Title
Analysis of a Bladder Cancer Experiment

Permalink
https://escholarship.org/uc/item/15j060b0

Author
Yin, Kevin

Publication Date
2013

Peer reviewed|Thesis/dissertation
Analysis of a Bladder Cancer Experiment

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Kevin Yin

2013
The effectiveness of cancer treatments has always been an issue to both doctors and patients, due to various types of cancer cells as well as their ability to mutate and replicate quickly. In this study, we mainly focus on the analysis of a bladder cancer experiment on the effectiveness of different bladder cancer cell treatments. The experiment has 125 combinations with 8 drugs in total, each with 6-10 dosage levels in result of treating 6 different bladder cancer cells, and the experimenter recorded the percentage of cancer cells remaining after the treatment. The experiment was performed in 6 batches. The purpose was to find the optimal drug combination in treating different cancer cells. The lower the percentage remains in a cancer cell, the better the combination is. The analysis procedure involves a variety of models and transformations, including linear model, generalized linear model with binomial, neural network and transformations such as logarithm, square root on the responses, and the results were compared in terms of the $R^2$, $MSE$ and the actual vs. predicted plots. The best model in terms of the fitness and the prediction accuracy is a second degree linear model with each batch being treated as a block.
The thesis of Kevin Yin is approved.

Nicolas Christou

Yingnian Wu

Hongquan Xu, Committee Chair

University of California, Los Angeles
2013
To my parents . . .

friends and the wonderful Statistics department

here at UCLA - all these could not have been done

without you . . .
# Table of Contents

1 Introduction .................................................. 1

2 Data ............................................................. 3
   2.1 Overview .................................................. 3
   2.2 Data Selection ............................................ 5

3 Method .......................................................... 7
   3.1 Linear Model ............................................... 7
   3.2 Generalized Linear Model - Binomial .................. 9
   3.3 Cross Validation - Prediction ......................... 10
   3.4 Neural Network ........................................... 11

4 Results ........................................................ 12
   4.1 Tables & Plots ............................................ 12
   4.2 Neural Network ........................................... 28

5 Conclusion ..................................................... 30
   5.1 Review of Procedure .................................... 30
   5.2 Results .................................................... 30

References ..................................................... 31
LIST OF FIGURES

1.1 Location of Bladder and the staging of Bladder Cancer  . . . . . . 2

2.1 A sample view of the actual data  . . . . . . . . . . . . . . . . 4

2.2 Absolute dosage level of the drugs  . . . . . . . . . . . . . . . . 5

4.1 Linear model with absolute dosage  . . . . . . . . . . . . . . . . 15

4.2 Absolute dosage with BIC selection  . . . . . . . . . . . . . . . . 16

4.3 Linear model with dosage level  . . . . . . . . . . . . . . . . . . 17

4.4 Dosage level with BIC selection  . . . . . . . . . . . . . . . . . . 18

4.5 Log transformation with absolute dosage  . . . . . . . . . . . . . 19

4.6 Log transformation with absolute dosage and BIC selection  . . . 20

4.7 Square root transformation with absolute dosage  . . . . . . . . . 21

4.8 Square root transformation with absolute dosage and BIC selection 22

4.9 Binomial model with absolute dosage  . . . . . . . . . . . . . . . . 23

4.10 Boxplots of the $\sqrt{MSE}$ with absolute dosage  . . . . . . . . . 25

4.11 Boxplots of the $\sqrt{MSE}$ on log transformation  . . . . . . . . . 26

4.12 Boxplots of the $\sqrt{MSE}$ on square root transformation  . . . . . 27

4.13 Neural Network Summary Table - J82  . . . . . . . . . . . . . . . . 28
## List of Tables

2.1 Summary of Drugs ................................................. 6
2.2 Summary of Cancer Cells ........................................... 6
4.1 $R^2$ for linear models ............................................. 12
4.2 $R^2$ for linear models with log and square root transformations . 13
4.3 $R^2$ for the binomial model ....................................... 14
ACKNOWLEDGMENTS

I wish to express my deepest gratitude to my thesis advisor Professor Hongquan Xu for his guidance, support and advice throughout my graduate study at UCLA. His intelligence and patience have improved my research skills.

I would also like to thank Professor Yingnian Wu and Professor Nicolas Christou for their encouragement and support.
CHAPTER 1

Introduction

Cancer begins when cells in part of the body start to grow out of control. Instead of dying, cancer cells grow and form new, abnormal cells and start invading other tissues.[1] Bladder cancer is a type of cancer that forms in tissues of the bladder, which is the organ that stores urine (Figure 1.1). What are the key statistics about bladder cancer? According to a recent estimate by the American Cancer Society for bladder cancer for 2013, there had been 15210 deaths reported from bladder cancer in 2013, leaving more than 500,000 people as survivors with an addition of 72,570 new cases being diagnosed in U.S. alone.[2] Thus, it is very crucial for both the patients and their families if a better treatment of bladder cancer can be developed by finding an optimal drug combination.

To get a better understanding of what bladder cancer is, we first want to investigate its causes. There are four main factors that predispose to bladder cancer: Occupational factors, such as in many cases, the patients being professionals working in chemical industry; Food factors, nitrites and nitrates in preserved foods; Chronic irritation of the bladder, such as lithiasis, chronic cystitis, schistosomiasis; and smoking along with drinking coffee and tea and analgesic abuse. The stages of bladder cancer are classified into TNM (tumor, lymph nodes, metastasis) system (Figure 1.1), each with several levels, and the extension of bladder cancer may occur locally or to invade nearby organs, or in other cases travel to further organs like brain, bone and liver. [3]

The question this analysis aims to answer is the following: Is there an ideal
drug combination that will effectively reduce the bladder cancer cell percentage, or even eliminate it? By answering this question, we will have a better understanding of whether we can find a model from the given dataset that results in optimization of drug combination.

Figure 1.1: Location of Bladder and the staging of Bladder Cancer
CHAPTER 2

Data

2.1 Overview

The data in this experiment consist of 125 combinations (runs) within 8 drugs, (namely D1 through D8), each with 6-10 absolute dosage levels in result of treating 6 different cancer cells. (Figure 2.1) The absolute dosage level has values taken with an increment of 5 times the previous dosage level, i.e. for D1, the absolute dosage level was from 0.00032 -> 0.0016 -> 0.008 -> 0.004 -> 0.2 -> 1. (Figure 2.2)

The dosage level is a logarithm transformation of the original absolute dosage level data. The main difference is that for the absolute dosage level, all drugs were recorded in the actual dosage level, (from 0.00004 to 12.5 depending on the drugs) whereas for the dosage level, we simply rescale the absolute dosage level into 0-6 for most of the drugs with few exceptions involved in decimals. i.e for D1, 0.00032 and 0.0016 absolute dosage level would be converted to dosage level 1 and 2.

The six types of cancer cells are: EJ, BIU87, J82, HBC, T24, c5637. The responses are recorded in a scale from 0 to 100 in terms of percentages. Moreover, this experiment was conducted under 6 batches, each with 15-25 combinations, it was designed in this way because each batch uses similar dosage levels, so that we can distinguish the drug effects in an intuitive manner and treat each batch as a block.
Figure 2.1: A sample view of the actual data
Figure 2.2: Absolute dosage level of the drugs, the column "Dosage Level" indicates what is the recorded value in terms of the dosage level after being transformed from the absolute dosage level. Column D1 - D8 simply shows each of the drugs has an increment of 5 times than the previous drugs along with some exceptions.

### 2.2 Data Selection

However, in order to reduce the complexity of our model, we chose to ignore the first batch, (the first 25 combinations), because it was the only batch that uses all 8 drugs, and the remaining 5 batches use D1, D2, D4, D6, D7 and D8. Another issue with the data was that for batch 6, only the first 3 cancer cells were recorded. This is caused by the researcher conducting this experiment, who thought it was not necessary to finish the experiment on the last 3 cancer cells for the very last batch based on the previous observation of the cancer cell’s response level. An interesting finding in this analysis was that there is a typo in combination 105 under D2, because 0.8 is way too big comparing with other dosage levels, thus we have chosen to ignore that particular combination. (Figure 2.1: combination in blue) And also for combination 80, the cancer cell remaining in J82 is -0.04, but...
we are certain that it was supposed to be +0.04. Thus we changed the sign of it. See (Table 2.1 and Table 2.2) for a summary of the data set (after being selected).

### Table 2.1: Summary of Drugs

<table>
<thead>
<tr>
<th></th>
<th>D1</th>
<th>D2</th>
<th>D4</th>
<th>D5</th>
<th>D7</th>
<th>D8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>1st Q.</td>
<td>0.00032</td>
<td>0.00128</td>
<td>0.01</td>
<td>0.60</td>
<td>0.0002</td>
<td>0.008</td>
</tr>
<tr>
<td>Median</td>
<td>0.0016</td>
<td>0.0064</td>
<td>0.05</td>
<td>0.3</td>
<td>0.001</td>
<td>0.2</td>
</tr>
<tr>
<td>Mean</td>
<td>0.121</td>
<td>0.04258</td>
<td>0.07303</td>
<td>0.6193</td>
<td>0.002454</td>
<td>0.3306</td>
</tr>
<tr>
<td>3rd Q.</td>
<td>0.008</td>
<td>0.032</td>
<td>0.05</td>
<td>1.5</td>
<td>0.005</td>
<td>1</td>
</tr>
<tr>
<td>Max</td>
<td>0.04</td>
<td>0.16</td>
<td>0.25</td>
<td>1.5</td>
<td>0.025</td>
<td>1</td>
</tr>
</tbody>
</table>

### Table 2.2: Summary of Cancer Cells

<table>
<thead>
<tr>
<th></th>
<th>EJ</th>
<th>BIU87</th>
<th>J82</th>
<th>HBC</th>
<th>T24</th>
<th>c5637</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min</td>
<td>10.47</td>
<td>1.063</td>
<td>0.00</td>
<td>0.08</td>
<td>3.184</td>
<td>0.511</td>
</tr>
<tr>
<td>1st Q.</td>
<td>32.25</td>
<td>4.776</td>
<td>8.412</td>
<td>42.88</td>
<td>14.195</td>
<td>30.793</td>
</tr>
<tr>
<td>Median</td>
<td>40.41</td>
<td>13.3</td>
<td>19.35</td>
<td>65.85</td>
<td>29.739</td>
<td>25.506</td>
</tr>
<tr>
<td>Mean</td>
<td>43.61</td>
<td>20.815</td>
<td>26.902</td>
<td>54.74</td>
<td>32.304</td>
<td>48.067</td>
</tr>
<tr>
<td>3rd Q.</td>
<td>51.77</td>
<td>31.865</td>
<td>43.881</td>
<td>74.03</td>
<td>50.476</td>
<td>65.583</td>
</tr>
<tr>
<td>Max</td>
<td>95.23</td>
<td>72.26</td>
<td>86.557</td>
<td>94.06</td>
<td>86.21</td>
<td>86.033</td>
</tr>
</tbody>
</table>

Table 2.1 is the summary of drugs that were used in this analysis, the numbers are recorded in absolute dosage level. Table 2.2 is the summary of cancer cells, the numbers are recorded in the percentage of cancer cells remaining after treatment.
CHAPTER 3

Method

There are several approaches in selecting models, first by looking into the summary of the dataset to get a sense of how the data behaves, then by performing different regression techniques such as linear model, generalized linear model with backward selection on both of the datasets (absolute dosage level vs dosage level in whole number scale), which allows us to eliminate the least significant terms one by one until no further improvement is possible. After that, log, square root and binomial transformation are taken on the response variable (cancer cells) for both datasets, and perform the same linear model regression with BIC stepwise selection to see if there are any improvements with the transformation. In order to compare the goodness of fit of all the models, $R^2$ is computed along with the observed vs fitted plot. Another approach is using the neural network model, since it can be used to perform nonlinear statistical modeling and provide new alternative regressions, and it is commonly used for developing predictive models in the medical field. [4]

3.1 Linear Model

In statistics, linear regression is the most widely used of all statistical techniques, it is the study of modeling the relationship between a dependent variable $y$ and one or more independent variables $X$, usually under an assumption of normally distributed errors. Note that, with more than one independent variables, it is called multiple linear regression. The goal of modeling using linear model is that, it can be fitted as a predictive model to an observed data set of $y$ and $X$ values,
if an additional value of $X$ is given without its accompanying value of $y$, then the fitted model can be used to make a prediction of the value $y$.

The general form of a linear model is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $x_i = (1, x_{i1}, \ldots, x_{ip})^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, $x_i^T \beta$ is the inner product between vectors $x_i$ and $\beta$, and $\varepsilon_i$ is the error term.

First, we construct a first-order linear model without any interaction effect or block effect, then we add on quadratic terms, interactions and block effects, but due to the fitness of our model in terms of $R^2$, a second-order model with linear, quadratic, interaction effects and block effect is more suitable in our case. It is important to include both the main effects and the interaction effects, as suggested by [5], main effects and interaction effects are called the factorial effects, and it is very efficient for studying two or more factors.

We consider a second-order model to systematically select the best model by using the backward selection technique. This will allow us to eliminate the least significant terms one at a time until there is no further improvement for the model. The general type of the backward selection model in terms of Rcode is:

```
model <- step(lm(EJ ~ (D1 + D2 + D4 + D6 + D7 + D8)^2 + I(D1^2) + I(D2^2) + I(D4^2) + I(D6^2) + I(D7^2) + I(D8^2) + Batch, data = training), k=log(nrow(training)))
```

The above Rcode is an example of the linear model with BIC stepwise selection. EJ is the cancer cell "EJ", D1- D8 are the drugs, Batch is the batch(block) effect, training is the 90% randomly selected training dataset from the original data, k
is the multiple of the number of degrees of freedom used for the penalty.
The coefficient of determination ($R^2$) is being compared for the goodness of fit. It provides information about how well a model fits given that the observed outcomes are replicated by the model, in terms of the proportion of total variation that is explained. That is:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\hat{y}_i = \hat{\mu}_i = x_i^T \hat{\beta}$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, and $n$ is the number of observations.

The risk function MSE (Mean Squared Error) corresponds to the expected value of the squared error loss, which is an estimator of the difference between values from an estimator and the true values of the quantity being estimated. In regression analysis, it is often referred to as the mean squared prediction error, which is the mean value of the squared deviations of the predictions from the actual true values.

The MSE formula is given:

$$MSE = \frac{\sum_{i=1}^{n}(\text{actual} - \text{predict})^2}{n}$$

where $n$ is the number of sample withhold in the cross validation. If $g(y_i)$ is the response in the linear model, then $\text{predict} = g^{-1}(\hat{y}_i) = g^{-1}(x_i^T \hat{\beta})$. We use log and square root transformations on the response as two different approaches on finding the best model.

### 3.2 Generalized Linear Model - Binomial

The general form of GLM is given by:

$$E(Y) = \mu = g^{-1}(X\beta)$$
Y is the dependent variable (response variable), $E(Y)$ is the expected value of Y, $\eta = X\beta$ is the linear predictor, and $g$ is the link function.

Binomial regression is often fitted as a generalized linear model with the predicted values $\mu$ are the probabilities of success, it is a technique in which the response followed by a Bernoulli trial with disjoint outcomes, ”success” and ”failure”. As a result, it is usually described primarily with a ”latent variable” indicating for the utility of making choices, an error variable distributed for the randomness. The likelihood of the prediction is given as:

$$L(Y|\mu) = \prod_{i=1}^{n}(1_{y_i=1}(\mu_i) + 1_{y_i=0}(1 - \mu_i))$$

$1_{y_i=1}$ is the indicator function which takes the value 1 when event $y_i$ occurs, and 0 otherwise. $\mu_i$ is the parameterised function of the explanatory variables.

In order to transform the expected value of the choice variable into a value that can be predicted by the linear predictor, a link function is introduced:

$$\eta = g(\mu)$$

$\eta$ is an intermediate variable representing a linear combination of the regression parameters in terms of the explanatory variables, and $g$ is a logit link function.

### 3.3 Cross Validation - Prediction

After selecting models, cross validation technique was used, 90% training and 10% testing used for all models. The main purpose was to see if our model generates accurate predictions, $R^2$, $MSE$ and the plots between predicted value vs. the actual values were being studied on determining the fitness of the model as well as how reliable it is when making predictions.
3.4 Neural Network

Unlike the linear model or generalized linear model, neural network was originally developed as an attempt to emulate the human brain. The main idea behind neural networks was to use a computer based model of the human brain to perform complex tasks, [6] including non-linear regression and discriminant models, data reduction models, and nonlinear dynamical systems.[7] The intuition behind this model is that we are allowed to use the artificial neural networks to simulate biological nervous system by combining neurons into a highly interconnected system (model).

A basic perceptron model of a neuron is:

\[ x_0 = f_0(\sum_{i \text{ inputs}} w_i x_i) \]

where \( f_0 \) is the activation function, \( w_i \) is the weights, and the value of \( x_0 \) depends on the inputs of \( x_i \).

A more advanced feed-forward neural network model with one hidden layer is:

\[ y_0 = \phi_0(\sum_h w_{ha} \phi_h(\sum_i w_{ih} x_i)) \]

where \( \phi_h \) is almost always logistic.
CHAPTER 4

Results

4.1 Tables & Plots

Tables 4.1, 4.2, 4.3 display the $R^2$ for each of the models described above as well as whether or not the BIC stepwise selection reduced any drugs.

<table>
<thead>
<tr>
<th></th>
<th>Absolute Dosage</th>
<th>Dosage Level</th>
<th></th>
<th>Absolute Dosage</th>
<th>Dosage Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear Model</td>
<td>BIC Selection</td>
<td>Linear Model</td>
<td>BIC Selection</td>
<td></td>
</tr>
<tr>
<td>EJ</td>
<td>0.8322</td>
<td>0.8042</td>
<td>0.8851</td>
<td>0.855</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remaining: D1, D2, D6, D8</td>
<td></td>
<td>Remaining: D1, D2, D4, D6, D8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BIU87</td>
<td>0.8511</td>
<td>0.8277</td>
<td>0.8219</td>
<td>0.7749</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remaining: all drugs</td>
<td></td>
<td>Remaining: D1, D2, D6, D7, D8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>J82</td>
<td>0.9256</td>
<td>0.9189</td>
<td>0.9164</td>
<td>0.8992</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remaining: all drugs</td>
<td></td>
<td>Remaining: D1, D2, D6, D7, D8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBC</td>
<td>0.8038</td>
<td>0.7548</td>
<td>0.8235</td>
<td>0.7802</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remaining: D4, D6, D8</td>
<td></td>
<td>Remaining: D2, D4, D6, D8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T24</td>
<td>0.7938</td>
<td>0.6871</td>
<td>0.824</td>
<td>0.769</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remaining: D4, D6, D8</td>
<td></td>
<td>Remaining: D1, D2, D4, D7, D8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c5637</td>
<td>0.8774</td>
<td>0.8535</td>
<td>0.9208</td>
<td>0.9063</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Remaining: D2, D4, D6, D8</td>
<td></td>
<td>Remaining: D2, D4, D6, D7, D8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1: $R^2$ for linear models
Table 4.1: $R^2$ for the absolute dosage and dosage level for both the linear model and the BIC stepwise selection along with the drugs that were remaining after the backward selection. Note: for some cancer cells like BIU87, no drugs were eliminated with the selection method.

<table>
<thead>
<tr>
<th></th>
<th>Log on Abs. Dosage</th>
<th>Sqrt on Abs. Dosage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Linear Model</td>
<td>BIC Selection</td>
</tr>
<tr>
<td>EJ</td>
<td>0.8213</td>
<td>0.7825</td>
</tr>
<tr>
<td></td>
<td>Remaining: D1, D2, D6, D8</td>
<td>Remaining: D1, D2, D6, D8</td>
</tr>
<tr>
<td>BIU87</td>
<td>0.9239</td>
<td>0.9159</td>
</tr>
<tr>
<td></td>
<td>Remaining: all drugs</td>
<td>Remaining: all drugs</td>
</tr>
<tr>
<td>J82</td>
<td>0.8632</td>
<td>0.8318</td>
</tr>
<tr>
<td></td>
<td>Remaining: D1, D6, D7, D8</td>
<td>Remaining: all drugs</td>
</tr>
<tr>
<td>HBC</td>
<td>0.718</td>
<td>0.6347</td>
</tr>
<tr>
<td></td>
<td>Remaining: D4, D6, D7, D8</td>
<td>Remaining: D4, D6, D8</td>
</tr>
<tr>
<td>T24</td>
<td>0.7378</td>
<td>0.6768</td>
</tr>
<tr>
<td></td>
<td>Remaining: D4, D6, D8</td>
<td>Remaining: D4, D6, D8</td>
</tr>
<tr>
<td>c5637</td>
<td>0.7828</td>
<td>0.7396</td>
</tr>
<tr>
<td></td>
<td>Remaining: D2, D4, D6, D8</td>
<td>Remaining: D2, D4, D6, D8</td>
</tr>
</tbody>
</table>

Table 4.2: $R^2$ for linear models with log and square root transformations

Table 4.2: $R^2$ for the log on absolute dosage and square root on absolute dosage of the response, for both the linear model and the BIC stepwise selection along with the drugs that were remaining after the backward selection. Note: for some cancer cells like BIU87, no drugs were eliminated with the backward selection method.
Table 4.3: $R^2$ for the binomial model

Table 4.3: $R^2$ for the binomial model with the response variable in the form of $(100X, 10000-X)$, where $X$ is the response of the cancer cell in the absolute dosage level. By looking at the $R^2$, it is difficult for us to determine which model is a better fit, since one model might be a good fit for a particular cancer cell but not others. Thus a follow up fitted vs. observed plots is needed in our case.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>EJ</td>
<td>0.8091</td>
</tr>
<tr>
<td>BIU87</td>
<td>0.896</td>
</tr>
<tr>
<td>J82</td>
<td>0.9208</td>
</tr>
<tr>
<td>HBC</td>
<td>0.793</td>
</tr>
<tr>
<td>T24</td>
<td>0.7994</td>
</tr>
<tr>
<td>c5637</td>
<td>0.8582</td>
</tr>
</tbody>
</table>
Figure 4.1: Linear model with absolute dosage
Figure 4.2: Absolute dosage with BIC selection
Figure 4.3: Linear model with dosage level
Figure 4.4: Dosage level with BIC selection
Figure 4.5: Log transformation with absolute dosage
Figure 4.6: Log transformation with absolute dosage and BIC selection
Figure 4.7: Square root transformation with absolute dosage
Figure 4.8: Square root transformation with absolute dosage and BIC selection
Figure 4.9: Binomial model with absolute dosage

Besides looking at the $R^2$ on the upper left corner, the first thing to do when comparing the observed vs. fitted plots is to see whether all points are closely attached to the abline, since ideally if all points are on the line, then there would have been a perfect fit. The second thing to look at is whether all points are equally distributed on the line. The third thing is to look if there are any outliers. Given the three conditions, we now take a close look on the comparison of two plots, Figure 4.2 and Figure 4.6. Figure 4.2 is the linear model with BIC stepwise
selection on the absolute dosage level, it clearly has a better fit than Figure 4.6 which is the log transformation on the response of the linear model with BIC stepwise selection on the absolute dosage level. Since all of the $R^2$ in Figure 4.2 is higher than what it is in Figure 4.6, the points are almost equally distributed and less scattered along with fewer outliers. Thus we conclude that Figure 4.2 has a better fit than Figure 4.6.

By observing the fitted vs. observed Plots, a linear model with absolute dosage and its stepwise selection seem to be the best model in term of the fitness, thus we will now focus on the predictions. The method of cross validation is that we randomly assign 90% of the data as ”training set” and the remaining 10% of the data as ”testing set”. We run both the linear model and the BIC stepwise selection with absolute dosage on the 90% training set, and then use this 90% to predict the remaining 10% training set, in order to see how accurate our predictions are in terms of the cancer cell remaining given the drug levels.

The way that the prediction works is that, we conduct 100 simulations, each randomly assigning 90% of the training set and then predicting on the remaining 10% of the testing set. To determine the goodness of the fit, we compute its $R^2$ and $MSE$, and then put them into boxplots to see if our predictions are accurate. Unfortunately, the $R^2$ for the prediction model turns out to be unacceptable, thus we will mainly focus on the $MSE$. Figure 4.10 - 4.12 display the boxplots of $\sqrt{MSE}$ for the cancer cell EJ.
Figure 4.10: Boxplots of the $\sqrt{MSE}$ with absolute dosage
Figure 4.11: Boxplots of the $\sqrt{MSE}$ on log transformation
To compare the MSE boxplot on the 100 simulations, the lower the value of a square root MSE, the better the predictions are. By observing Figure 4.10, both the linear model and the BIC stepwise selection for the MSE are within a reasonable range, since the median is around 10, whereas the median for Figure 4.11 is generally above 50. Although all figures have outliers. The difference between the top three plots and the bottom three are that the top three are in the range of $0 - 500$ on the $y$–axis, which is the the value of $\sqrt{MSE}$, whereas the bottom three are in the range of $0 - 50$, thus they give a better view in terms of the median, and the IQR.
4.2 Neural Network

According to the article "Systematic Quantitative Characterization of Cellular Responses Induced by Multiple Signals". [8] Neural network should be an efficient model for biological related regression analysis. The authors argue that neural network models performed generally better for fitting the data with fewer available data points to fit the model, we were in a similar case of not having many data points. A summary of the neural network model result in terms of the correlation square between the actual and the predicted value is provided in Figure 4.13.

Figure 4.13: Neural Network Summary Table - J82, this table returns the correlation square between the actual and the predicted value. Linout is the switch for linear outputs, decay is the parameter for weight decay, maxit is the maximum number of iterations, size is the number of units in the hidden layer, sd = 0 means that the fitted values are constant and the correlation cannot be calculated. By observing the table, linout=F is not a good option, thus normally a model will be fitted as linout=T, decay=0 and maxit=100 along with size 1 or 2.

Since it was not easy to compute the $R^2$ for neural network models, we chose to calculate the correlation square between the actual and the predicted values, and then compare it with the correlation square between the actual and the predicted...
value for linear models. It seems that neural network model did not provide any improvements in terms of the fitness, but complicated the model in its way. An example would be the correlation square of the actual vs. the fitted value for J82 in the linear model with absolute dosage is 0.87732 vs. 0.82728 in the NNET model given size 2, linout=T, decay=5e-4, and maxit=100.
CHAPTER 5

Conclusion

5.1 Review of Procedure

We first ran second-order linear models on both the absolute dosage and the dosage level, we then took log, square root and binomial transformations on the response values (cancer cell) of the absolute dosage level. By comparing the $R^2$ and the predicted vs. fitted plots, we narrowed down to the linear model and the BIC stepwise selection for the absolute dosage level. After that, we took cross-validation with 90% as the training set and 10% as the testing set, to test the fitness of our model as well as its prediction. Another approach taken was the neutral network, although it did not improve the model.

5.2 Results

By comparing all the models and the predictions, we found that linear model and the BIC stepwise selection on the absolute dosage level is the best fitted model in terms of both the prediction and the fitness.

We tried to find a stabilized model in both the prediction and the fitness, however the best model we found was not stabilized, since it gives unusual large values when cross validation was used.

Further study might be implied with more than 2 degree order linear model, but then the model would be much more complex.
REFERENCES


