important to our work such as control values may also be left out because the experimenter has already deemed the values to be sufficiently within range and therefore useless for any other purposes.

However if the dataset is extremely large or has the potential to grow enormously, then this raw format may not be the best. If the data can be correctly loaded and managed on a relational database, we can pull out the data we require efficiently and without the redundancy in the long data format of a typical R data frame.

Exploring the data

As with all data, we must take a look at it before we attempt to model anything. Exploratory analysis can involve numerical summaries such as taking means and standard deviations by different factors and various plots to uncover any patterns.

We suggest that the first plot to do is the data of one unit. This is a check that the data is stored correctly and to see the general trend in the observations across the concentrations of compound. The x-axis is the logged concentration while the y-axis is the response, typically the optical density of the spots or wells we are measuring. Then it should be easy to extend to a grid of plots organized by any important factors such as specific compounds or enzyme conditions. It may be possible to detect some patterns this way, but it may be difficult unless there are strong effects in the data. This is more for looking over the data and seeing if there are any strange outlying units. We also need to be careful in dismissing units as outliers especially if some seem to be non-monotonic. Even though none of the datasets discussed here are like this, it is indeed possible due to the complexity of the reactions involved. In such situations, the use of logistic curves and the like will not be appropriate; some useful methods include smoothing spline anova models [10].

Individual nonlinear fitting

Once we can assume that the logistic curve will work relatively well on the dataset, we try fitting it on each unit separately. Choosing the three or four parameter model depends on whether we think there is sufficient evidence to assume that the lower asymptote of the curve is zero. We can always start with looking at the results of the four-parameter model and checking if the estimate is near zero, although there may be increased difficulty in fitting a more complex model.

Because each set of estimates depends on only each unit's data, it is often that these observations may not be enough to be able to estimate a full curve. The recurring issue in all the datasets discussed in this thesis that causes the motivating factor for labs to contact a statistician after large portions of the data has been produced is the less than optimal range of concentrations of compounds used to interrogate the cellular activity. As we have seen in the plots of the raw data, too much real estate on the plates, chips, or slides have been wasted on interrogating the flat areas where there is not enough compound to affect anything and where there is too much and not utilized efficiently for the important stage of the activity.

Fitting it robustly may help a little, but it should be clear from this dissertation that individual nonlinear fitting is insufficient on its own for analyzing this data, although it provides a good foundation to base subsequent global analysis on.

Through analyzing the set of parameter estimates that we are able to obtain from individual fitting, we can design a model that will encompass the important characteristics of the dataset as a whole. This is done through boxplots of the estimates by specific factors and simple analysis of variance.

Global nonlinear fitting

Armed with what the ideal model should likely be, we can attempt to fit it on the dataset using several different methods. Constraining the parameters is often a good and easy method of fitting the data jointly. If the results from this is not ideal, mixed effects models can offer much more flexibility in balancing between forcing parameters to be equal and letting their individual characteristics be represented. However it is quite typical that this is unsuccessful, either due to the fitting methods or the fact that the data does not have enough information to support such a model.

Determining what error messages mean and how exactly the model is not fitting is difficult. Minimization/maximization algorithms can often get trapped in local optima. Compromises are sometimes necessary to be able to fit anything at all. We have to determine whether such changes are appropriate and keep on iterating through many different models to see what is best. The work in this dissertation illustrates a few examples of model building, but it is highly dependent on each individual dataset. The incorporation of external data that is pertinent to the issues at hand can also be helpful.

Further improvements to the suggested models in this dissertation include finding appropriate transformations of the parameters [19] and using better methods to deal with the heteroscedasticity and robustness, such as transforming both sides [15].

5.2 Practical issues

Even if the advantages of fitting data jointly are clear in the results we've seen, an extremely important consideration is the practical issues in computation. Data is always continually being generated and should be processed quickly for input in further analysis and experiments. If allowing colloborators to use the program, will it run reasonably and relatively quickly for the majority of the time? Or will it result in strange errors and halt that analysis until the writer of the code or other statisticians are able to figure it out and either fix it simply or change the existing model to adapt to the new data that it choked on? Ideally, we want programs that are efficient and near fail-safe, but significant time and effort is needed to accomplish this.

Although current programs out there for applying nonlinear mixed effects models to data are thoughtfully coded and their small examples run absolutely well, general utilization is often unsuccessful. At this point, each dataset must be considered on its own very carefully to ensure that reasonable results, if any, will be obtained. This has proved true for the datasets involved in this dissertation, as well as for other datasets in the public literature [4].

5.3 Conclusion

The most important tool for the analysis of such data is perseverance. Nonlinearity makes things complicated, and fitting nonlinear mixed effects models is no exception. We should always be ready with new ideas to implement and improve the analysis at each stage even if roadblocks abound.

Bibliography

[1] National Cancer Institute COMPARE Methodology. `http://dtp.nci.nih.gov/docs/compare/compare_methodology.html`.

[2] Wikipedia, the free encyclopedia. `http://www.wikipedia.org`.

[3] C.K. Augustine, J.S. Yoo, A. Potti, Y. Yoshimoto, P.A. Zipfel, H.S. Friedman, J.R. Nevins, F. Ali-Osman, and D.S. Tyler. Genomic and molecular profiling predicts response to temozolomide in melanoma. *Clinical Cancer Research*, 15(2):502, 2009.

[4] L.A. Baharith, A. Al-Khouli, and G.M. Raab. Cytotoxic assays for screening anticancer agents. *Statistics in Medicine*, 25(13), 2006.

[5] D.M. Bates and S. DebRoy. Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, 91(1):1–17, 2004.

[6] R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[7] A. Citri and Y. Yarden. Egf-erbb signalling: towards the systems level. *Nature Reviews Molecular Cell Biology*, 7(7):505–516, 2006.

[8] SPM Crouch, R. Kozlowski, KJ Slater, and J. Fletcher. The use of atp bioluminescence as a measure of cell proliferation and cytotoxicity. *Journal of immunological methods*, 160(1):81–88, 1993.

[9] M. Davidian and D.M. Giltinan. *Nonlinear models for repeated measurement data Monographs on statistics and applied probability.* Chapman & Hall, 1995.

[10] C. Gu. *Smoothing spline ANOVA models.* Springer Verlag, 2002.

[11] M.Y. Lee and J.S. Dordick. High-throughput human metabolism and toxicity analysis. *Current opinion in biotechnology*, 17(6):619–627, 2006.

[12] M.Y. Lee, J.S. Dordick, and D.S. Clark. Metabolic enzyme microarray coupled with miniaturized cell-culture array technology for high-throughput toxicity screening. *Methods in molecular biology (Clifton, NJ)*, 632:221, 2010.

[13] M.Y. Lee, R.A. Kumar, S.M. Sukumaran, M.G. Hogg, D.S. Clark, and J.S. Dordick. Three-dimensional cellular microarray for high-throughput toxicology assays. *Proceedings of the National Academy of Sciences*, 105(1):59, 2008.

[14] M.Y. Lee, C.B. Park, J.S. Dordick, and D.S. Clark. Metabolizing enzyme toxicology assay chip (metachip) for high-throughput microscale toxicity analyses. *Proceedings of the National Academy of Sciences of the United States of America*, 102(4):983, 2005.

[15] A. Oberg and M. Davidian. Estimating data transformations in nonlinear mixed effects models. *Biometrics*, 56(1):65–72, 2000.

[16] J.C. Pinheiro and D.M. Bates. *Mixed-effects models in S and S-PLUS*. Springer, 2000.

[17] Promega Corporation. *CellTiter-Glo® Luminescent Cell Viability Assay Technical Bulletin TB288.*

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2009. ISBN 3-900051-07-0.

[19] G.J.S. Ross. *Nonlinear estimation.* Springer Series in Statistics, 1990.

[20] T. Speed. [That BLUP is a Good Thing: The Estimation of Random Effects]: Comment. *Statistical Science*, pages 42–44, 1991.

[21] H. Wickham. *ggplot2: elegant graphics for data analysis.* Springer-Verlag New York Inc, 2009.