Title
Modeling Multiple Drugs on Lung Cancer and Normal Cells using Regression

Author
Chen, Michelle

Publication Date
2013

Peer reviewed|Thesis/dissertation

University of California

Los Angeles

Modeling Multiple Drugs on Lung Cancer and Normal Cells using Regression

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Statistics

by

Michelle Chen

2013

Abstract of the Thesis

Modeling Multiple Drugs on Lung Cancer and Normal Cells using Regression

by

Michelle Chen

Master of Science in Statistics

University of California, Los Angeles, 2013

Professor Hongquan Xu, Chair

Multiple drugs is much more effective in a cancer treatment than single drugs; however, it is challenging to find a reliable model that describes the relationship between drugs and the cell activities, as well as to construct efficient designs for reducing the number of drug combinations needed to be tested. We analyze data that consists of cellular ATP-levels on lung cancer cells and normal cells with 3 inhibitor drugs, AG490, U0126 and I-3'-M, each at 8 dosage levels. We construct models with different number of predictors and methods as well as different subsets of the full data. We find that large dataset does not guarantee the accuracy of a model's prediction, but an appropriate design and a good model building method are the keys. Square root transformation on the response gives a model prediction surpassing the performance among all other models we built. A second order model constructed with a 27-run factorial design performs remarkably well for both cancer cells and normal cells. It is in general much harder to model or predict cancer cells than normal cells.

The thesis of Michelle Chen is approved.

Ying Nian Wu

Frederic Paik Schoenberg

Hongquan Xu, Committee Chair

University of California, Los Angeles

2013

To my father . . . who — willingly accepts me as I am —

supports my decision and believes in me more than I believe in myself.

Dad, at down times, I know you always have my back.

In the years to come, I just want you to know that I have yours.

Table of Contents

List of Figures

List of Tables

Acknowledgments

My sincere gratitude to Professor and Graduate Vice Chair, Hongquan Xu, who guided my research diligently from beginning to end, answered my questions with great patience and gave me valuable critics and feedbacks regarding my chosen methods used to analyze this data. I can't possibly accomplish what I had without his undivided support. A special thanks to Professor Ren Sun from Department of Molecular and Medical Pharmacology for sharing this lung cancer data.

I would also like to thank Professor and Chair, Frederic Paik Schoenberg, and Professor Ying Nian Wu for serving my committee. Professor Wu, he shows me alternative explanations regarding my questions raised from his *Theoretical Statistics* class. He makes difficult concepts sound so simple and easy. I'm truly inspired. Professor Schoenberg, he goes above and beyond showing his students derivations and proofs in his *Regression Analysis* class. This experience makes reading related literatures much easier later on.

My special thanks to the Department Student Affair Officer, Glenda Jones, who takes great care of every student in this department. Her continuous support and strong belief in every student's potential to succeed, truly make our department feels like home. Our department wouldn't be the same without her.

Lastly, to my fellow graduate students, thank them all for making this challenging and demanding life as a graduate student here at UCLA quite unforgettable. Their company and laughter always brighten up my day when I need a quick escape from textbooks and research articles.

CHAPTER 1

Introduction

1.1 Motivation

Inhibition of cell proliferation has been a widely used approach in cancer treatment. Certain drugs can change the signal pathway mechanism of a cellular response; they can increase or decrease chance of survival for a targeting cell in a biological system. You can refer to [EV11] if you are interested in learning more about the biology behind cancer treatments. A common problem observed in a single drugs treatment is its toxicity. Since the effective dosage is usually high in concentration, it is toxic in a biological system. Research shows a multiple drugs treatment has advantages over a single drugs treatment in killing cancer cells. With multiple drugs treatment, the effective dosage for each drug in a combination is usually lower in comparison to the effective dosage in a single drugs treatment. Thus, there is a genuine interest to how multiple drugs interact within a cellular system such that inhibition of a target cellular proliferation can be achieved.

The challenge lies in explaining the complex chain of reactions initiated by multiple drugs in a biological system. Since drugs are also known to inhibit/activate targets aside from the desired target, this may lead to unknown reactions, may inhibit/activate some signal pathways, directly or indirectly, and change the desired response. Together the combinatory effects is difficult to predict and that the best dosage combination from a multiple drugs treatment giving the optimal

performance, killing most cancer cells yet preserving most normal cells, in a cancer treatment is thus hard to find.

The lack of detailed knowledge on a cellular system and its response to simultaneous effects from multiple drugs suggest an alternative approach: *in silico* experiment. This approach eliminates the intermediates, namely activation/inhibition of signal pathways and undesired targets, and gives a direct observation to the fundamental cellular response. In any typical *in silico* experiment, cellular ATP-level is generally accepted as an indicator for assessing inhibition of cellular proliferation.

Currently, scientists need many trial and error experiments to find the dosage with maximal effect from an individual drug. The found dosage for each drug will be the initial starting dosage in a multiple drugs treatment. And from there, biologists must decide the most appropriate range for each drug in a combination from a multiple drugs treatment through yet another set of trial and error experiments. Keep in mind that the co-existence of other drugs will almost always change the drug effect observed from the individual drug experiment, together the drug under observation may have an increased or decreased potency. This searching process can be time consuming and costly; hence, a good model which describes the relationship between drug to drug interaction and an individual drug effect on cellular response coexisting with other drugs will make the search process less tedious and more efficient.

Above gives us a strong motivation to come up with a mathematical model, which can successfully capture this intricate relationship among the drugs; how they change each other's potency and change the response in inhibition of a target cellular proliferation. Moreover, we are interested to find a minimal yet effective

data subset which suffices to build a good model, so the experiment is also cost effective.

1.2 Data

This data taken from [AYF11] have 3 drugs, AG490, U0126, I-3'-M, and all are inhibitors. Each drug has 8 concentrations giving a total of 512 combinations. Table 1.1 gives the concentrations used for the 3 drugs, which are abbreviated as A, B and C. Whenever a drug combination or drug treatment is referred in this paper, it could be: drug A at 300.00 (μM), drug B at 100.00(μM) and drug C at 1.00(μM) or at any of the concentrations shown in Table 1.1. A drug treatment's potency on A549 non-small lung cancer cells (yc) and AG02603 normal fibroblast cells (yn) are determined through the measurement of cellular ATP-level. Cellular ATP-level for all combinations were measured 72 hours after initiation of the drug treatments, and they were normalized against untreated cellular ATP-level. This is done by recording the percentage of cell survival in a scale from 0 to 1 in a well of cancer cells or a well of normal cells, where 0 implies no cells survive, and 1 implies all cells survive. A few drug treatments from the dataset is shown in Table 1.2.

Table 1.1: Drug Concentration

Abbrev.	Drug Name	Drug Concentration(μM)
A	AG490	0,0.3,1,3,10,30,100,300
B	U0126	0,0.1,0.3,1,3,10,30,100
C	I-3'-M	0,0.3,1,3,10,30,100,300

Table 1.2: Drug Combination

Obs Index	A	B	C	yn	yc
95	30.00	0.10	0.30	0.88	0.82
359	3.00	3.00	300.00	0.16	0.00
293	300.00	10.00	100.00	0.29	0.05
86	1.00	1.00	0.30	0.98	0.95
480	100.00	0.00	1.00	0.64	0.49

1.3 Plots

1-Way Interaction

Normal cell's ATP-level, as shown in boxes $A|A$, $B|B$, $C|C$ from Figure 1.1, and cancer cell's ATP-level, as shown in boxes $A|A$, $B|B$, $C|C$ from Figure 1.2, both drop as drug A concentration increases. But notice, the cancer cell's ATP-level drops in a faster rate in comparison to the normal cell's ATP-level, which is in accord with our objective: "kill cancer cells while preserving normal cells" and is considered a favorable outcome. The same is true for drug B and drug C; we see that all three drugs show desired outcome in a lung cancer treatment.

2-Way Interaction

Normal Cell

When drug A, drug B, or drug C is fixed at a particular concentration from any of the 8 concentrations listed in Table 1.1, we notice the ATP-level drop begins within intervals (30μM, 100μM) and (10μM, 30μM) when drug C is interacting with drug A and drug B respectively, within interval (10μM, 30μM) when drug

B is interacting with drug A and drug C, within intervals (10μM, 30μM) and (30μM, 100μM) when drug A is interacting with drug B and drug C respectively. Despite high concentration from drugs A, B and C (concentration lines 100, 300 for A, C and lines 30, 100 for B) appears to have initially a much lower ATP-level, the slope of these lines when interacting with other drugs is not as steep as the others, especially noticeable in boxes $B|A$ and $C|A$ from Figure 1.1. The worst cases observed are in boxes $A|C$ and $A|B$ from Figure 1.1 when the lines 100 and 300 reverse in direction as the concentration of drug A increased in dosage.

Cancer Cell

Similar trend is observed like we have seen in normal cells; however especially worth notice is box $B|C$ from Figure 1.2 because when drug C is fixed at concentration line 300, as it interacts with drug B, increasing drug B dosage does not lower ATP-level, it increases ATP-level.

Density

Normal Cell

Density distribution of normal cell from original data looks multi-modal as shown on Figure 1.3, and it is hard to determine exactly how many there are. After log transformation on the response, it becomes clear that it is tri-modal, but it is very left skewed. Mean from each modal increases in the power of 2 from left to right, and spread from each modal is similar in magnitude. With a square root transformation on the response, it still is tri-modal, however, not as left skewed.

Cancer Cell

Density distribution of cancer cell looks bi-modal on Figure 1.4. After log transformation on the response, it still looks bi-modal, but it is very left skewed like observed in normal cell. Notice, one modal has a mean that is twice the value of the other; however, spread for both are comparable. With a square root transformation on the response, it becomes tri-modal, where each has a comparable spread, but their means seem to increase in the power 2 from left to right.

1.4 Approach

Our approach can be divided into two parts. First, we are interested in building an effective model including just enough terms; hence, we consider 3 models consisting sets of terms. They will be specified further in *Section 2.1 Model Prototype*. Here we briefly describe the terms each model includes - model g1 contains only linear predictors; model g2 contains the linear and 2-way interaction predictors; model g3 contains linear, 2-way interaction and quadratic predictors. Higher order terms were tested and did not show significant importance. We thus limit ourselves to focus on these 3 particular models.

Second, we are also interested in exploring a cost effective experiment, an optimal number of runs. That is with the least number of drug combinations yet still proves to be effective in building reliable models. We choose 3 specific subsets of drug combinations from the full dataset for comparison purpose. These data subsets are given the following names: n_{512}, n_{80}, n_{27}. Detail on the selection of these subsets is covered in *Section 2.2 Data Subset Selection*.

Now that we have determined the models we intend to investigate, namely, g1, g2

and g3, as well as the subsets, n_{512}, n_{80}, and n_{27}, that will be used to construct these models. We applied different model building methods like Ordinary Least Squares, Response Transformation and Generalized Linear Model as covered in *Chapter 2* to all the combinations from mixing model prototypes and data subsets. After these models are built we assess them with several criteria: a model's goodness of fit is assessed with R^2 and $\widehat{\sigma}$, a model's performance on prediction is assessed with $RMSE$, $\widehat{R}$ and $\widehat{R}$ Plot. Detailed definitions regarding all criterion mentioned are explained explicitly in *Chapter 3*. Finally, we conclude our finds in *Chapter 4*. We will discuss an effective model from a few perspectives: what is an appropriate number of terms to include, what is a reasonable number of runs that is cost and prediction effective, lastly, error diagnostics of chosen models are checked to see if they are statistically valid.

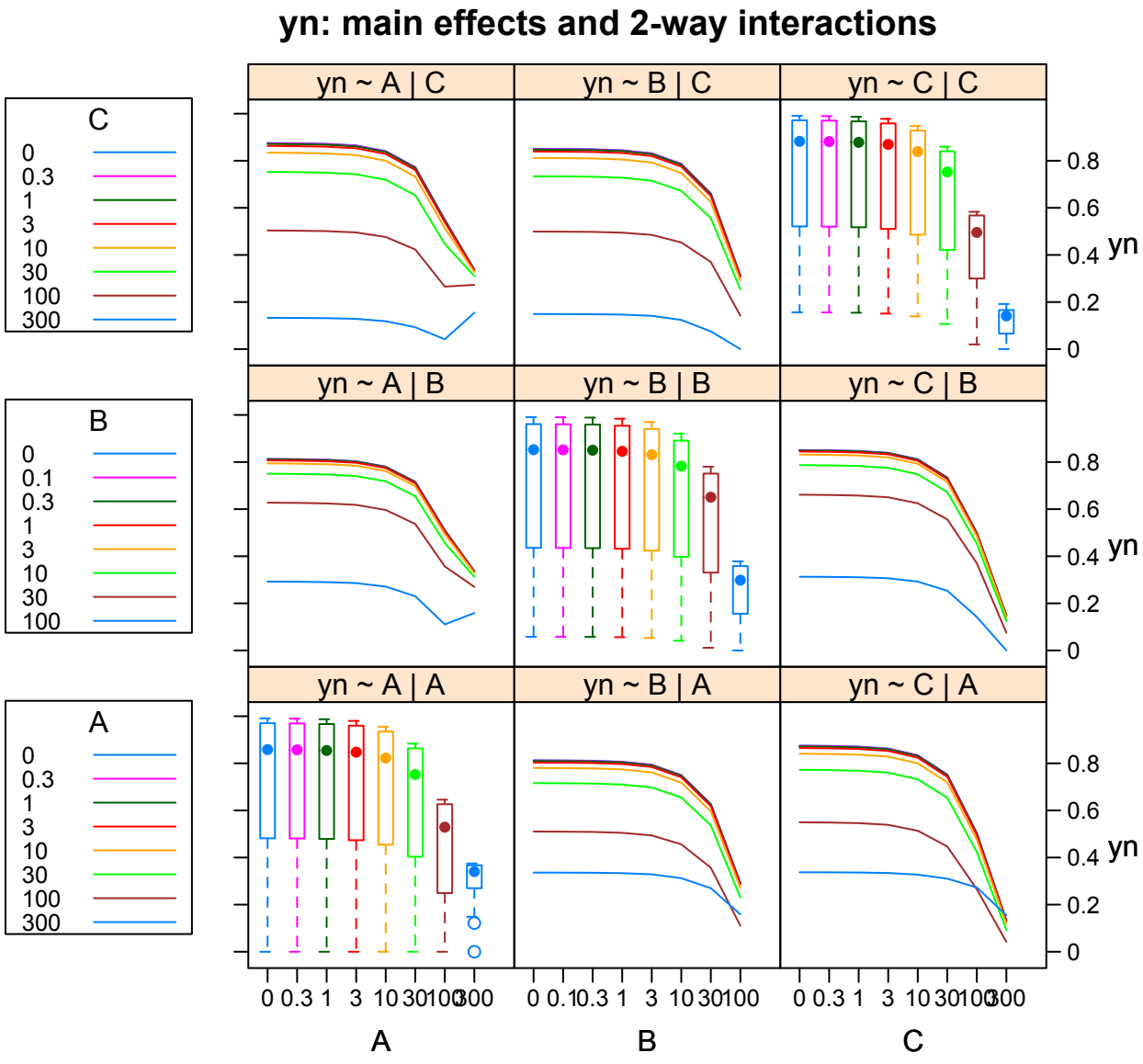


Figure 1.1: Interaction Plots for Normal Cells

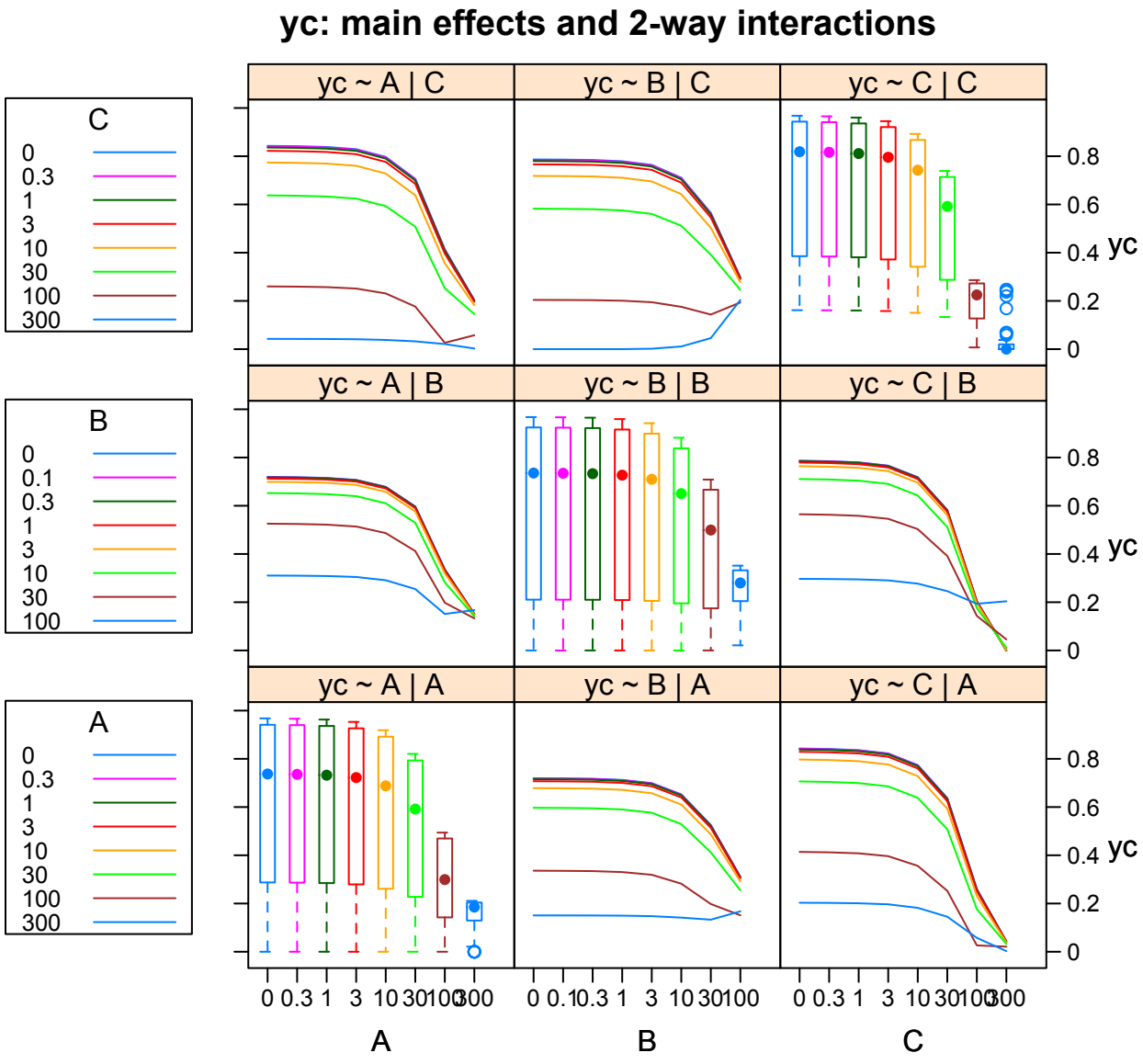


Figure 1.2: Interaction Plots for Cancer Cells

Figure 1.3: Density Distributions for Normal Cells

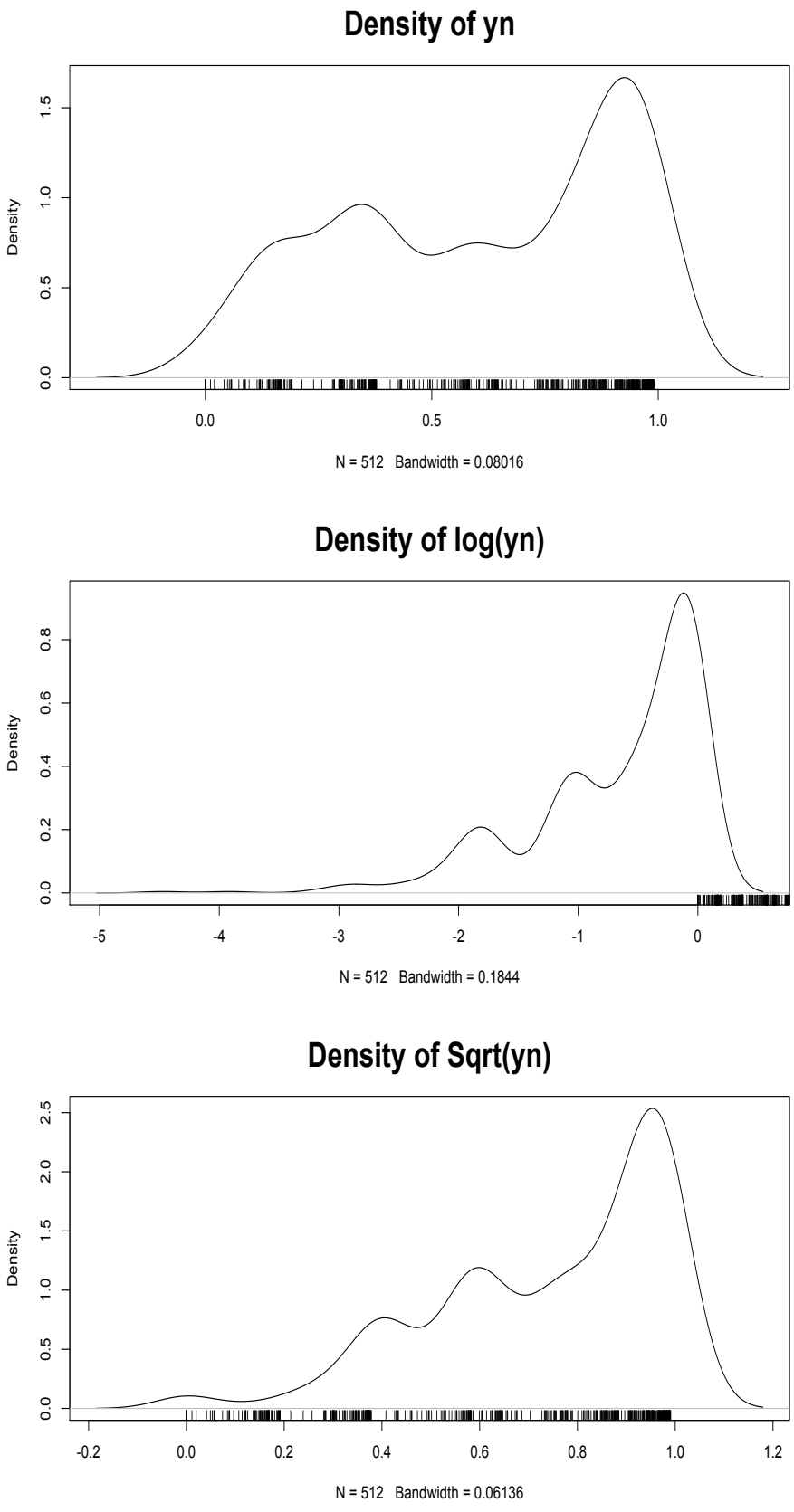
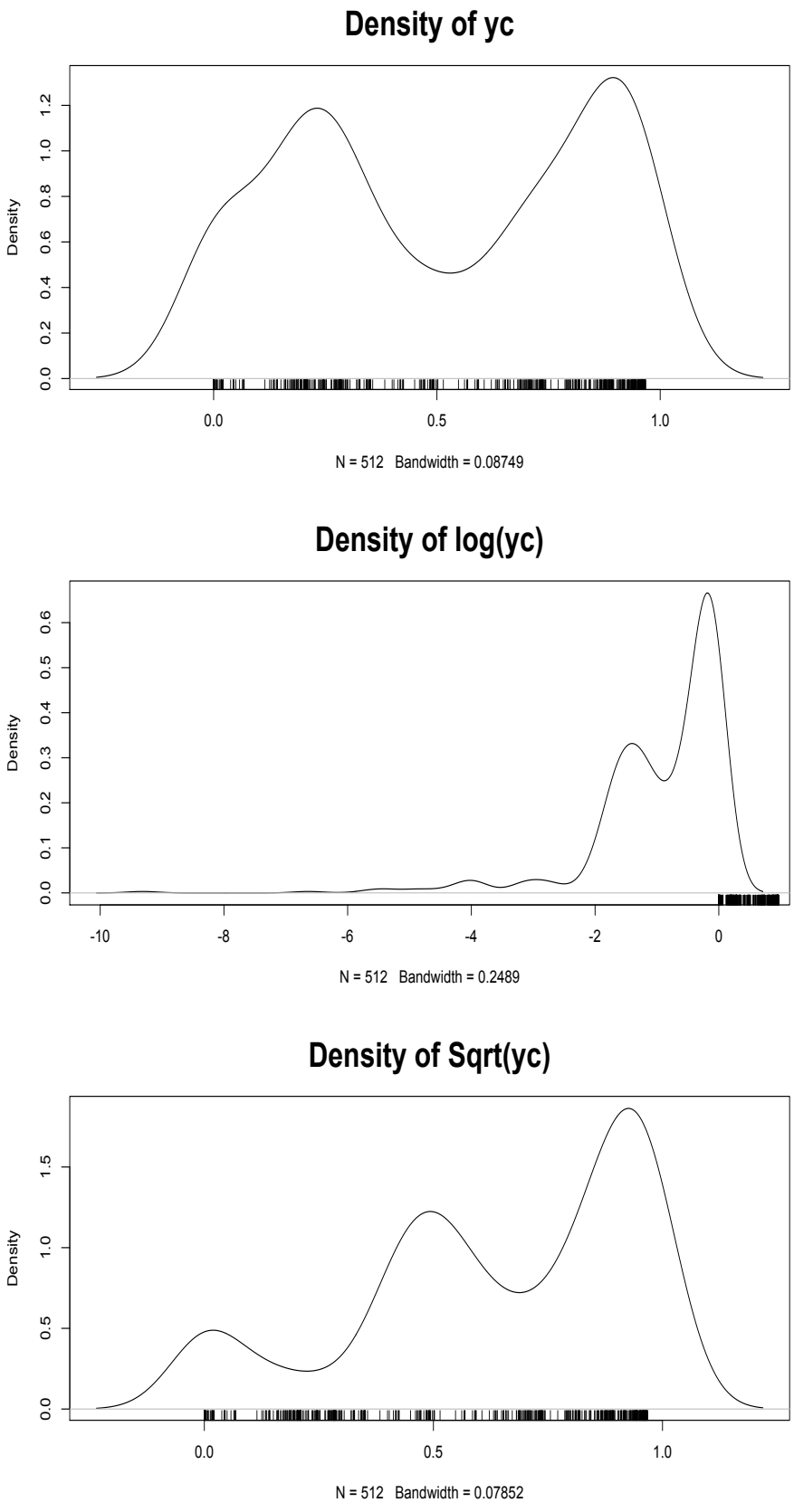


Figure 1.4: Density Distributions for Cancer Cells



CHAPTER 2

Method

2.1 Model Prototype

An effective model comes with just enough terms to describe a relationship between objects of interest. In our case, the objects are the cells and the three drugs. After fitting models with different number of terms, we have narrowed our focus to primarily three models. Also, after some experimentation with this dataset, it appears that predictors beyond the quadratic terms are not necessary. We thrive to avoid over fitting the models; hence, after careful evaluation only the following models will be built and discussed for both normal cell and cancer cell. These models are called: g1, g2 and g3 and have the following prototype where g1 contains only first order terms, g2 contains first order and 2-way interaction terms, and g3 contains first order, 2-way interaction and quadratic terms. Below is a model's prototype defined in a mathematical expression,

$$y_j = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

$$p \in (\mathbb{Z} \mid 2 < p < 10), \quad j = (cancer\,cell, normal\,cell)$$

Depending on the specific model prototype chosen, the total number of β changes accordingly, see Table 2.1 for the correspondence between variable x_p and the predictors: drugs A, B and C. Variables from the above expression will be explained thoroughly in *Section 2.3 Linear Regression: Least Squares Estimation.*

		Model		
p	x_p	$g1$	$g2$	$g3$
0	Intercept	✓	✓	✓
1	A	✓	✓	✓
2	B	✓	✓	✓
3	C	✓	✓	✓
4	AB		✓	✓
5	AC		✓	✓
6	BC		✓	✓
7	A^2			✓
8	B^2			✓
9	C^2			✓

Table 2.1: Model Prototype

2.2 Data Subset Selection

When we construct our models the maximum of n is 512, where n stands for the number of drug combinations. This is true simply because that concludes all the possible drug combinations from an experiment consists of 3 drugs and each consists of 8 concentrations. In fact, we will build g1, g2 and g3 with the following data subsets from the full dataset: a large set with 512 drug combinations abbreviated as n_{512}, a medium set with 80 randomly chosen drug combinations abbreviated as n_{80} like seen in Table 2.2, and a small set with 27 drug combinations abbreviated as n_{27}.

We choose these 3 subsets specifically for a few reasons. First, n_{512} is chosen because it is simply the complete dataset, and we are interested to study the behavior of the cells when interacting with the drugs using all available data and to extract as much information as we possibly can. Since [AYF11] claims that models built with 80 randomly chosen drug combinations and some methods of estimation always give satisfactory prediction, we also include n_{80}. It is randomly chosen using R, so that we can compare models built using n_{80} against other two subsets and investigate this claim. Lastly, n_{27} is a subset selected with a specific factorial design which improves resolution of a model. See [WH09], [Mon05] and [DXH12] for more information on topics regarding experimental design. Basically, n_{27} contains only concentration level 0, 5, and 7 from all 3 drugs. Refer back to Table 1.1, each drug has 8 available concentrations listed in ascending order and each concentration from that list corresponds to a number from 0 to 7, with the lowest in concentration corresponds to 0 and highest in concentration corresponds to 7. Hence, there should be no confusion for readers regarding concentration level 0, 5, and 7.

Table 2.2: n_{80} Selection

Observation Index							
4	6	7	19	30	37	55	64
66	72	73	74	77	80	81	82
86	91	95	98	112	117	120	122
126	127	136	137	140	163	170	174
175	191	196	202	206	220	233	236
239	246	248	277	278	279	282	293
300	315	317	322	326	331	349	359
372	376	378	381	385	389	401	402
411	417	421	425	427	442	450	455
464	466	467	479	480	485	496	511

2.3 Linear Regression: Least Squares Estimation

In this section, we will go over a model's mathematical representation and explanation of the notations used. Since the same representation will be used throughout the next few sections when we cover methods for building models, we like to explore the representation in detail for the reader, so there is no ambiguity. Note, it is assumed that readers are familiar with linear regression in general and that this section does not mean to serve as a detail introduction to the least squares estimation. One should refer to [MS03] for a thorough introduction on *Linear Regression*. Also you can refer to [Far05] and [Far06] on statistical software *R* which we utilized for our computation.

This dataset has **two response variables**: *yn: ATP-level of Normal Cell*, *yc: ATP-level of Cancer Cell*, and **three drugs**: *drug A*, *drug B* and *drug C*. They

can all be presented in a tabular form which look like the below,

$$\begin{array}{ccccc} yc_1 & yn_1 & A_1 & B_1 & C_1 \\ yc_2 & yn_2 & A_2 & B_2 & C_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ yc_n & yn_n & A_n & B_n & C_n \end{array}$$

where $n = 1, 2, 3, ..., 512$ is the index of observation in the dataset and $yc = (ATP - level\, of\, Cancer\, Cell)$ and $yn = (ATP - level\, of\, Normal\, Cell)$. Let's say we want to find a linear relationship between **Response Variable**: yc and **Predictors**: A, B, C using a linear model; we write the following:

$$yc_i = \beta_0 + \beta_1 A_i + \beta_2 B_i + \beta_3 C_i + \epsilon_i \; where \; i = 1, ..., n$$

Thus, there are n equations, and β_0, β_1, β_2, β_3 are unknown parameters to be estimated. Note, $max(n) = 512$; hence, any n less than 512 implies we are only using a subset of our data like n_{80} and n_{27}.

We can further simplify the notation and leave behind excessive subscripts if we adapt the matrix representation. Now these n equations are simply:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and we refer $\mathbf{X}\boldsymbol{\beta}$ as the deterministic component and $\boldsymbol{\epsilon}$ as the random component of a probabilistic model where

$$\mathbf{y} = (yc_1, \cdots, yc_n)^T, \quad \boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^T$$

and,

$$\mathbf{X} = \begin{array}{cccc} 1 & A_1 & B_1 & C_1 \\ 1 & A_2 & B_2 & C_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & A_n & B_n & C_n \end{array}$$

as we change the number of parameters that we like to estimate, size of the matrix $\mathbf{X}$ and vector $\boldsymbol{\beta}$ would change accordingly. In general, the response lies in an $n - dimensional$ space while $\boldsymbol{\beta}$ lies in a $p - dimensional$ space, where p is the number of parameters, and that $p < n$ is required for a unique $\boldsymbol{\beta}$ estimation.

Suppose we like to estimate the β's for predictors A, B, C and ABC (interaction among drug A, drug B and drug C), then $\mathbf{X}$ and $\boldsymbol{\beta}$ are now defined as:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)^T \qquad \mathbf{X} = \begin{matrix} 1 & A_1 & B_1 & C_1 & A_1B_1C_1 \\ 1 & A_2 & B_2 & C_2 & A_2B_2C_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & A_n & B_n & C_n & A_nB_nC_n \end{matrix}$$

where β_0, β_1, β_2, β_3, and β_4 are estimates of the effect for intercept, drug A, drug B, drug C and interaction among drug A, B, C. We apply the least squares estimation to estimate these $\boldsymbol{\beta}$ parameters. The least squares estimate of $\boldsymbol{\beta}$ is called $\widehat{\boldsymbol{\beta}}$, which minimizes the residual sum of squares:

$$\boldsymbol{\epsilon^T \epsilon} = \boldsymbol{(y - X\beta)^T (y - X\beta)}$$

such that it explains as much of the response as possible, and that the structure in the data can be successfully captured in the parameters leaving least variation in the residual. Some people use error as an alternative term for residual. To obtain $\widehat{\boldsymbol{\beta}}$, we differentiate the above expression with respect to $\boldsymbol{\beta}$ and setting the resulting expression to zero, we should find this $\widehat{\boldsymbol{\beta}}$ satisfies the normal equations below,

$$\boldsymbol{X^T X \widehat{\beta} = X^T y}$$

Now, if $\boldsymbol{X^T X}$ is invertible then we have our parameter estimation:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{(X^T X)^{-1} X^T y}$$

All the steps described above can be done using available commands from a statistical software called R.

2.4 Ordinary Least Squares Model

Ordinary **L**east **S**quares models are constructed with the original response using the least squares estimation to estimate $\boldsymbol{\beta}$ parameters. Model prototype g1, g2 and g3 are built with subsets n_{512}, n_{80}, n_{27} for both yc and yn. Thus, there are 9 OLS yn models and 9 OLS yc models. We show models built using n_{27} with yn in Table 2.3, and models built using n_{27} with yc in Table 2.4. We only intend to do this once for illustration purpose. Models built with other methods will have similar forms but a different set of $\boldsymbol{\beta}$ parameter estimations.

In *Chapter 3*, when we defined our criterions used to assess models, you will repeatedly encounter the following notations:

(1) y_i, this is an observed response with observation index i. Refer back to Table 1.2 when $i = 95$, y_{95} is 0.88 for yc and 0.82 for yn.

(2) $\widehat{y}_i$, this is a model predicted response with observation index i. Again, looking back at Table 1.2 when $i = 95$, we have the following drug combination: drug A at 30 μM, drug B at 0.10μM and drug C at 0.30μM. To get $\widehat{y}_{95}$ for $Model_{g3}$ from Table 2.3, we multiply the appropriate predictors to their corresponding coefficients and sum up the products, then there we have $\widehat{y}_{95}$ for $Model_{g3}$ predicted yn. Note, $AB = 30 * 0.10$, $A^2 = 30 * 30$, and likewise for other 2-way interaction and quadratic terms.

(3) $\overline{y}$ this is mean generated from a data subset using the original responses, and $\overline{\widehat{y}}$ this is mean generated from a data subset using the model predicted responses. The definition of a mean should be apparent, but we list it below anyway.

$$\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

where $\overline{Y}$ can be replaced by $\overline{y}$ or $\overline{\widehat{y}}$, and that Y_i can be replaced by y_i or $\widehat{y}_i$ to get the desired mean value. n is the size of the dataset.

(4) $\widehat{\epsilon}_i$ this is called the i^{th} residual or the i^{th} error. It is simply the difference between an observed response and a model predicted response for observation index i, and it is defined as,

$$\widehat{\epsilon}_i = y_i - \widehat{y}_i$$

Table 2.3: $OLS_{n_{27}}$ yn Model Coefficients

	$Coefficients$		
$Term$	$Model_{g1}$	$Model_{g2}$	$Model_{g3}$
$Intercept$	0.778268	0.912015	0.955943
A	-0.000837	-0.001700	-0.003472
B	-0.003388	-0.005389	-0.006575
C	-0.001550	-0.002452	-0.003705
AB		0.000009	0.000009
AC		0.000005	0.000005
BC		0.000010	0.000010
A^2			0.000006
B^2			0.000011
C^2			0.000004

2.5 Response Transformation Model

The estimation from the OLS model depends heavily on satisfactory error assumptions. The initial regression diagnostics on the OLS models reveal a clear problem of non-constant variance. By observing the $\widehat{\epsilon}_i$ vs. $\widehat{y}_i$ plot, note $max(i)$ can be 27, 80 or 512 depending on the data subset used for model construction, we

Table 2.4: $OLS_{n_{27}}$ yc Model Coefficients

	$Coefficients$		
$Term$	$Model_{g1}$	$Model_{g2}$	$Model_{g3}$
$Intercept$	0.647165	0.804114	0.880431
A	-0.001060	-0.001866	-0.003761
B	-0.001506	-0.004375	-0.006866
C	-0.001411	-0.002503	-0.005720
AB		0.000009	0.000009
AC		0.000004	0.000004
BC		0.000017	0.000017
A^2			0.000006
B^2			0.000024
C^2			0.000010

notice various patterns, which are indicators of violation to the constant variance assumption.

Our approach to dealing with non-constant variance is transforming the response variables y to $h(y)$, so that when appropriate $h()$ is chosen the variance of $h(y)$ should be constant. However, such $h()$ is usually an art of guess work, and we make our guesses by separately applying the following $h()$ transformations on our response: (1) Log (2) Logit (3) Square Root. Note Square Root will be expressed as Sqrt on Tables and Figures. Log and Square Root should not be foreign to readers, and Logit has the following definition:

$$logit(y_i) = log(\frac{y_i}{1 - y_i})$$

Basically, the steps toward building a model is identical to those in OLS with an exception of a prior response transformation of our choice. Special attention need

to be taken when calculating some of the criterions in *Chapter 3*. Since we use the transformed response to build a model, it is important to transform back the result so that observed and predicted responses can be compared on the same scale.

Here, $\widehat{y}_i$ is a model predicted response with observation index i that is not scaled. An extra step of inverse transformation is required for a scaled model predicted response with observation index i which we denote as $\tilde{y}_i$. We notated inverse transformation as $h^{-1}()$. Like before, we refer to $i = 95$ on Table 1.2, we have the following drug combination: drug A at 30 μM, drug B at 0.10μM and drug C at 0.30μM, and to get $\widehat{y}_{95}$ for a model, we multiply the appropriate predictors to their corresponding coefficients and sum up the products. Yes, this result is a model predicted response value, but keep in mind, this model predicted response value is not the same scale as the observed response value; thus, we need to perform $h^{-1}(\widehat{y})$ and make sure a fair comparison can be done between a scaled model predicted response and observed response. Depending on the choice of $h()$, the inverse transformation also differs, see below:

(1) If $h()$ is Log, then the transformed back response is

$$\tilde{y}_i = exp(\widehat{y}_i)$$

(2) If $h()$ is Logit, then the transformed back response is

$$\tilde{y}_i = \frac{exp(\widehat{y}_i)}{1 + exp(\widehat{y}_i)}$$

(3) If $h()$ is Square root, then the transformed back response is

$$\tilde{y}_i = \widehat{y}_i^2$$

Last remark, a total of 54 models will be constructed in this section: 9 Log yn models and 9 Log yc models; 9 Logit yn models and 9 Logit yc models; 9 Sqrt yn models and 9 Sqrt yc models.

2.6 GLM: Binomial Regression Model

Another problem observed from the OLS model is that roughly 2% of the model predicted responses is less than 0. This result however is poor in practice because the smallest response value is 0 implying all cells killed. Thus, we find it intrigued to introduce a model that takes care of this issue. Thus, our choice for *Binomial Regression Model*, a member within the GLM family (**G**eneralized **L**inear **M**odels).

Suppose the response variable Y_i for $i = 1, ..., n$ is binomially distributed with $B(n,\, p_i)$ so that:

$$P(Y_i = y_i) = \binom{n}{y_i} p_i^{y_i} (1 - p_i)^{n - y_i}$$

And, assume that Y_i are independent, which is usually the case in the nature of most experiments. Each of the drug combinations that compose the response Y_i are all subject to the same q predictors $(x_{i1}, ..., x_{iq})$ which are drugs A, B and C in our case. We need a linear model that describes the relationship of $x_1, ..., x_q$ to p; hence, we need to construct a linear predictor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_q x_{iq}$$

Due to the nature of p_i which needs to be $0 \leq p_i \leq 1$, setting $\eta_i = p_i$ in general will not work well. Instead, we should use a *link function:* $\boldsymbol{g}$ which makes describing the relationship between η_i and p_i feasible, such that $\eta_i = \boldsymbol{g}(p_i)$. There are many choices for such $\boldsymbol{g}$ and all need to meet the following requirements: it must be monotone, and it must satisfy $0 \leq \boldsymbol{g}^{-1}(\eta) \leq 1$ for any η. Our choices are:

(1) Link Logit: $\eta_i = log \frac{p_i}{1 - p_i}$
(2) Link Probit: $\eta = \Phi^{-1}(p_i)$ where Φ^{-1} is the inverse normal CDF

The estimation method for GLM is different from OLS. The β parameters are estimated using the method of maximum likelihood. It involves an algorithm to

perform maximization which we will not discuss here. Further, there are more than one algorithm that gets the job done, and the one we utilized is called *Fisher Scoring Method*. For your own interest you can refer to [Far06] for further detail.

In a binomial regression model, we interpret our response, y_i a little differently than we would in an OLS model. Since y_i is really the proportion of cell survival and that 1 in a proportion is the maximum number of cell survival. The new perspective on viewing the data response should be: "Proportion of cell killed" $1 - y_i$ and "Proportion of cell survived" y_i. Lastly, the p_i that is related to parameter β's through η is a predicted probability of cell survival. The above notes may seen trivial, but they are important concepts that need to be clarified before building a binomial regression model because one must know the right input for a variable. If such step were done incorrectly, the parameters estimated will be nonsense.

Like the Response Transformation Model, to obtain $\tilde{y}_i$, a scaled model predicted response with observation index i; the $h^{-1}()$ step is required. We provide the inverse transformation below:

(1) If $h()$ is Link Logit, then the transformed back response is

$$\tilde{y}_i = \frac{exp(\widehat{y}_i)}{1 + exp(\widehat{y}_i)}$$

(2) If $h()$ is Link Probit, then the transformed back response is

$$\tilde{y}_i = Probability(Y_i \leq \widehat{y}_i) \quad generated \quad from \quad N(0,1)$$

To wrap up, we conclude this section with the general form of GLM. A GLM is defined by specifying two components. First, the response should be a member of exponential family distribution. Second, there should be a link function that describes how the mean of the response and a linear combination of the predictors

are related.

If the distribution of Y is a member from the exponential family, then it should have the following form:

$$f(y|\theta,\, \phi) = exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right]$$

The θ is called the *canonical parameter* and represents the location while ϕ is called the *dispersion parameter* and represents the scale. Various members of the family is defined by specifying functions a, b, and c.

Lastly, we show how a binomial response can be rewritten and is indeed a member of the exponential family.

$$\begin{aligned} f(y|\theta,\, \phi) =& \binom{n}{y}\mu^y(1-\mu)^{n-y} \\ =& exp\left[ylog\mu + (n-y)log(1-\mu) + log\binom{n}{y}\right] \\ =& exp\left[ylog\frac{\mu}{1-\mu} + nlog(1-\mu) + log\binom{n}{y}\right] \end{aligned}$$

where $\theta = log\frac{\mu}{1-\mu}$, $b(\theta) = -n\cdot log(1-\mu) = n\cdot log(1+exp\theta)$ and $c(y,\, \phi) = log\binom{n}{y}$

CHAPTER 3

Measure Criteria

3.1 R^2

Coefficient of Determination is a measure used to check a model's goodness of fit, with value closer to 1 indicating a better fit. It is defined as the following:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2}$$

Here $n = 512$, 80, 27 for n_{512}, n_{80}, and n_{27} respectively. Notice, since the definition for R^2 differ for a binomial regression model, we did not include it in Table 3.1. We find it may be injustice to compare it against others because it is not calibrated on the same basis like others.

Now, let us take a look at Table 3.1, *The R^2 Comparison Table*. We are really simultaneously considering three objectives in mind. How does number of predictors, size of dataset and model building method change the desired result? We will consider the listed objectives within each cell type then compare both against each other.

Normal Cell's R^2 Values

Increasing number of predictors has a positive impact on R^2 value for all methods generated with any data subset, as seen on the number change from model prototype g1 to model prototype g3. In general, model prototype g3 has the highest R^2 for all methods generated with any data subset. If we choose the highest 3

R^2 values within each data subset, we notice that response transformation with square root always makes it to the top 3.

Cancer Cell's R^2 Values

The same result like above is observed with only a slight lower magnitude on the R^2 in comparison to normal cell as seen in Table 3.1.

3.2 $\widehat{\sigma}$

$\widehat{\sigma}$ obtained from $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2}$ is the estimated standard deviation of the errors. This quantity serves as an indicator to the quality of β estimates obtained. However, its accuracy depends heavily on satisfactory error assumptions. It is defined with the following mathematical expression.

$$\widehat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}\widehat{\epsilon}_i^2 \quad where \quad \widehat{\epsilon}_i^2 = (\widehat{y}_i - y_i)^2$$

Here $n = 512, 80, 27$ for n_{512}, n_{80}, and n_{27} respectively. Since we easily have more than 50 models, it is infeasible to check and fix each individual model's regression diagnostics and its error assumptions. Instead, we simply assume the quality of $\widehat{\sigma}$ is acceptable. When decision of our best models are made, thorough diagnostics will be done for selected models.

This quantity often associates with statistical inferences; thus, a smaller $\widehat{\sigma}$ indicates a better precision. With that in mind, let's take a look at Table 3.2.

Normal Cell's $\widehat{\sigma}$ Values

Increasing number of predictors has a positive impact on $\widehat{\sigma}$ value for all meth-

Table 3.1: R^2 Comparison

	$Normal\,Cell$			$Cancer\,Cell$		
$Method$	$g1$	$g2$	$g3$	$g1$	$g2$	$g3$
$\underline{n_{512}}$						
OLS	0.8868	0.9583	0.9949	0.7543	0.8662	0.9741
Log	0.8987	0.9488	0.9702	0.8223	0.9344	0.9454
$Logit$	0.8224	0.8748	0.9485	0.8064	0.9231	0.9612
$Sqrt$	0.9348	0.9713	0.9937	0.8302	0.9305	0.9826
$\underline{n_{80}}$						
OLS	0.9198	0.9657	0.9966	0.8474	0.8996	0.9844
Log	0.8708	0.9323	0.9694	0.9177	0.9618	0.9715
$Logit$	0.8271	0.8689	0.9477	0.8944	0.9410	0.9776
$Sqrt$	0.9362	0.9724	0.9957	0.9152	0.9498	0.9912
$\underline{n_{27}}$						
OLS	0.8241	0.9726	0.9788	0.6902	0.9120	0.9323
Log	0.8774	0.9912	0.9940	0.7500	0.9498	0.9546
$Logit$	0.8480	0.9287	0.9500	0.7289	0.9368	0.9482
$Sqrt$	0.9099	0.9844	0.9879	0.7563	0.9392	0.9450

ods generated with any data subset, notice how the number changes from model prototype g1 to model prototype g3. Model prototype g3 has the smallest $\widehat{\sigma}$ for all methods generated with any data subset. If we choose the smallest $\widehat{\sigma}$ value within each data subset, we notice that response transformation with square root is always the choice.

Cancer Cell's $\widehat{\sigma}$ Values

The same result like the above is observed except with $\widehat{\sigma}$ values that are twice in magnitude in comparison to normal cell for some models as seen in Table 3.2.

3.3 $RMSE$

RMSE stands for **R**oot **M**ean **S**quare **E**rror, it measures how well a model performed in predicting the observation. It is defined as the following:

$$\sqrt{\frac{\sum_{i=1}^{n}(\tilde{y}_i - y_i)^2}{n}}$$

where y_i is an observed response, $\tilde{y}_i$ is a scaled model predicted response, and $n = 512$. A model which gives a better prediction performance than the other model will have a smaller $RMSE$ value. Now let's look at Table 3.3

Normal Cell's $RMSE$

Increasing number of predictors has a positive impact on $RMSE$ value for all methods generated with n_{512} and n_{27}. $RMSE$ decreases from model prototype g1 to model prototype g3. Model prototype g3 has the lowest $RMSE$ for all methods generated with these two data subsets. Also, if we choose the lowest $RMSE$ value within each of the three data subsets, we notice that response transformation with square root is always the choice. Unlike the others, Log n_{80} shows an unfavorable trend whilst g3 has $RMSE = 0.0649$, g1 has $RMSE = 0.0572$, and g2 has $RMSE = 0.0463$.

Table 3.2: $\widehat{\sigma}$ Comparison

	$Normal\,Cell$			$Cancer\,Cell$		
$Method$	$g1$	$g2$	$g3$	$g1$	$g2$	$g3$
$\underline{n_{512}}$						
OLS	0.1047	0.0637	0.0223	0.1683	0.1245	0.0550
Log	0.2745	0.1957	0.1499	0.7842	0.4779	0.4372
$Logit$	0.8472	0.7135	0.4588	1.1671	0.7377	0.5259
$Sqrt$	0.0596	0.0396	0.0186	0.1231	0.0790	0.0396
$\underline{n_{80}}$						
OLS	0.0873	0.0582	0.0187	0.1359	0.1125	0.0453
Log	0.2749	0.2030	0.1394	0.5557	0.3865	0.3408
$Logit$	0.8107	0.7201	0.4647	0.8907	0.6794	0.4274
$Sqrt$	0.0568	0.0382	0.0154	0.0905	0.0710	0.0303
$\underline{n_{27}}$						
OLS	0.1413	0.0598	0.0571	0.1797	0.1027	0.0977
Log	0.4798	0.1376	0.1238	1.2262	0.5891	0.6078
$Logit$	0.8841	0.6493	0.5898	1.5818	0.8190	0.8038
$Sqrt$	0.0859	0.0383	0.0365	0.1607	0.0860	0.0888

Cancer Cell's $RMSE$

Again, increasing number of predictors has a positive impact on $RMSE$ value for all methods generated with n_{512} and n_{27} but Log Transformation . $RMSE$ decreases from model prototype g1 to model prototype g3. Model prototype g3 has the lowest $RMSE$ for all methods generated with these two data subsets with an exception on method Log Transformation. Unlike normal cell, if we choose the lowest $RMSE$ value within each of the three data subsets, we notice that the method of choice varies, but it is likely between Link Logit, Link Probit or Sqrt. Lastly, $RMSE$ from Log, Link Logit, and Link Probit built with n_{80} show random trend and no association between number of predictors and method of model generation.

3.4 $\widehat{R}$

$\widehat{R}$ served as a measure of the correlation between the observed response y_i and a scaled model predicted response $\tilde{y}_i$. We defined it such that it measures the strength of linear dependence on the matter of our interest: y_i response from observation, $\tilde{y}_i$ response from a scaled model prediction. It is computed with the following expression through statistical software R:

$$\widehat{R} = \frac{\sum_{i=1}^{n}[(y_i - \overline{y})(\tilde{y}_i - \overline{\tilde{y}})]}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{(\sum_{i=1}^{n}\tilde{y}_i - \overline{\tilde{y}})^2}}$$

Where $n = 512$, and with a $\widehat{R}$ closer to 1 indicating a better prediction performance from a model, now let's look at Table 3.4.

Normal Cell's $\widehat{R}$ Values

Increasing number of predictors in a model also increases $\widehat{R}$ value for all methods from all 3 data subsets, with one exception: method Log Transformation from n_{80}. It shows no improvement from model g1 to g2 to g3. In most cases, g3 is required

Table 3.3: $RMSE$ Comparison

	$Normal\,Cell$			$Cancer\,Cell$		
$Method$	$g1$	$g2$	$g3$	$g1$	$g2$	$g3$
$\underline{n_{512}}$						
OLS	0.1764	0.1280	0.0838	0.1676	0.1237	0.0544
Log	0.0580	0.0392	0.0427	0.1535	0.0703	0.0738
$Logit$	0.1044	0.0973	0.0509	0.1249	0.0901	0.0436
$Sqrt$	0.0628	0.0438	0.0172	0.1213	0.0797	0.0274
$LinkLogit$	0.0803	0.0758	0.0372	0.0951	0.0755	0.0298
$LinkProbit$	0.0813	0.0742	0.0338	0.1133	0.0764	0.0301
$\underline{n_{80}}$						
OLS	0.1105	0.0746	0.0252	0.1764	0.1280	0.0838
Log	0.0572	0.0463	0.0649	0.0829	0.6181	0.4689
$Logit$	0.1051	0.1010	0.0599	0.0991	0.1225	0.1082
$Sqrt$	0.0582	0.0481	0.0239	0.0967	0.0792	0.0595
$LinkLogit$	0.0831	0.0794	0.0422	0.0863	0.0792	0.0957
$LinkProbit$	0.0827	0.0777	0.0388	0.0927	0.0839	0.0971
$\underline{n_{27}}$						
OLS	0.1401	0.0703	0.0341	0.2054	0.1360	0.0763
Log	0.1131	0.0434	0.0433	0.3644	0.0800	0.1152
$Logit$	0.1187	0.0933	0.0583	0.2021	0.0954	0.0621
$Sqrt$	0.1145	0.0517	0.0224	0.2069	0.1007	0.0557
$LinkLogit$	0.0987	0.0795	0.0430	0.1495	0.0826	0.0347
$LinkProbit$	0.1059	0.0770	0.0377	0.1637	0.0910	0.0367

to obtain a 99 % correlation. If we choose the top 3 highest values, square root transformation is always included

. Cancer Cell's $\widehat{R}$ Values

Like what is observed in normal cell, method of Log Transformation from n_{80} is the only exception. This one exhibits an odd trend where g1 has $\widehat{R} = 0.9709$, g2 has $\widehat{R} = 0.4391$, and g3 has $\widehat{R} = 0.5763$.

3.5 The $\widehat{R}$ Plots

The $\widehat{R}$ Plots are scatterplots. We plot observed response, y_i , on the horizontal axis and model scaled model predicted response, $\tilde{y}_i$, on vertical axis and see how well this scatterplot resembles a linear relationship.

We show the layout of a typical $\widehat{R}$ Plots in Table 3.5 to indicate what each panel in a $\widehat{R}$ Plots from Figure 3.1 to Figure 3.12 represents. You can easily locate the model building method from the caption of each $\widehat{R}$ plots.

Recall the numbers we observed in Table 3.4, making scatter plots and showing association between y_i and $\tilde{y}_i$ for $i = 1, .., 512$ reveal much more information than the numbers alone.

Normal Cell's $\widehat{R}$ Plots

There are six $\widehat{R}$ Plots, each for the following methods: OLS, Log, Logit, Sqrt, Link Logit and Link Probit. Despite the fact that model prototype g3 built from some methods within each data subset can achieve a 99% correlation between scaled model predicted responses and observed responses, there is a concern on outliers as seen in Figure 3.1 to Figure 3.6. With n_{80} giving many remote data points away from the linear line, some outliers observed for n_{512} and rarely observed for

Table 3.4: $\widehat{R}$ Comparison

	$Normal\,Cell$			$Cancer\,Cell$		
$Method$	$g1$	$g2$	$g3$	$g1$	$g2$	$g3$
$\underline{n_{512}}$						
OLS	0.9417	0.9789	0.9975	0.8685	0.9307	0.9870
Log	0.9836	0.9921	0.9917	0.8936	0.9816	0.9826
$Logit$	0.9527	0.9590	0.9886	0.9346	0.9667	0.9924
$Sqrt$	0.9821	0.9901	0.9985	0.9502	0.9763	0.9967
$LinkLogit$	0.9407	0.9687	0.9920	0.8529	0.8493	0.9216
$LinkProbit$	0.9408	0.9725	0.9936	0.8614	0.8587	0.9433
$\underline{n_{80}}$						
OLS	0.9403	0.9720	0.9967	0.8646	0.9258	0.9694
Log	0.9842	0.9896	0.9820	0.9709	0.4391	0.5763
$Logit$	0.9506	0.9564	0.9830	0.9583	0.9387	0.9522
$Sqrt$	0.9827	0.9878	0.9972	0.9652	0.9746	0.9846
$LinkLogit$	0.9404	0.9506	0.9883	0.8432	0.8706	0.9147
$LinkProbit$	0.9398	0.9587	0.9909	0.8522	0.8747	0.9200
$\underline{n_{27}}$						
OLS	0.9383	0.9774	0.9952	0.8653	0.9266	0.9789
Log	0.9487	0.9907	0.9902	0.4190	0.9719	0.9449
$Logit$	0.9324	0.9610	0.9825	0.8243	0.9635	0.9830
$Sqrt$	0.9622	0.9885	0.9982	0.8838	0.9683	0.9927
$LinkLogit$	0.9382	0.9512	0.9741	0.8632	0.8333	0.8880
$LinkProbit$	0.9362	0.9610	0.9822	0.8598	0.8576	0.9099

Table 3.5: $\widehat{R}$ Plot Panel Interpretation

g_1 n_{512}	g_2 n_{512}	g_3 n_{512}
g_1 n_{80}	g_2 n_{80}	g_3 n_{80}
g_1 n_{27}	g_2 n_{27}	g_3 n_{27}

n_{27}.

Cancer Cell's $\widehat{R}$ Plots

Similar problem is observed here as seen in Figure 3.7 to Figure 3.12. Many show obvious flaw in a model's prediction performance. Even though $\widehat{R}$ values seen to be satisfactory, the $\widehat{R}$ Plots speak otherwise. Lastly, prototype g3 build with n_{80} and any method gives many outliers in comparison to n_{512} and n_{27}.

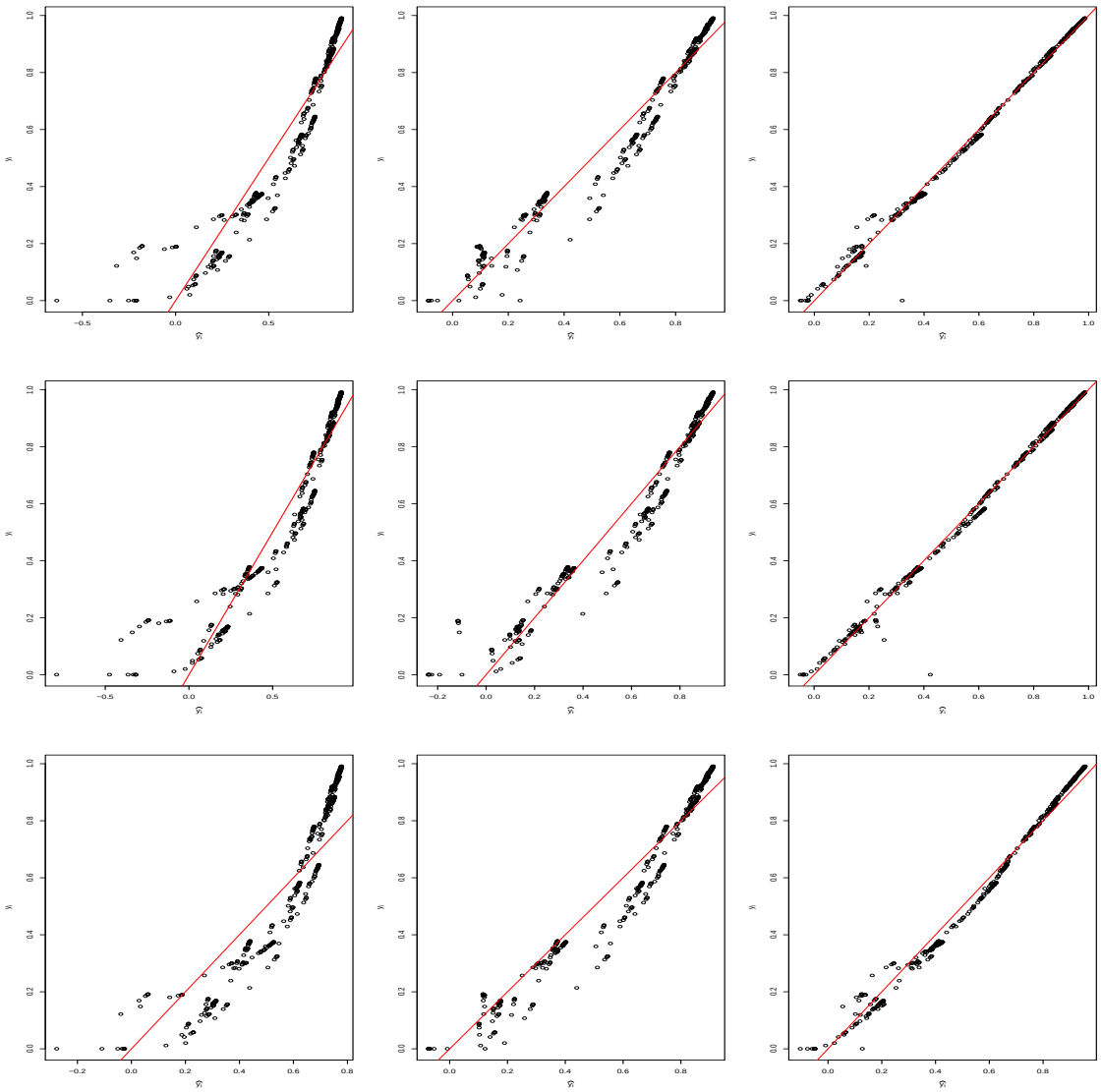


Figure 3.1: OLS $\widehat{R}$ Plots for Normal Cells

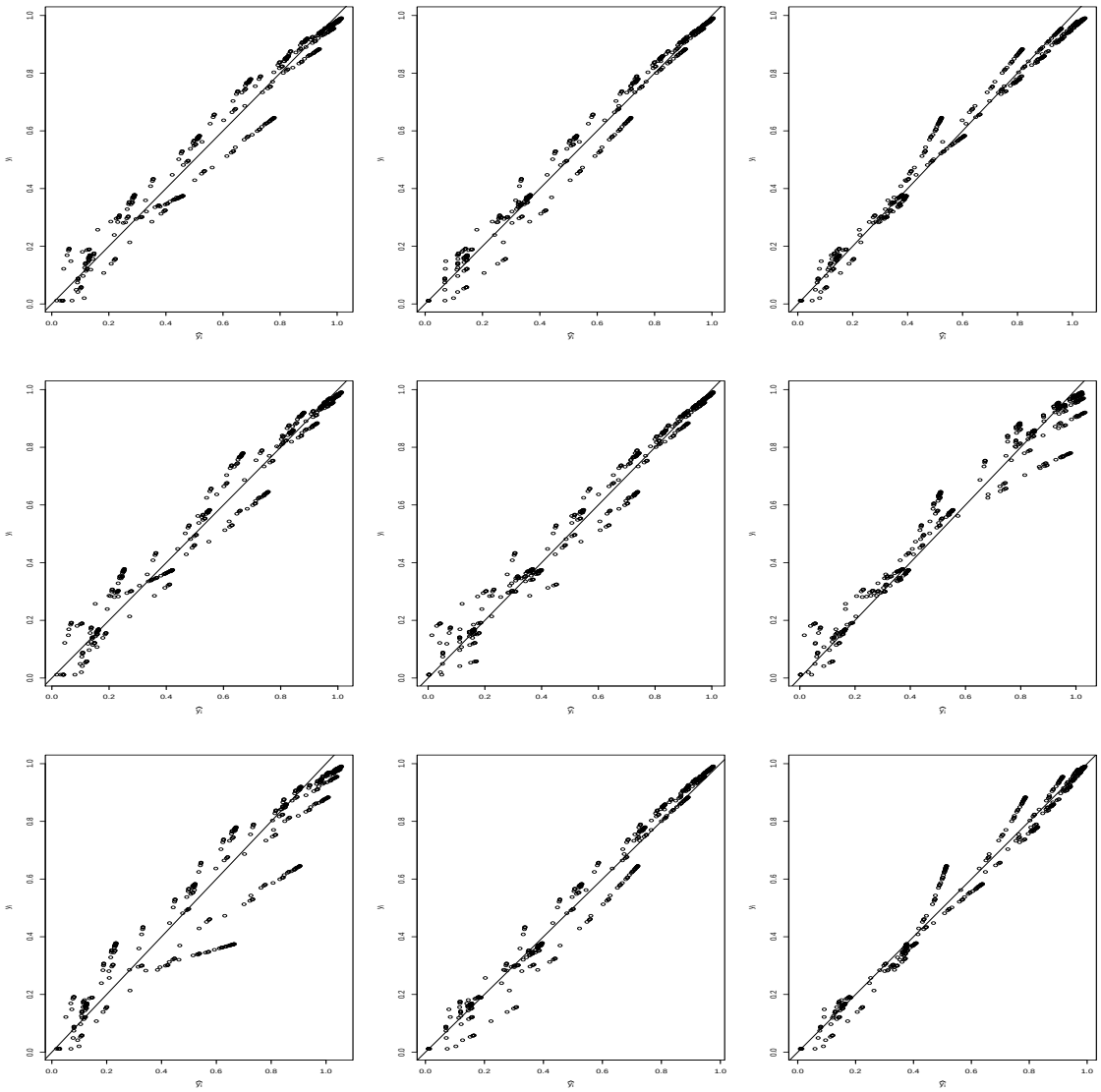


Figure 3.2: Log Transformation $\widehat{R}$ Plots for Normal Cells

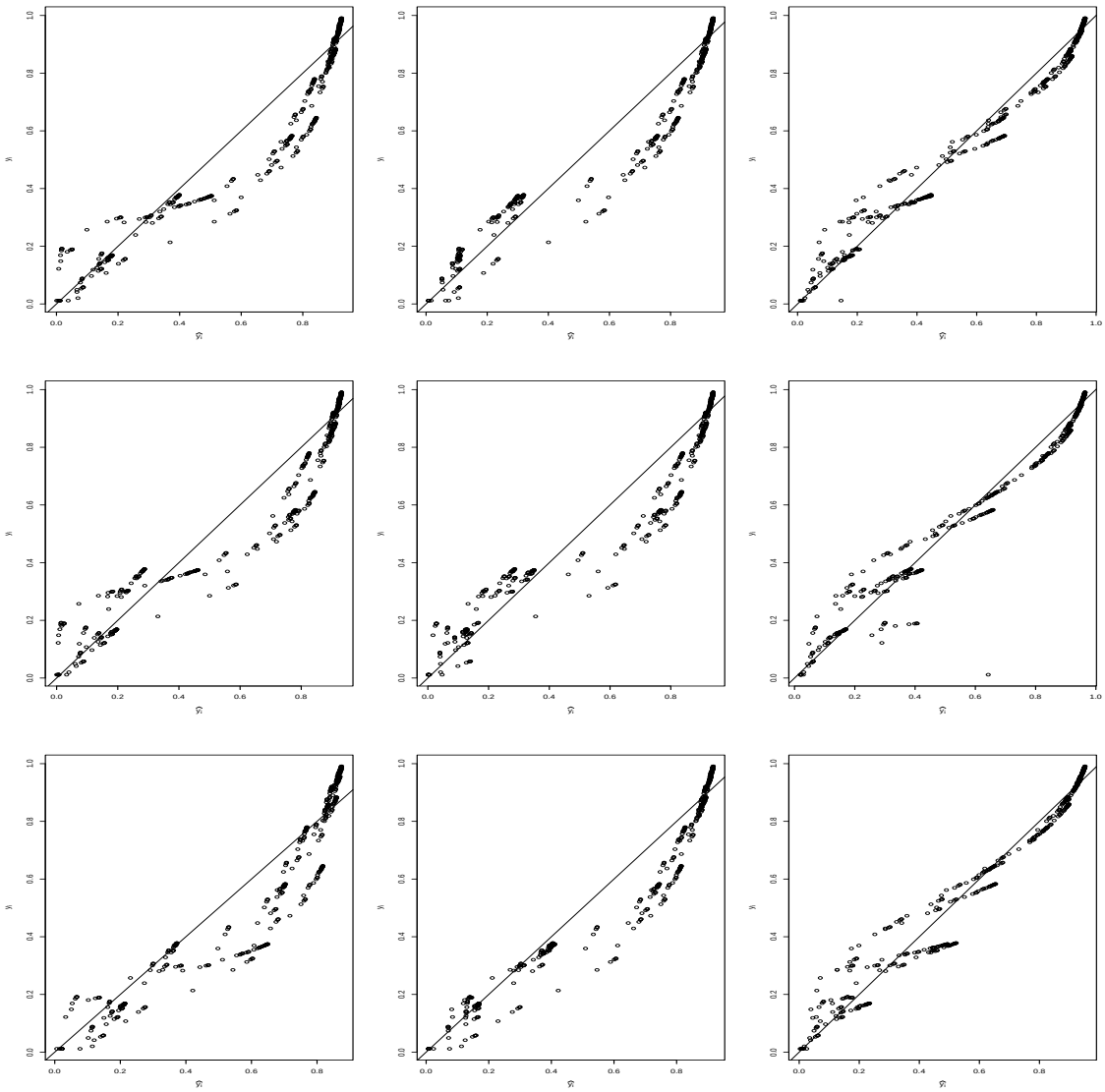


Figure 3.3: Logit Transformation $\widehat{R}$ Plots for Normal Cells

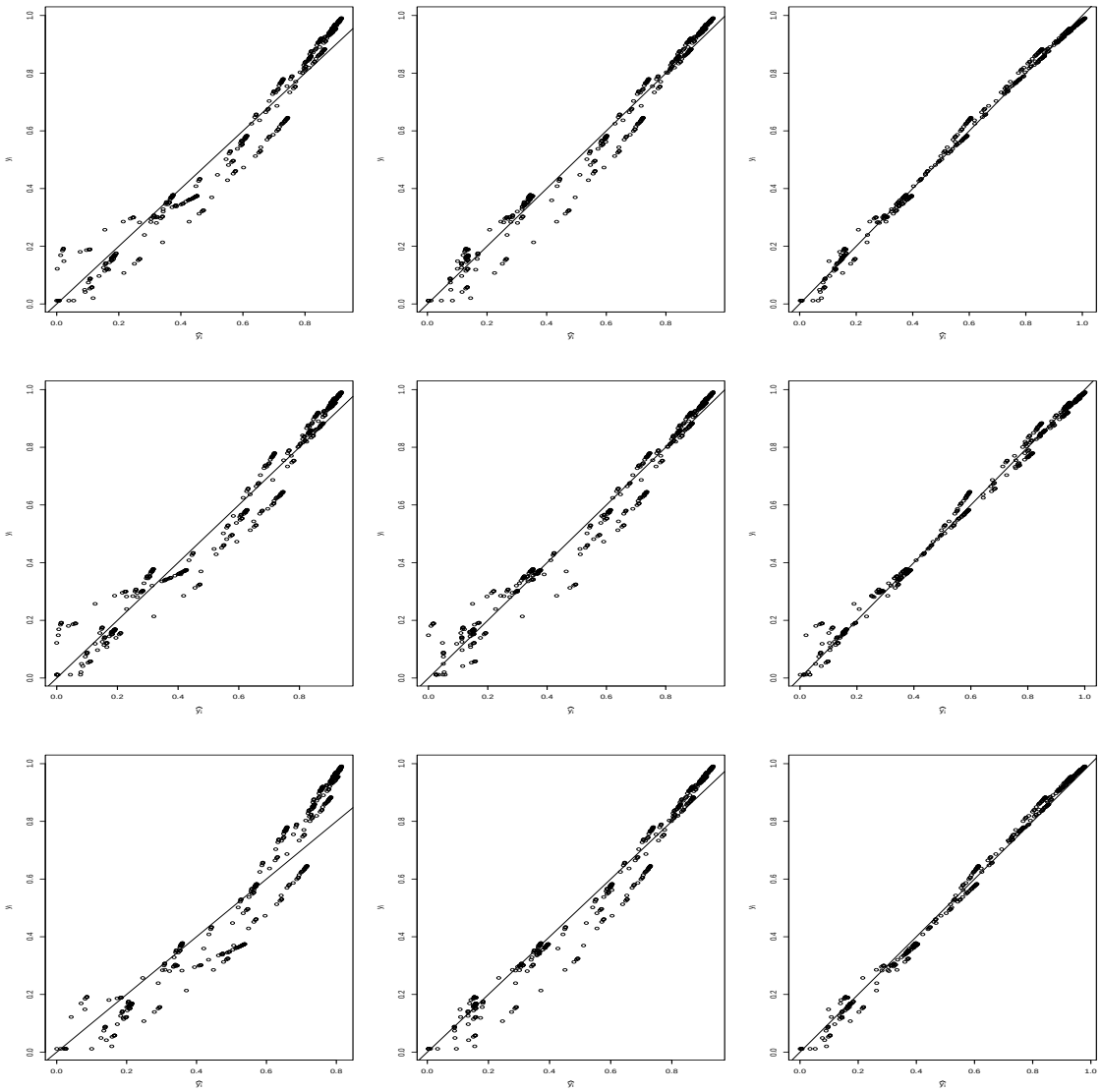


Figure 3.4: Sqrt Transformation $\widehat{R}$ Plots for Normal Cells

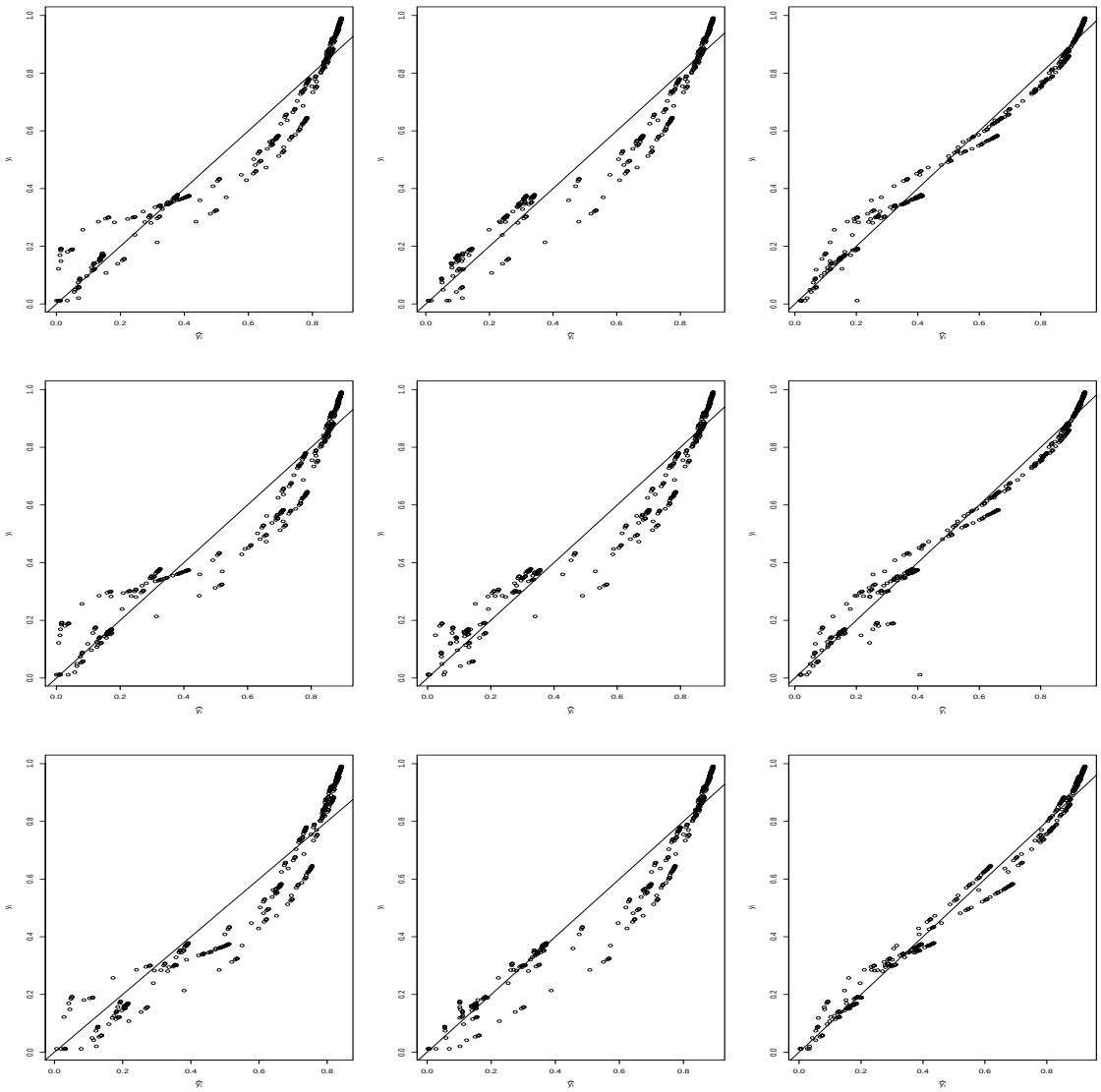


Figure 3.5: Binomial Error Link Logit $\widehat{R}$ Plots for Normal Cells

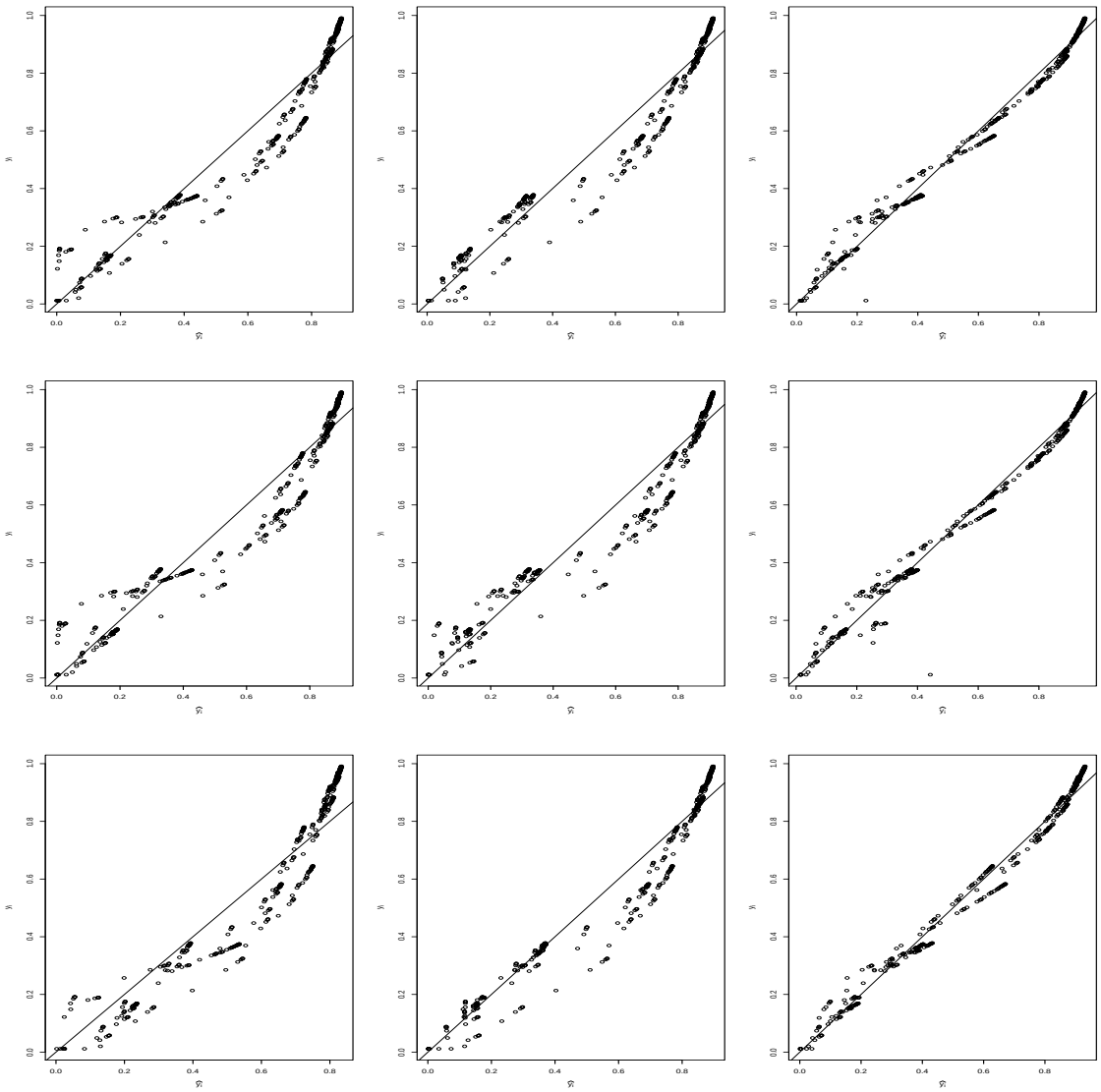


Figure 3.6: Binomial Error Link Probit $\widehat{R}$ Plots for Normal Cells

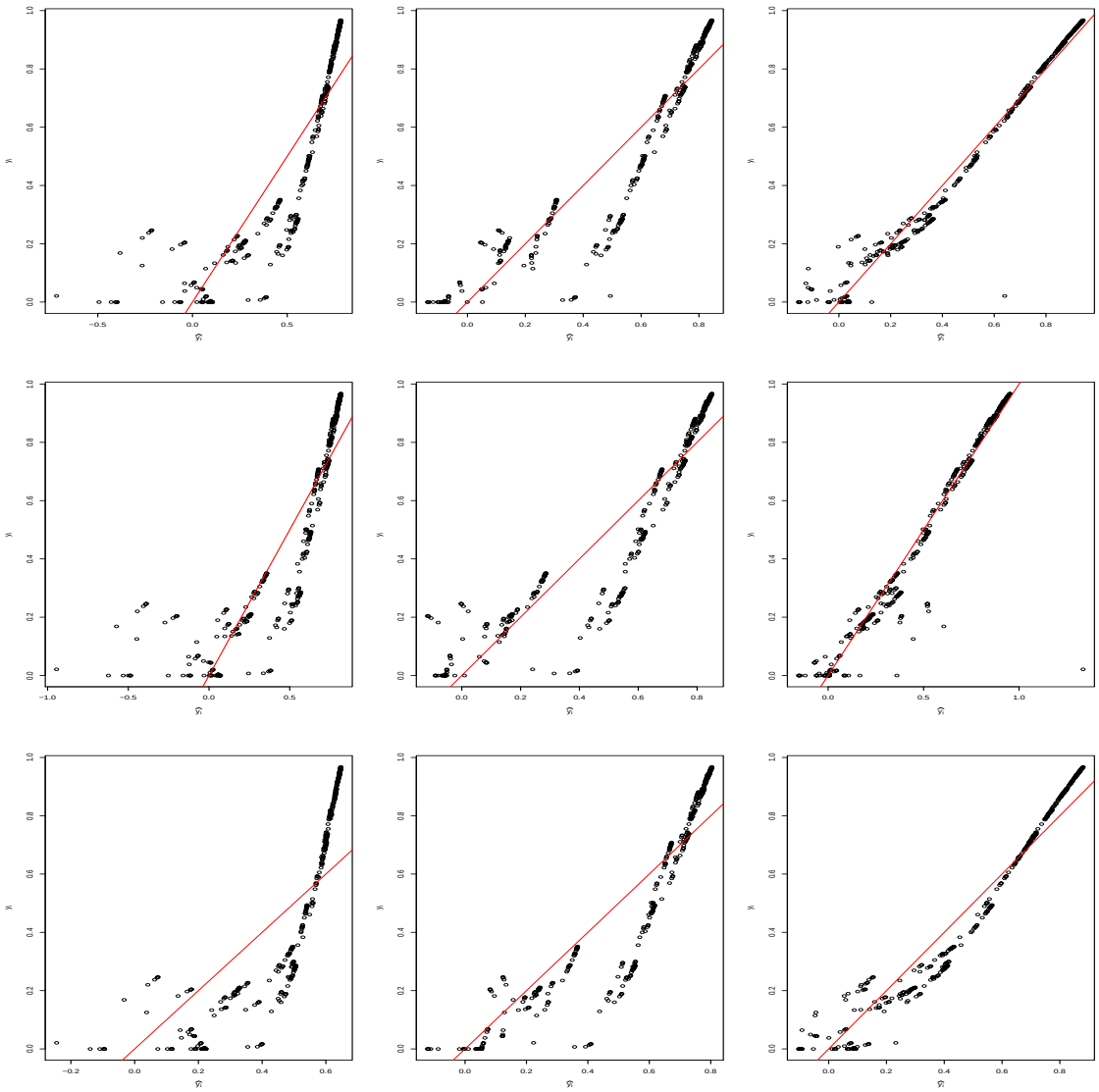


Figure 3.7: OLS $\widehat{R}$ Plots for Cancer Cells

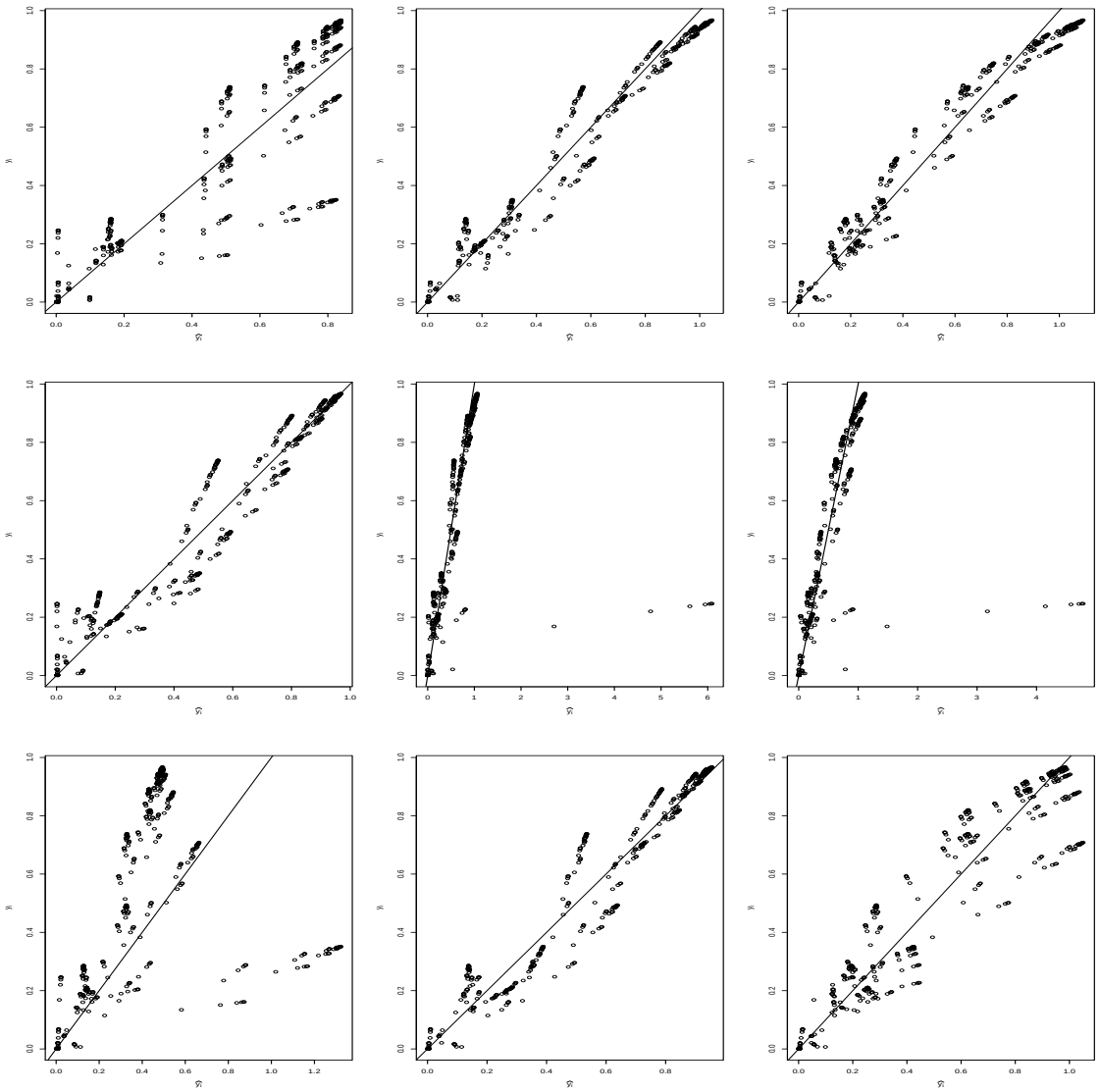


Figure 3.8: Log Transformation $\widehat{R}$ Plots for Cancer Cells

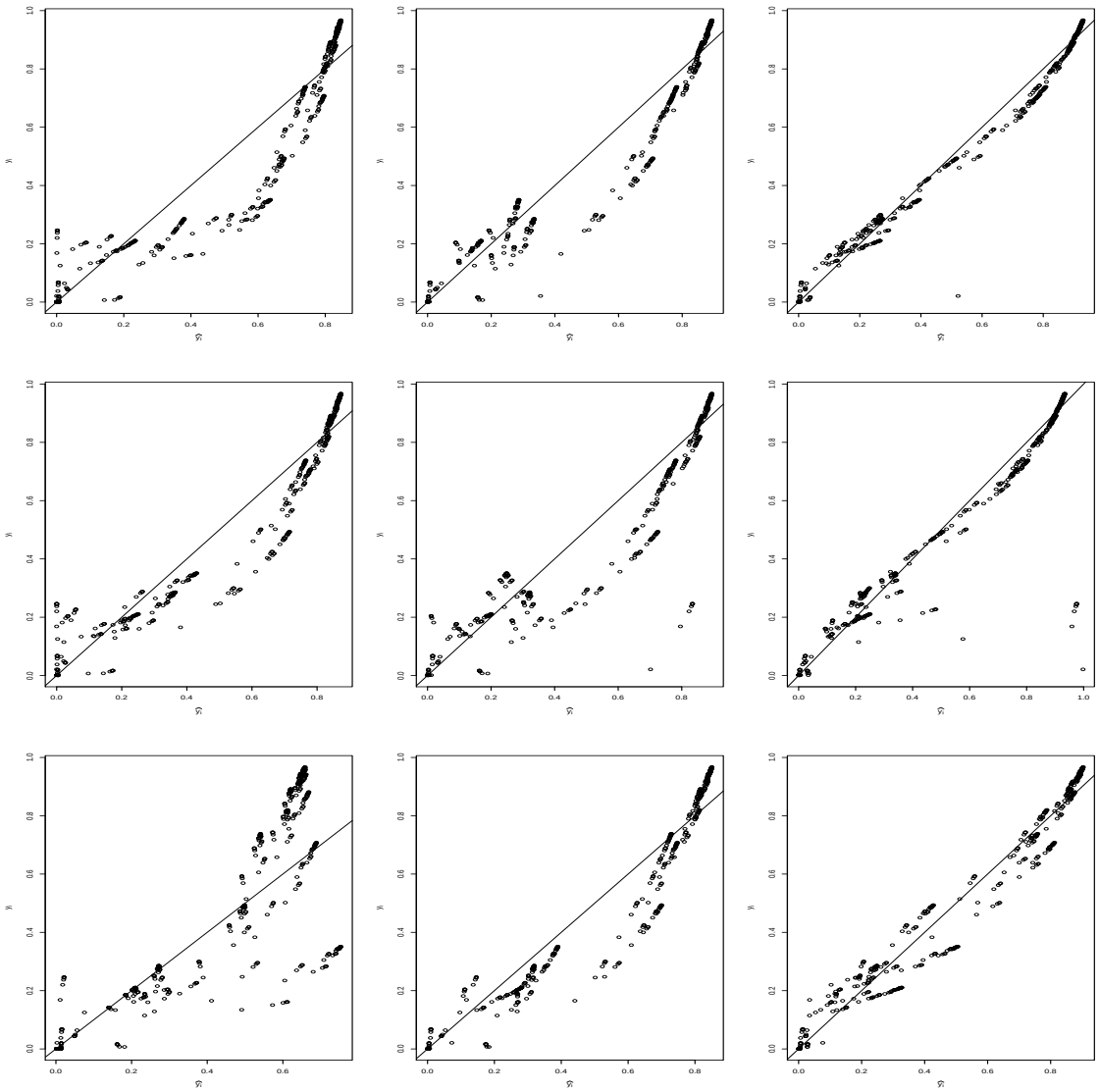


Figure 3.9: Logit Transformation $\widehat{R}$ Plots for Cancer Cells

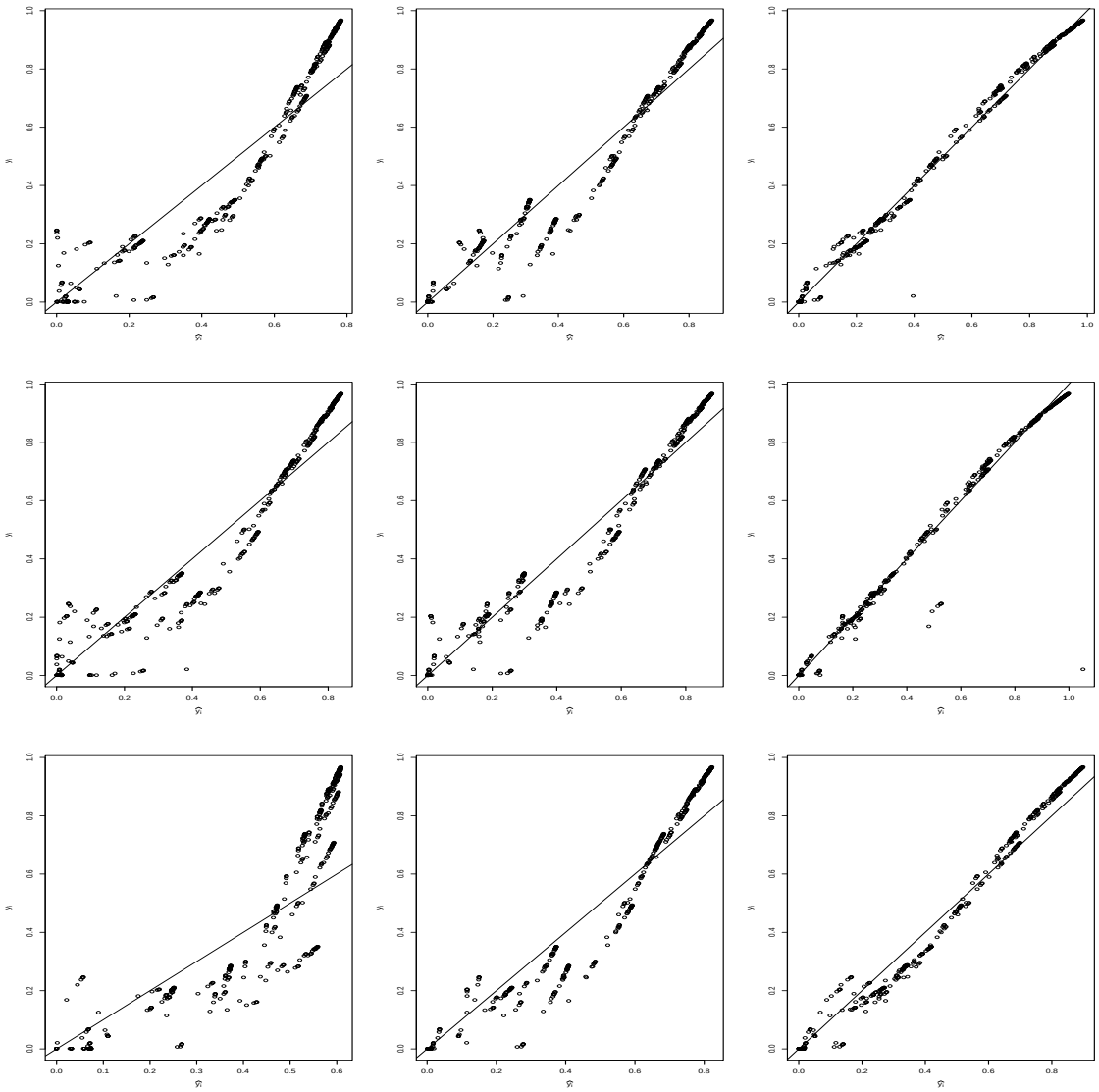


Figure 3.10: Sqrt Transformation $\widehat{R}$ Plots for Cancer Cells

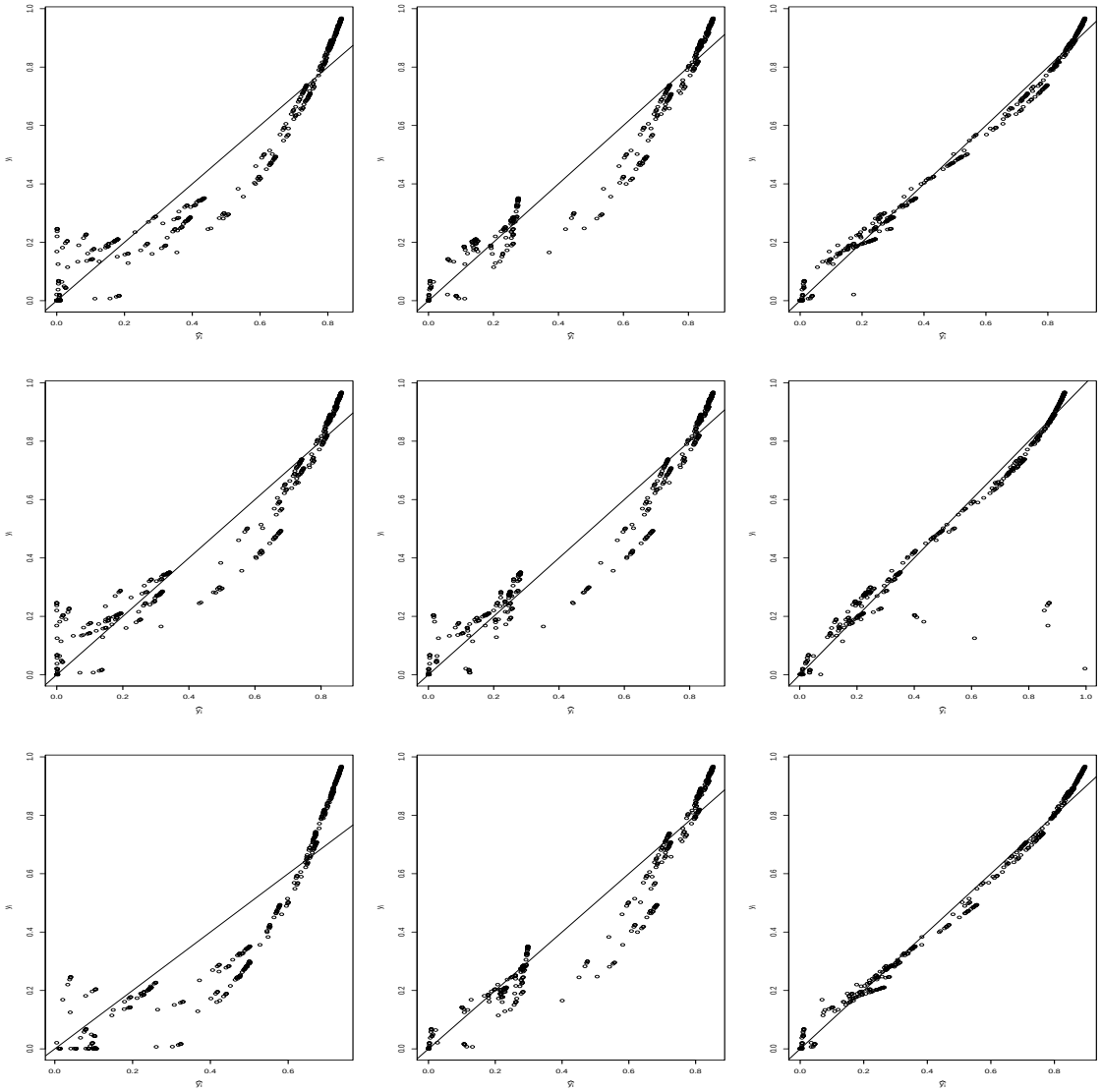


Figure 3.11: Binomial Error Link Logit $\widehat{R}$ Plots for Cancer Cells

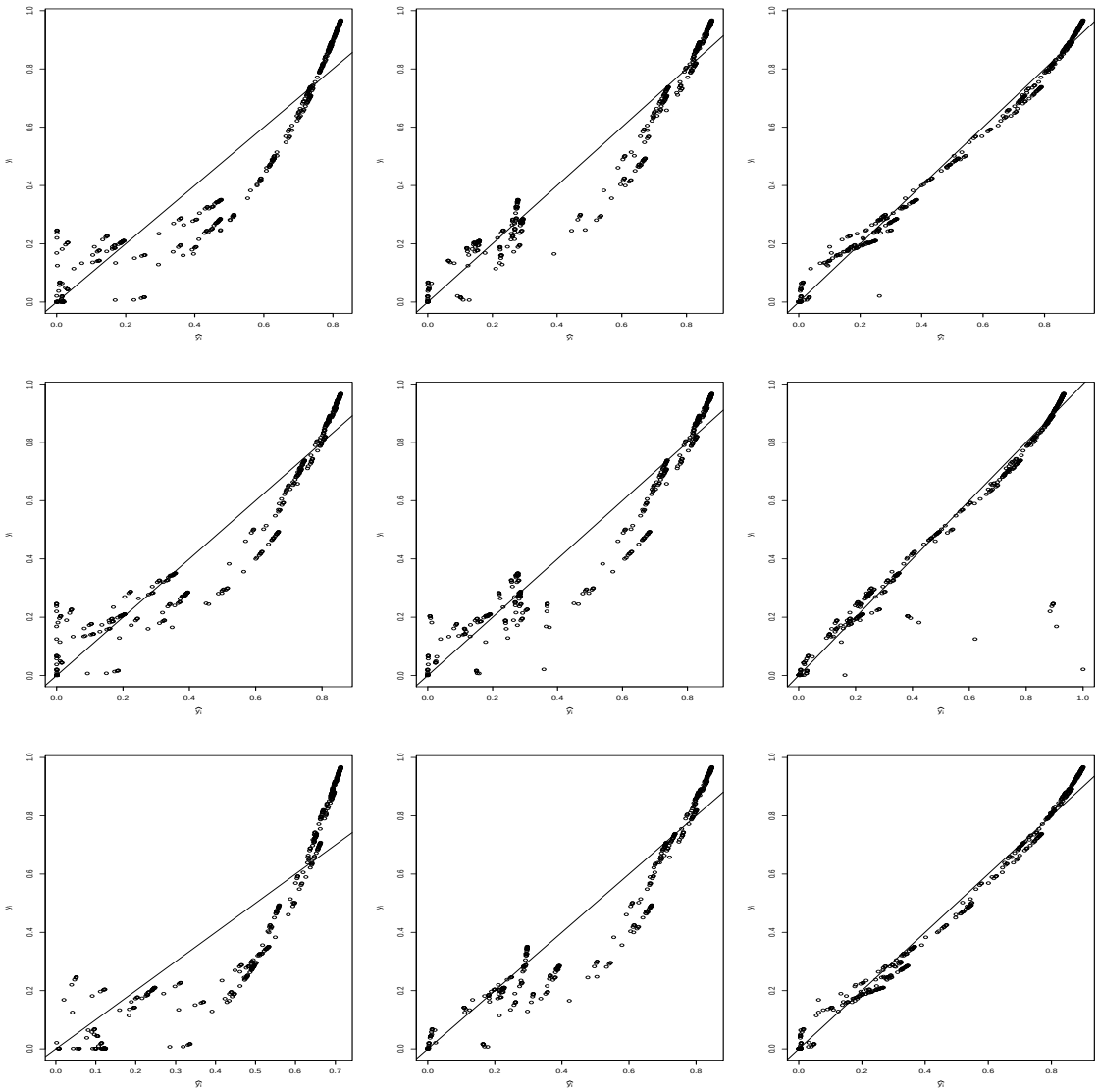


Figure 3.12: Binomial Error Link Probit $\widehat{R}$ Plots for Cancer Cells

CHAPTER 4

Discussion

4.1 Model Prototype Comparison

Through the combination from various methods and data subsets, we built our prototypes $g1$, $g2$, and $g3$. Simply interpret Table 3.1 to Table 3.4 by columns, you can see, in all cases, prototype $g3$ gives better goodness of fit and prediction performance than $g1$ and $g2$; hence, we believe building an effective model requires at least first order, 2-way interaction and quadratic predictors so that data structure can be captured.

Furthermore, model constructed with square root transformation on the response appear to give a good $\widehat{R}$ value and a satisfactory $\widehat{R}$ Plots when built using data subset n_{512} and n_{27}. The problem with n_{80} is despite the impressive $\widehat{R}$ values, it gives many outliers on the $\widehat{R}$ Plots.

Thus, we believe models constructed using square root transformation could be valuable in describing the relationship between cells and drugs in a lung cancer treatment. However, it's worth to note that all models are capable of describing the relationship between normal cells and drugs much better than between cancer cells and drugs.

4.2 Subset Comparison

Now if we look at Table 3.1 to Table 3.4 by blocks: n_{512}, n_{80}, and n_{27}. You may have a second thought on the claim from [AYF11] regarding what is an appropriate number of drug combinations that should be used to construct an effective model. We argue that a factorial design with 27-runs appears to be more powerful than random selection. We make our argument based on our observations below.

First, let's look at normal cells, method Log Transformation for generating $g1$, $g2$, $g3$ from n_{80} shows no improvement in $\widehat{R}$ values. And, in cancer cells, models built from n_{80} give unfavorable $\widehat{R}$ values. Increment of predictors from prototype $g1$ to $g3$ results decrease in $\widehat{R}$ value from method Log Transformation. And, from method Logit Transformation, we observe a fluctuating $\widehat{R}$ value from prototype $g1$ to $g2$.

Second, careful inspection of the $\widehat{R}$ Plots reveal some problems: despite the promising number shown on the $\widehat{R}$ Table within n_{80}, models built from n_{80} suffer from severe outlier problems. Thus, we promote careful selection of drug combinations and factorial design when building a model.

4.3 Best Model Error Assumption Diagnostics

We have chosen our best models to be g3 built from Response Transformation Sqrt with n_{27}. Thus, we want to make sure $\epsilon \sim N(0, \sigma^2 I)$ is indeed valid, and we check the following: (1) errors are independent, (2) errors have equal variance, (3) errors are normally distributed. Figures 4.1 and 4.2 show no significant departures from the assumptions.

4.4 Conclusion

We find drug effects appears to be nonlinear and non-additive; there are complex drug interactions. Factorial design is more effective than random selection of runs. A second order model with square root transformation on the response fits the lung cancer data the best among the models we fitted. Cancer cell's interaction with drugs is more difficult to model than normal cell's.

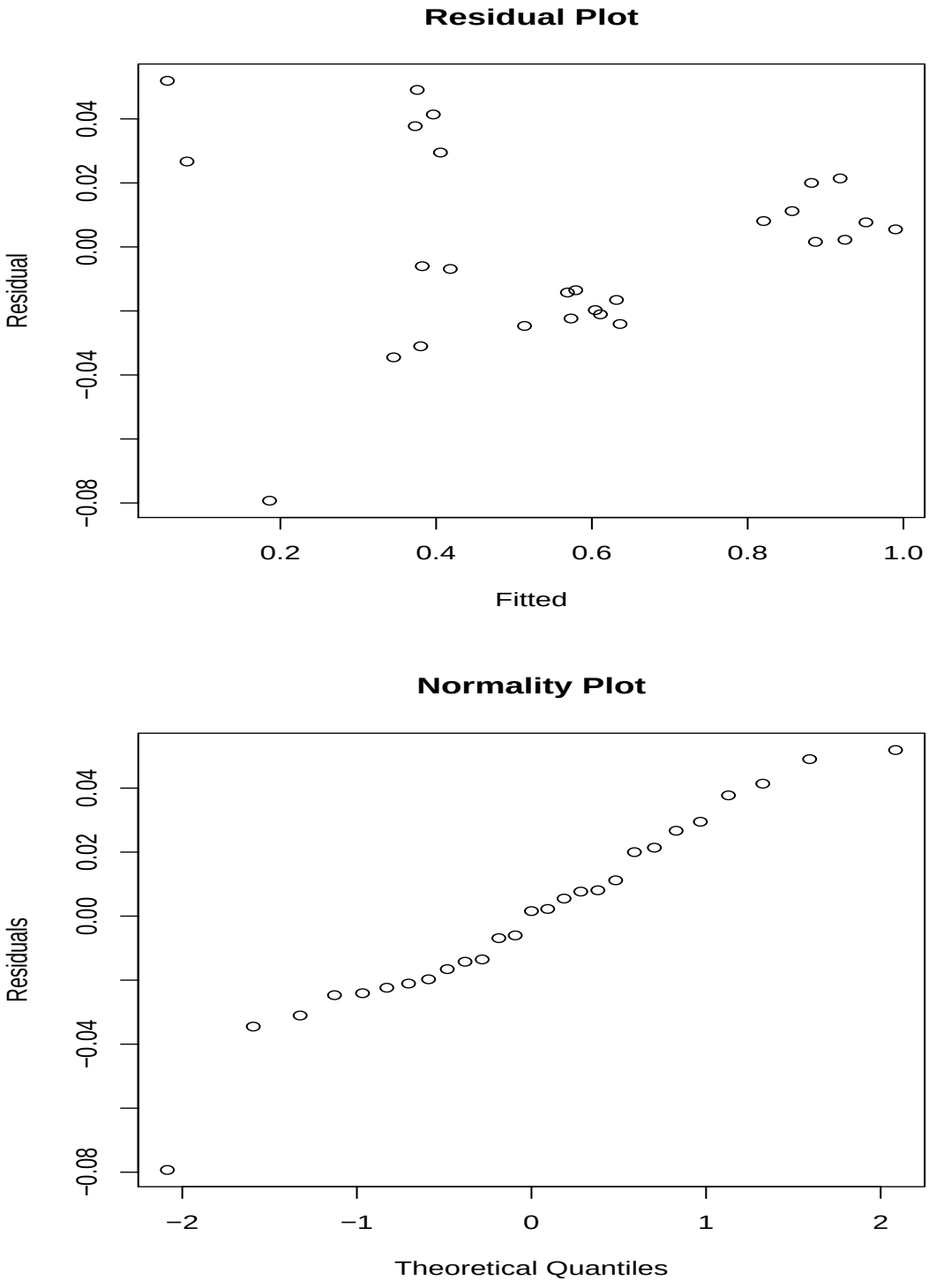


Figure 4.1: Model Diagnostics yn with Prototype g3, Method Sqrt, Subset n_{27}

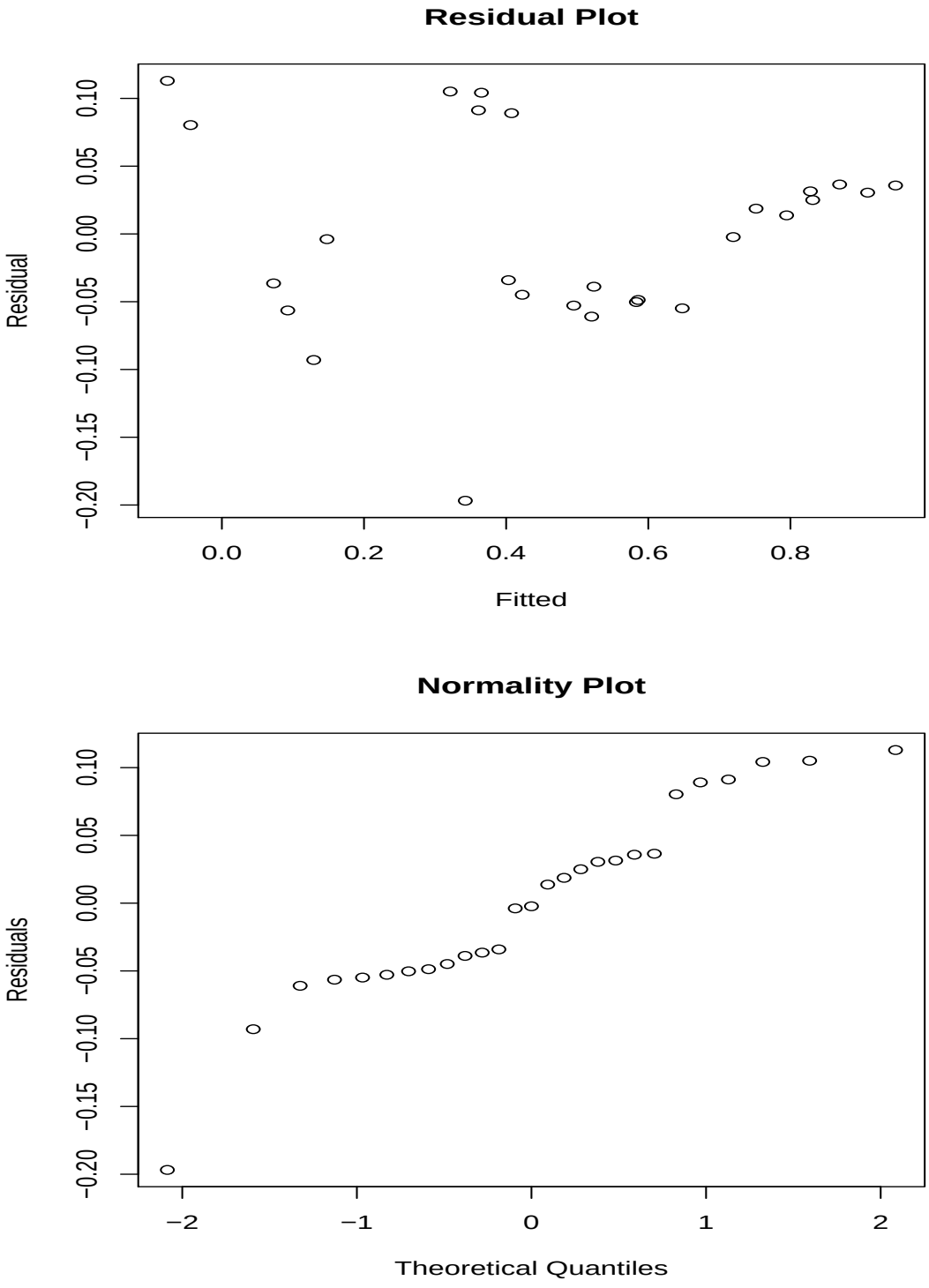


Figure 4.2: Model Diagnostics yc with Prototype g3, Method Sqrt, Subset n_{27}

References

[AYF11] Ibrahim Al-Shyoukh, Fuqu Yu, Jiayin Feng, Karen Yan, Steven Dubinett, Chih-Ming Ho, Jeff S Shamma, and Ren Sun. "Systematic quantitative characterization of cellular responses induced by multiple signals." *BMC Systems Biology*, **5**(88), May 2011.

[DXH12] Xianting Ding, Hongquan Xu, Chanelle Hopper, Jian Yang, and Chih-Ming Ho. "Use of Fractional Factorial Designs in Antiviral Drug Studies." *Quality and Reliability Engineeringn International*, Feburary 2012.

[EV11] Gerard I. Evan and Karen H. Vousden. "Proliferation, cell cycle and apoptosis in cancer." *Nature*, **411**:342–348, May 2011.

[Far05] Julian J Faraway. *Linear Models with R.* Taylor & Francis Group, LLC, Boca Raton, Florida, 2005.

[Far06] Julian J Faraway. *Extending the Linear Model with R.* Taylor & Francis Group, LLC, Boca Raton, Florida, 2006.

[Mon05] Douglas C. Montgomery. *Design and Analysis of Experiments.* John Wiley & Sons, Inc., Hoboken, New Jersey, sixth edition, 2005.

[MS03] William Mendenhall and Terry Sincich. *A Second Course in Statistics Regression Analysis.* Pearson Education, Inc., Upper Saddle River, New Jersey, sixth edition, 2003.

[WH09] C. F. Jeff Wu and Michael S. Hamada. *Experiments Planning, Analysis and Optimization.* John Wiley & Sons, Inc., Hoboken, New Jersey, second edition, 2009.