OPTIMAL DYNAMIC TREATMENT STRATEGIES

DETADEEN ALI ALSHIBANI

Thesis submitted for the degree of

Doctor of Philosophy

School of Mathematics and Statistics

Newcastle University

United Kingdom

NOVEMBER 2010

Acknowledgements

I am extremely grateful to my supervisor Professor Robin Henderson, for his guidance, encouragement, advice and support throughout the development and completion of this research. I would also like to thank my advisor Dr Phil Ansell for his assistance regarding dynamic programming and bandit problem chapters. Thanks also go to both the past and present postgraduate students and the staff in the Mathematics and Statistics Department at Newcastle University. Special praise must be reserved for my father, mother, wife, daughter Leen and sons Ali, Nael, Omar and Mohammed for their personal support. Also I thank all my friends for their constant inspiration.

Sections within Chapter 4 have been published in paper in Biometrics in 2009 titled *Regret-Regression for Optimal Dynamic Treatment Regimes* by Robin Henderson, Phil Ansell and Deyadeen Alshibani.

Abstract

Dynamic treatment regimes are functions of treatment and covariate history which are used to advise on decisions to be taken. Murphy (2003) and Robins (2004) have proposed models and developed semi-parametric methods for making inferences about the optimal dynamic treatment regime in a multi-interval study that provide clear advantages over traditional parametric approaches.

The main part of the thesis investigates the estimation of optimal dynamic treatment regimes based on two semi-parametric approaches: G-estimation by James Robins and Iterative Minimization by Susan Murphy. Moodie et al. (2006) show that Murphy's model is a special case of Robins' and that the methods are closely related but not equivalent.

In this thesis we first describe and demonstrate the current theory, then present an alternative method. This method proposes a modelling and estimation strategy which incorporates the regret functions of Murphy (2003) into a regression model for observed responses. Estimation is fast and diagnostics are available, meaning a variety of candidate models can be compared. The method is illustrated using two simulation scenarios taken from the literature and using a two-armed bandit problem. An application on determination of optimal anticoagulation treatment regimes is presented in detail.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview of Dynamic Treatment Regimes

The subject of dynamic treatment regime estimation has attracted the attention of many researchers. It is a method for optimizing a regime which is changing over time. It plays an important role in making decisions in different fields. For example a physician needs to adjust dynamically a treatment strategy for his or her patients at each time point and has to prescribe an appropriate treatment strategy based on the diagnosis carried out for the patients and the treatment information. The decision taken might be to continue to treat a patient with the current treatment or to increase/ decrease it. A dynamic treatment regime is a sequence of decision rules. It introduces an ethical and flexible set of formal rules studying the effects of treatment which are adjusted over time according to the response to treatment.

Murphy (2003) defined a dynamic treatment regime as a list of decision rules, one per time interval, for how the level of treatment will be tailored through time to an individual's changing status. It is also defined as a function that takes treatment, covariate history and baseline covariates as inputs and returns a decision to be taken (Moodie et al. 2005).

The treatment regime strategies play a critical role in the method or course of remedial treatment of several diseases such as AIDS or cancers. For such diseases, physicians face the difficult problem of deciding when and which drugs to administer to a patient. Furthermore, for such diseases, the strength and interaction of the treatment with the immune system are so complex that the design of the optimal treatment strategy may require some complicated analysis.

1.2 Notation

At the start, we need some notation. Our development extends to the general case.

- K denotes the number of intervals;
- j is a specific interval so that $j = 1, 2, 3, ..., K$.
- M_j represents a status variable available at the start of the j^{th} interval, in particular, M_1 represents baseline covariates and M_2 includes time-varying covariates which may depend on treatment received in the earlier intervals. M_j may be scalar or multivariate.
- T_j, is the treatment at interval j given subsequent to observing M_j.
- Y is the outcome observed at the end of the K^{th} interval, and large values of Y are preferred. So, the occurrence is $(M_1, T_1, M_2, T_2, \cdots, M_K, T_K, Y)$.
- $\bar{M_j}$ denotes a status variable at time j and its history, e.g. $\bar{M_j} = (M_1, M_2, ..., M_j)$. Also $\bar{T_j} = (T_1, T_2, ..., T_j)$.
- Specific values are denoted with the lower case, e.g. $m_1, t_1, m_2, \cdots, t_K, m_K$. Also $\bar{t_j} = (t_1, t_2, ..., t_j)$.

- d_j a rule or regime and $\underline{d}_{j+1}^{opt}$ is interpreted to mean that optimal rules are followed from time $j+1$ onward.

1.3 Blood Clot

Blood clotting is the body's natural protection against extreme bleeding. Anticoagulation is used for patients with either a history or a risk of thrombosis (abnormal formation of blood clots) in many disorders. The general aim is to prevent thrombotic complications while keeping the hemorrhage risk low (Landefeld et al. 1993).

1.3.1 Anticoagulation dosage

Warfarin is an anticoagulant. It is classified to be efficient and reasonably safe for preventing thrombosis. It is the most commonly prescribed anticoagulant drug. It interacts with many widely-used drugs, and the response of warfarin varies seriously between patients. Some foods have been reported to interact with warfarin. Its activity has to be monitored by regular blood testing for the international normalized ratio (INR) to make sure an adequate yet safe dose is taken (Holbrook et al. 2005) . When initiating warfarin therapy, the specialist will decide how strong the anticoagulant treatment needs to be. The target INR level will vary from person to person depending on the clinical indicators, but tends to be 2-3 in most conditions. Target INR may be 2.5-3.5 (or even 3.0-4.5) in patients with one or more mechanical heart valves (Baglin 2006). The common side effect of warfarin is bleeding. The risk of severe bleeding is small but clearly visible (the average annual rate is 0.9 to 2.7 %), and clearly benefit should outweigh the risk when warfarin is considered as a therapeutic measure. Risk of bleeding is augmented if the INR is out of range due to accidental or deliberate overdose or due to interactions (Horton 1999).

1.3.2 Warfarin data

Rosthøj et al (2006) used data from 350 patients given Warfarin anticoagulation from one hospital between February 1995 and August 2000. Treatment periods varied from 16 days to almost five years and involved from two to 124 clinic visits. Covariates available include age, sex, and diagnosis. For this first analysis we decided to concentrate on a subset of the data, made up of the first 14 clinic visits for the 303 Warfarin-treated patients with at least that number of visits. The first four visits are considered as induction and the analysis concentrates on the remaining 10. In the analysis given in this thesis we do not consider the effect of covariates or time intervals between visits. As will be seen, the consequent marginal approach brings a range of practical difficulties to be overcome before we can consider conditional analyses. We have to decide what to define as state M_j at visit j . After consultation we selected $M_j = 0$ if INR in range, or $M_j = D_j/R$, where D_j is the difference (positive or negative) between INR at time j and the nearest boundary of the target range, and R is the width of that target range. Half of the visits had INR in range, $M_j = 0$. If not, the distribution was positively skew with range -1.53 to 5.00, lower and upper quartiles -0.19 and 0.80, and median 0.25.

For the actions, we decided to define T_j to be the change in prescribed dose at visit j, since usually a decision consists of two stages: first, whether or not to change dose; second, if changing, to what value. Actually there is also a third stage, a recommendation as to the time interval until the next clinic visit, but this will not be considered. The dose level (and change in dose) is a discrete variable, determined by the 0.5, 1, 3 and 5 mg Warfarin tablet sizes, but a fairly large number of combinations are used in practice. Some 61 per cent of visits result in a decision to leave dose unchanged. Otherwise, the dose change distribution is fairly symmetric about zero, with standard deviation close to 1 mg but there were occasional very large changes, the range being -9 to +8 units.

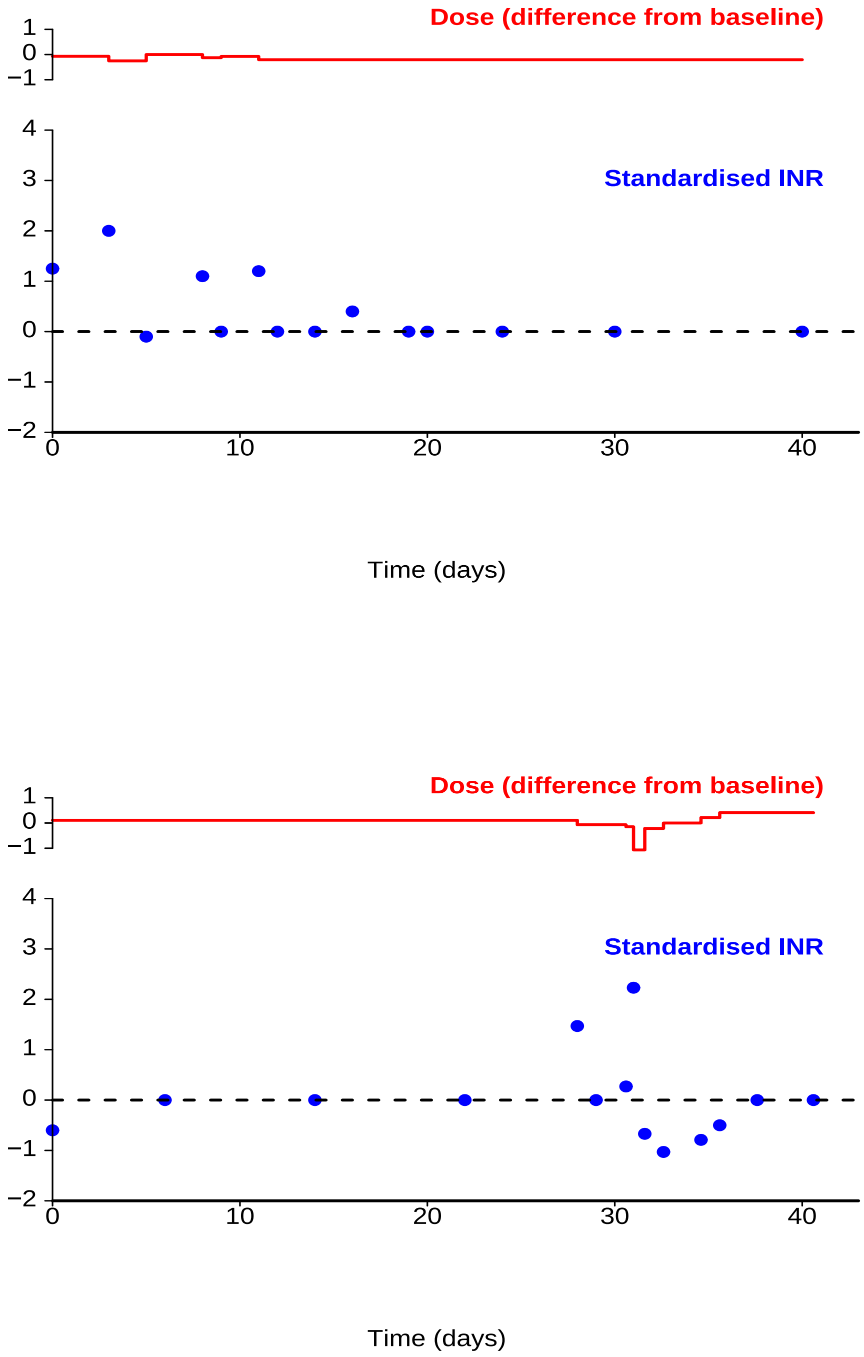

Figure 1.1: Two illustrations of anticoagulant data.

Figure 1.1 illustrates the type of data for two patients. The top part of each plot shows how dose was (or was not) changed at clinic visits. The lower part of each plot shows the standardised International Normal Ratio (INR), a measure of blood clotting speed. The standardised version we use has a mixed distribution with a point mass at zero if clotting speed is within the target range, and otherwise a positive value if clotting time is too long, negative if clotting is too quick. The upper plot shows data from a patient with the quite common pattern of initially unstable INR being brought under control by a small number of modest dose changes. The lower plot shows the less common, but not unusual, situation of a patient whose clotting time suddenly increases after a stable period. Several dose changes are needed to regain control and at one time there is overadjustment, causing INR to fall below target and subsequently have to be increased again. In both cases trial-and-error combined with the clinicians' experience and judgment were used to determine dose changes: there are as yet no accepted decision rules. Since the achieved quality of anticoagulation control is often poor, and with the use of anticoagulants increasing worldwide, there is a need for more objective and routine procedures.

1.4 Thesis outline

In order to improve optimal dynamic treatment regimes methodology we propose a modelling and estimation strategy which incorporates the regret functions of Murphy (2003) into a regression model for observed responses. Estimation is quick and diagnostics are available, meaning a variety of candidate models can be compared.

We initially describe the basics of dynamic programming and causal inference, to increase our understanding of the response to treatment and the effects of various covariates. Then we first describe and demonstrate the current theory. We investigate the estimation of optimal dynamic treatment regimes based on a regret function approach introduced by Susan Murphy and a G-estimation method by James Robins. These are compared first for simulated sequential randomization trials taken from the literature. We then present regret-regression for optimal dynamic treatment regimes as alternative method.

The second chapter presents background on dynamic programming and causal inference. We discuss some basic assumptions which are needed to estimate dynamic treatment regimes. Chapter 3 introduces the reader to Murphy and Robins methods and investigates their techniques to discover any characteristics or any important relationships between them.

Chapter 4 poses the main idea of this thesis. We introduce regret-regression for optimal dynamic treatment regimes with discussion and comparison with Murphy and Robins methods and illustrate the method via simulations and an application data.

Our extension is presented in the next two chapters. In Chapter 5 we use another alternative method called inverse probability of treatment weighting to estimate the optimal dynamic treatment regimes then compare its results with the regret-regression and prove that theoretically. This is then extended in chapter 6 to illustrate the regret-regression estimation method using a common decision stochastic problem from the literature called multi-armed bandit problem.

Chapter 7 uses the regret-regression method and knowledge gathered from the last few chapters and presents some diagnostics for choosing the correct model to estimate optimal dynamic treatment regimes via Murphy (2003) scenario and with an application to the anticoagulant example. Conclusion is then presented in the closing chapter.

Chapter 2

Background

2.1 Dynamic Programming

2.1.1 What is dynamic programming?

Dynamic programming is the problem of optimizing a sequence of decisions in which each decision must be made after the result of the previous decision becomes known (Upton and Cook 2002). It is a method for optimizing a system changing over time that has been successfully applied in manufacturing systems, environmental engineering, business, and many other fields. Due to the infamous curse of dimensionality, exact solutions are only possible for small problems or under very limiting restrictions. However, recent advances in computing power have given rise to many approximate dynamic programming methods. These advances now provide the potential for the application of DP to complicated dynamic decisions, such as adaptive interventions or dynamic treatment regimes. The key advantage of dynamic programming is its ability to account for future decisions when optimizing a current decision. Dynamic programming is an optimization procedure that is designed to efficiently search for the global optimum of a function; it is an algorithmic technique based on a recurrent formula and some starting states (Bellman and Dreyfus 1962).

The problem can usually be divided into stages with a decision required at each stage. Each stage has a number of states associated with it. There exists a recursive relationship that identifies the optimal decision for stage i, given that stage $i + 1$ has already been solved, and the final stage must be solvable by itself (Yong et al. 2007). The dynamic programming solutions have a complicated polynomial function which requires a much faster running time than other techniques. The most important and difficult issue in dynamic programming is how to determine stages and states so that all of the previous characteristics hold. More details and examples can be found in Cormen et al. (2001).

The uniqueness of dynamic programming resides in the principle of optimality and it is on this principle that the whole of dynamic programming is based. Just as in the calculus we use the basic idea of solving a function after differentiating and equating to zero to find its minima or maxima (remembering to evaluate the function at the end points) so in dynamic programming we use the principle of optimality expressed in the functional equation. The principle is easy enough to recite. Bellman put it this way: "An optimal policy has the ownership that, whatever the initial state and initial decision are, the remaining decisions must form an optimal policy with regard to the state resulting from the first decision". As Bellman says in his first book on dynamic programming, "this observation has all the dangerous simplicity of half-truth and it can be proved quickly enough by a proof by contradiction. It is not however, an easy principle to understand and it is worthwhile spending some time on a simple illustration of the ideas (Norman 1975)". So, we can define dynamic programming as an approach to optimization. Optimization means finding a best solution among several feasible alternatives in each stage of a multistage decision. The term "a best solution" is used because there may be more than one optimal solution (Nemhauser 1967).

Most medical decision problems are extremely complex and contain a large number of variables. Abstraction simplifies the process of building a decision model by allowing a model builder to work at a level of detail that the builder is most comfortable with. It is also useful in time critical situations or when there is not enough data to support complete specification of probabilities of the uncertain events. Creation of formal models for decision-making involves selecting the set of relevant factors to consider and the level of detail at which to represent them. Often the best choice is not obvious at the outset. A model builder may begin constructing a model and realize that certain portions need to be refined. Or he may become overwhelmed by the complexity of the developing model and decide to abstract away some detail, at least temporarily, to simplify his task (Sundaresh et al 1999).

2.1.2 Elements of the dynamic programming model

The basic components of the model are:

- T. This is a variable that can be manipulated to achieve the desired objective. Write $T_j = (t_1, t_2, \ldots.., t_K)$. These variables are commonly referred to as independent or decision variables.

- The factor $M = (m_1, m_2, \ldots.., m_K)$, affects the objective but these are not controllable.

- The measure of effectiveness Y, can be the utility, or return associated with particular values of the decision variables and parameters. The measure of effectiveness, alternatively called the utility measure, criterion function, objective function, or return function, is a real-valued function of the decision variables and parameters.

  There is a wide variety of commonly used measures of utility, such as cost, profit and rate of return. It will be assumed that a specific measure of effectiveness can always be chosen that will adequately reflect the important differences among different values of the decision variables (Eric 1982).

- Any T satisfying the constraints (which is conditional on the model) is known as feasible solution to the model. The decision-making problem is to find a feasible solution that yields high value or return. An optimal solution (T^{opt}) is defined as a feasible solution producing the greatest possible return, that is,

$$Y(T^{opt}, M) = \max_{T} Y(T, M)$$

- $C_j(T_j)$ and $R_j(T_j)$ are respectively cost and reward the decision variable.

2.2 Deterministic Dynamic Programming

Deterministic dynamic programming is characterised by the fact that, once the decision rule has been selected, the outcomes are known in advance. For naturally sequential processes it means that the change of state at any stage is completely determined by the action at the preceding stage and by the state at that stage. As a consequence, for naturally sequential deterministic processes, one can either treat the decisions one by one or all together (Taha 1992).

To illustrate the steps of the dynamic programming model, let us consider the following example (basics were taken from Taha (1987), but we modified it for fixing ideas). Suppose a physician wishes to treat a patient. He or she tries to avoid side effects (e.g., medicine toxicity) by using only 6 dosage units for allocation to all four time points. Each time point is requested to submit its treatment giving cost $C_j(T_j)$ of alternative treatments $T_j = \{1, 2, 3\}$ at $j = 1, 2, 4$ and $T_3 = \{1, 2\}$ (e.g., total dosage units used at time j) and rewards $R_j(T_j)$ at time j (e.g., total treatment utility units) for each treatment T_j. Table 2.1 summarises the costs and the rewards. The zero proposals are introduced the possibility of not allocating dosage units to individual time points. The aim is to maximize the total treatment rewards resulting from the allocation of the 6 units to the four time points.

	Time 1		Time 2		Time 3		Time 4	
T_j	$C_1(T_1)$	$R_1(T_1)$	$C_2(T_2)$	$R_2(T_2)$	$C_3(T_3)$	$R_3(T_3)$	$C_4(T_4)$	$R_4(T_4)$
1	0	0	0	0	0	0	0	0
2	3	10	1	3	2	4.2	2	5.6
3	4	14	4	12	–	–	3	7.2

Table 2.1: Costs and rewards

The problem has $3 \times 3 \times 2 \times 3 = 54$ possible treatment policies, some of them are infeasible because they require more dosage units than the 6 dosage units available. If the total cost for any of the 54 combinations does not exceed the 6 dosage units, its total reward is computed. The optimal treatment policy is the feasible combination yielding the highest total rewards. For example, a feasible treatment policy of $\{2, 1, 1, 2\}$ is to treat the patient at time points 1 and 4 using 3 and 2 dosage units respectively. They will cost 5 dosage units and yield a total reward of 15.6 treatment utility units.

2.2.1 Forward recursive equation:

Let us introduce that $f_j(M_j)$ is the maximum rewards at time j given the state M_j and

$$\max Y = \sum_j f_j(M_j),$$

is the maximum total reward at the end of the final time point. To follow the forward procedure, the computations are carried out in order f_1, f_2, f_3, f_4. Computations advance from the first to the last time point. Now, let M_j = units of dosage allocated to time points 1 up to j. We thus write the recursive equation for the dosage limited example as,

$$\begin{aligned} f_1(M_1) &= \max_{C_1(T_1)\leq M_1} [R_1(T_1)] \\ f_j(M_j) &= \max_{C_j(T_j)\leq M_j} [R_j(T_j) + f_{j-1}(M_{j-1})] \qquad j = 2, 3, 4. \end{aligned}$$

Since we have a deterministic model we can write $M_1 = C_1(T_1)$ say. Similarly $M_2 = C_1(T_1)+C_2(T_2)$, $M_3 = C_1(T_1)+C_2(T_2)+C_3(T_3)$ and $M_4 = C_1(T_1)+C_2(T_2)+C_3(T_3)+C_4(T_4)$. Thus $M_1 = M_2 - C_2(T_2)$, $M_2 = M_3 - C_3(T_3)$ and $M_3 = M_4 - C_4(T_4)$. We can define

$$M_{j-1} = M_j - C_j(T_j),$$

then,

$$\begin{aligned} f_j(M_j) &= \max_{C_j(T_j)\leq M_j} [R_j(T_j) + f_{j-1}(M_j - C_j(T_j))] \qquad j = 2, 3, 4. \\ f_1(M_1) &= \max_{C_1(T_1)\leq M_1} [R_1(T_1)] \\ f_2(M_2) &= \max_{C_2(T_2)\leq M_2} [R_2(T_2) + f_1(M_2 - C_2(T_2))]. \end{aligned}$$

The computations are carried out as shown in Table 2.2.

	$R_1(T_1)$			Optimal solution			$R_2(T_2) + f_1(M_2 - C_2(T_2))$			Optimal solution	
M_1	$T_1 = 1$	$T_1 = 2$	$T_1 = 3$	$f_1(M_1)$	T_1^{opt}	M_2	$T_2 = 1$	$T_2 = 2$	$T_2 = 3$	$f_2(M_2)$	T_2^{opt}
0	0	-	-	0	1	0	0+0=0	-	-	0	1
1	0	-	-	0	1	1	0+0=0	3+0=3	-	3	2
2	0	-	-	0	1	2	0+0=0	3+0=3	-	3	2
3	0	10	-	10	2	3	0+10=10	3+0=3	-	10	1
4	0	10	14	14	3	4	0+14=14	3+10=13	12+0=12	14	1
5	0	10	14	14	3	5	0+14=14	3+14=17	12+0=12	17	2
6	0	10	14	14	3	6	0+14=14	3+14=17	12+0=12	17	2

Table 2.2: Costs and rewards for time points 1 and 2

As shown the values of M_1 and M_2 are discrete, they as well as M_3 in Table 2.3, may only assume the values $\{0, 1, \cdots, 6\}$ but are not known exactly. On the other hand, M_4, which is the total treatment units allocated to all time points, is equal to 6. The solution of the problem is to start with the time 1 (the left side of Table 2.2). Given the value of M_1, we obtain conditional decisions and choose the best alternative whose cost does not exceed M_1. The right side of Table 2.2 shows time point 2 calculations. The idea now is to choose the alternative in time 2 given M_2 that yields the best reward for time points 1 and 2. We now describe the details for time 2 computations.

- When $M_2 = 0$ the only feasible alternative given M_2 is $T_2 = 1$ whose cost and reward units are both equal to zero.
- $M_2 = 1$. Here we have two feasible alternatives given M_2. They are $T_2 = 1$ and $T_2 = 2$ costing 0 and 1 and yielding rewards of 0 and 3 respectively. Thus, the values of $M_1 = M_2 - C_2(T_2)$ corresponding to $T_2 = 1$ and $T_2 = 2$ are $1 - 0$ and $1 - 1$. The corresponding best rewards from time 1 given $M_1 = 1$ and 0 are both equal to zero. Thus $f_2(M_2 = 1) = \max(0 + 0, 3 + 0) = 3$, corresponding to $T_2 = 2$, which is the optimal treatment at this time point when $M_2 = 1$.
- When $M_2 = 4$. Feasible alternatives are $T_2 = 1$, $T_2 = 2$ and $T_2 = 3$ costing 0, 1 and 4 and yielding rewards of 0, 3 and 12 respectively. Thus, the values of $M_1 = M_2 - C_2(T_2)$ corresponding to $T_2 = 1$, $T_2 = 2$ and $T_2 = 3$ are $4 - 0$, $4 - 1$ and $4 - 4$. The corresponding best rewards from time 1 given $M_1 = 4$, 3 and 0 are 14, 10 and zero respectively. Thus $f_2(M_2 = 4) = \max(0 + 14, 3 + 10, 12 + 0) = 14$ and $T_2^{opt}(M_2 = 4) = 3$.

At time 3,

$$f_3(M_3) = \max_{C3(T_3)\leq M_3} [R_3(T_3) + f_2(M_3 - C_3(T_3))] \qquad\qquad T_3 = 1, 2$$

	$R_3(T_3) + f_2(M_3 - C_3(T_3))$		Optimal Solution	
M_3	$T_3 = 1$	$T_3 = 2$	$f_3(M_3)$	T_3^{opt}
0	0+0=0	-	0	1
1	0+3=3	-	3	1
2	0+3=3	4.2+0=4.2	4.2	2
3	0+10=10	4.2+3=7.2	10	1
4	0+14=14	4.2+3=7.2	14	1
5	0+17=17	4.2+10=14.2	17	1
6	0+17=17	4.2+14=18.2	18.2	2

Table 2.3: Costs and rewards for time 3

Then for time 4

$$f_4(M_4) = \max_{C4(T_4) \leq M_4} [R_4(T_4) + f_3(M_4 - C_4(T_4))] \qquad T_4 = 1, 2, 3$$

	$R_4(T_4) + f_3(M_4 - C_4(T_4))$			Optimal solution	
M_4	$T_4 = 1$	$T_4 = 2$	$T_4 = 3$	$f_4(M_4)$	T_4^{opt}
6	0+17=17	5.6+14=19.6	7.2+10=17.2	19.6	2

Table 2.4: Costs and rewards for stage 4

Now we can read the optimal solution directly starting from time 4, we can choose $T_4 = 2$, which cost 2 dosage units. Then M_3 from time 3 will be 6-2=4. From Table 2.3, we see that optimal alternative given $T_3 = 1$. Since $T_3 = 1$ cost zero units, we have again, from the right side of Table 2.2, we obtain $T_2 = 1$ as the optimal alternative time 2. Finally since $T_2 = 1$ cost zero units and using the left side of Table 2.2, we obtain $T_1 = 3$ as the optimal alternative at the first time point. Thus the optimal combination of proposals for time points 1, 2, 3 and 4 is (3, 1, 1, 2), which yields response 19.6

2.2.2 Backward recursive equation

In the dynamic programming literature, the recursive equation is set up such that the computations start at the last time point and then proceed back to time 1. This method is called the backward procedure. The main difference between the forward and the backward methods occurs in the way we define the state of the system.

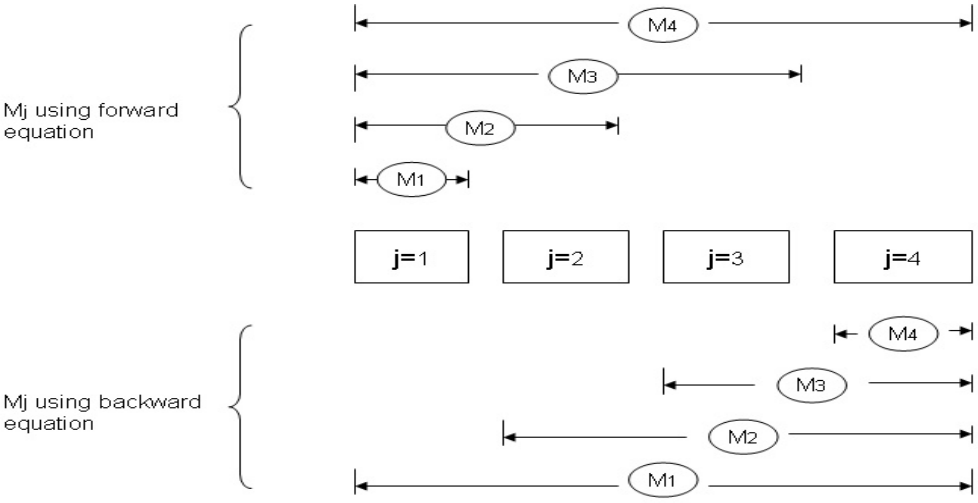

Figure 2.1: States M_j using the forward and backward methods.

We define the state M_j as the amount of dosage T_j allocated to time j onward and $f_j(M_j)$ as the corresponding optimal reward at time j, for $j = \{1, 2, 3, 4\}$. The order of time point computations is thus: f_4, f_3, f_2 and f_1 .

$$\begin{aligned} f_4(M_4) &= \max_{C_4(T_4)\leq M_4}[R_4(T_4)], \\ f_j(M_j) &= \max_{C_j(T_j)\leq M_j}[R_j(T_j)+f_{j+1}(M_j-C_j(T_j))] \ \ j=1,2,3. \\ f_3(M_3) &= \max_{C_3(T_3)\leq M_3}[R_3(T_3)+f_4(M_3-C_3(T_3))]. \end{aligned}$$

The computations are carried out as follows. First: at time points 4 and 3

		$R_4(T_4)$		O.S			$R_3(T_3)+f_4(M_3-C_3(T_3))$		Optimal solution	
M_4	$T_4=1$	$T_4=2$	$T_4=3$	$f_4(M_4)$	T_4^{opt}	M_3	$T_3=1$	$T_3=2$	$f_3(M_3)$	T_3^{opt}
0	0	-	-	0	1	0	0+0=0	-	0	1
1	0	-	-	0	1	1	0+0=0	-	0	1
2	0	5.6	-	5.6	2	2	0+5.6=5.6	4.2+0=4.2	5.6	1
3	0	5.6	7.2	7.2	3	3	0+7.2=7.2	4.2+0=4.2	7.2	1
4	0	5.6	7.2	7.2	3	4	0+7.2=7.2	4.2+5.6=9.8	9.8	2
5	0	5.6	7.2	7.2	3	5	0+7.2=7.2	4.2+7.2=11.4	11.4	2
6	0	5.6	7.2	7.2	3	6	0+7.2=7.2	4.2+7.2=11.4	11.4	2

Table 2.5: Costs and rewards for time points 4 and 3

Next: at time 2

where

$$f_2(M_2)=\max_{C_2(T_2)\leq M_2}[R_2(T_2)+f_3(M_2-C_2(T_2))].$$

And then: at time 1

$$f_1(M_1)=\max_{C_1(T_1)\leq M_1}[R_1(T_1)+f_2(M_1-C_1(T_1))].$$

In time 1, the optimal solution is determined by starting with M_4 at time 4 and proceeding to M_1 at time 1. Naturally, the solutions are identical with those of the forward method.

	$R_2(T_2) + f_3(M_2 - C_2(T_2))$			Optimal solution	
M_2	$T_2 = 1$	$T_2 = 2$	$T_2 = 3$	$f_2(M_2)$	T_2^{opt}
0	0+0=0	-	-	0	1
1	0+0=0	3+0=3	-	3	2
2	0+5.6=5.6	3+0=3	-	5.6	1
3	0+7.2=7.2	3+5.6=8.6	-	8.6	2
4	0+9.8=9.8	3+7.2=10.2	12+0=12	12	3
5	0+11.4=11.4	3+9.8=12.8	12+0=12	12.8	2
6	0+11.4=11.4	3+11.4=14.4	12+5.6	17.6	3

Table 2.6: Costs and rewards for time 2

	$R_1(T_1) + f_2(M_1 - C_1(T_1))$			Optimal solution	
M_1	$T_1 = 1$	$T_1 = 2$	$T_1 = 3$	$f_1(M_1)$	${T_1}^{opt}$
6	0+17.6=17.6	10+8.6=18.6	14+5.6=19.6	19.6	3

Table 2.7: Costs and rewards for time 1

2.3 Stochastic Dynamic Programming :

When the states and the returns at each time point are dependent on probabilities then we use the technical terms *stochastic decision making* to describe the decisions under risk or just uncertainty. For example, let us consider a single-time stochastic return function $R(T, M)$, where

T: is a decision variable.

M: is a discrete random variable.

For a fixed set of decisions T, we define $R(T, M)$ as expectation $[E_M[r(T, M)]$. Now, we consider two different related examples

Example 1

Suppose a physician has to decide (within 4 time points), when he or she should treat a patient with leukemia using a chemotherapy drug. Each time point, the physician is informed of a new state opportunity which he or she must either treat the patient or wait. If wait an opportunity is lost in the sense that the physician can not treat past events. The opportunity appearing at each time point is independent of the ones appearing in other time points. It is known that at the start of each time point, it may be one of three states $m_j = \{1, 2, 3\}$. m_j occurs with probability $p(m_j)$. It is given that: $p_1 = p(m_j = 1) = 0.3, p_2 = p(m_j = 2) = 0.5, p_3 = p(m_j = 3) = 0.2$. The drug provides a return, (e.g., a ratio of natural cells) equal to $r(m_j)$ where $j = 1, 2, 3, 4$ if the paient was treated. The return is given as $r1 = r(m_j = 1) = 0.10, r2 = r(m_j = 2) = 0.20, r3 = r(m_j = 3) = 0.30$. We want to establish the strategy that the physician should follow to maximise his or her expected return. Note that the probability distribution is the same at all time points.

To solve this problem we note that any time point j, the physician has to decide a binary decision T_j, either to treat $t_j = 1$ or wait $t_j = 0$. Define $f_j(m_j) =$ maximal expected return when starting in time point j and following an optimal strategy up to the end. Optimal strategy at time point j can be followed by choosing $t_j^{opt} = 1$ if a current return $r(m_j)$ is bigger than expected return $E[f_{j+1}(m_{j+1})]$. The maximal expected return if starting in time point j is:

$$E[f_{j+1}(m_{j+1})] = \sum_{m_{j+1}=1}^{3} f_{j+1}(m_{j+1})p(m_{j+1})$$

and the DP recursion is given by

$$f_j(m_j) = \max\{r(m_j), E[f_{j+1}(m_{j+1})]\} \quad \text{where} \qquad j = 1, 2, 3, 4.$$

We start at the end.

Time 4:

$$\begin{aligned} f_4(m_4) &= \max[r(m_4), E[f_5(m_5)] \\ &= \max(r(m_4), 0) \\ &= r(m_4) \quad \text{where} \qquad r(m_4) = \{r_1, r_2, r_3\} \end{aligned}$$

In this time point $d^{opt} = 1$ for all $m_4 = \{1, 2, 3\}$ as it is the final time point.

Time 3:

Here we calculate the expected value of returns at time 4 and compare it with $r(m_3)$

$$\begin{aligned} E[f_4(m_4)] = \sum_{m_4=1}^{3} \quad f_4(m_4)p(m_4) &= f_4(r_1) \times p_1 + f_4(r_2) \times p_2 + f_4(r_3) \times p_3 \\ &= 0.1 \times 0.3 + 0.2 \times 0.5 + 0.3 \times 0.2 = 0.19 \\ f_3(m_3) &= \max[r(m_3), E[f_4(m_4)] \\ &= \max[r(m_3), 0.19] \\ \text{Hence} \qquad f_3(m_3) &= \{r_2, r_3\} \quad \text{when} \;\; m_3 = \{2, 3\}, \\ \text{and} \qquad f_3(m_3 = 1) &= 0.19 \end{aligned}$$

Because 0.19 only bigger than 0.10. Thus optimal decisions are $d_3^{opt} = 0$ if $m_3 = 1$. Otherwise the physician must treat the patient.

Time 2:

$$\begin{aligned} E[f_3(m_3)] = \sum_{m_3=1}^{3} \quad f_3(m_3)p(m_3) &= f_3(r_1) \times p_1 + f_3(r_2) \times p_2 + f_3(r_3) \times p_3 \\ &= 0.19 \times 0.3 + 0.2 \times 0.5 + 0.3 \times 0.2 = 0.217 \\ f_2(m_2) &= \max[r(m_2), E[f_3(m_3)] \\ &= \max[r(m_2), 0.217] \\ f_2(m_2) &= 0.217 \quad \text{if} \;\; m_2 = \{1, 2\}, \\ f_2(m_2 = 3) &= 0.3 \end{aligned}$$

Time 1:

$$\begin{aligned}
E[f_2(m_2)] = \sum_{m_2=1}^{3} f_2(m_2)p(m_2) &= f_2(r_1) \times p_1 + f_2(r_2) \times p_2 + f_2(r_3) \times p_3 \\
&= 0.217 \times 0.3 + 0.217 \times 0.5 + 0.3 \times 0.2 = 0.2336 \\
f_1(m_1) &= \max[r(m_1), E[f_2(m_2)]] \\
&= \max[r(m_1), 0.2336] \\
f_1(m_1) &= 0.2336 \quad \text{when} \quad m_1 = \{1, 2\}, \\
f_1(m_1 = 3) &= 0.3
\end{aligned}$$

Optimal decisions at time point 1 are the same with those in time point 2. Table 2.8 explains the expected return from time point one to four and what the optimal decisions are.

		$d_j^{opt}(m_j)$			
Time point j	$E[f_{j+1}(m_{j+1})]$	$m_j = 1$	$m_j = 2$	$m_j = 3$	Reason
1	0.2336	0	0	1	$0.3 > 0.2336$
2	0.217	0	0	1	$0.3 > 0.217$
3	0.19	0	1	1	$0.2, 0.3 > 0.19$
4	0	1	1	1	$0.1, 0.2, 0.3 > 0$

Table 2.8: Maximum expected return and optimal decision strategies of the three states.

Here the physician must treat the patient at any time point if $m_j = 3$, at time points 3 and 4 if $m_j = 2$, but only at time point 4 if $m_j = 1$.

Example 2

Suppose a physician has to treat a patient randomly at three time points, either $(T = 1)$ with probability $1 - p$, or $(T = 2)$ with probability p. Treatment t which is binary $\{0, 1\}$, is to be given subsequent to observing the status. If the patient was treated at time j, after observed M_j, then different random returns (depending on the probability of $\theta(T)$) can be occurred. Let us assume that $R_j(t_j|T, \theta)$ be a set of treatment return at time point j, where $j = \{1, 2, 3\}$. The patient's status starts with the baseline covariate M_1 units. So we can obtain that $M_{j+1} = M_j + R_j$. The next figure shows four potential outcomes at the end of each time point j

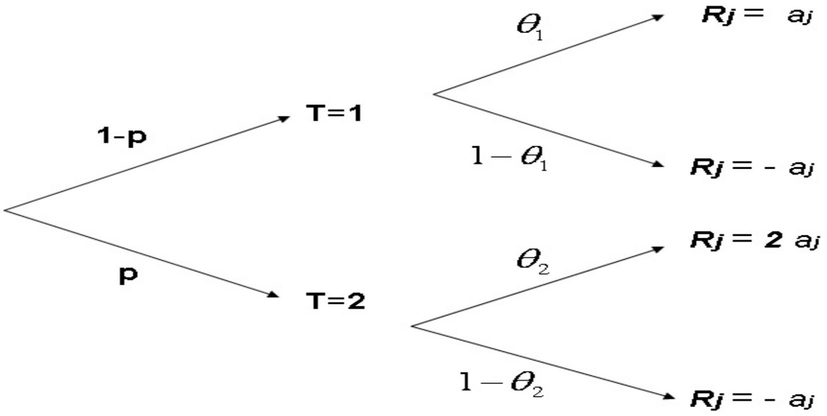

Figure 2.2: Treatment random returns at each time point j

$$\begin{aligned}
R_j(t_j &= 1|T = 1, \theta_1) = a_j \\
R_j(t_j &= 1|T = 1, 1 - \theta_1) = -a_j \\
R_j(t_j &= 1|T = 2, \theta_2) = 2a_j \\
R_j(t_j &= 1|T = 2, 1 - \theta_1) = -a_j,
\end{aligned}$$

and $R_j(t_j = 0|T, \theta) = 0$. Can we find the optimal strategy which maximises the patient expected final response?

We use the following optimality equations:

$$M_{j+1} = \max \; [\, M_j, E(M_{j+1})]$$

$$E(M_{j+1}) \quad = \quad M_j + E(R_j)$$

The idea of this example is to treat the patient only if there is a positive return $E(R_j) > 0$, but if $E(R_j) < 0$ then $M_j > E(M_{j+1})$ and the optimal decision is not to treat the patient. Note that patient status will be fixed if the patient is not treated, then $M_{j+1} = M_j$.

$$\begin{aligned} E(R_j) &= [(1-p)(1-\theta_1)(-a_j) + (1-p)\theta_1 a_j + (p)(1-\theta_2)(-a_j) + p\theta_2(2a_j)] \\ &= a_j[(1-p)(2\theta_1 - 1) + p(3\theta_2 - 1)] \end{aligned}$$

Case 1

Supppose a_j is an amount that depends on the current state M_j say. $M_1 = 800, p = 0.5, \theta_1 = 0.4, \theta_2 = 0.7, a_j = 0.25M_j$,

$$\begin{aligned} E(R_j) &= a_j[(1-p)(2\theta_1 - 1) + p(3\theta_2 - 1)] \\ &= a_j[0.5 \times (2 \times 0.4 - 1) + 0.5 \times (3 \times 0.7 - 1)] = 0.45a_j. \end{aligned}$$

As seen $E(R_j) = 0.45a_j$ and $a_j = 0.25M_j$. So based on these details, $E(R_j) > 0$ at all j, then the optimal decision is to treat the patient. Because $E(M_{j+1}) > M_j$ at all j as shown below

At time point 1,

$$\begin{aligned} M_2 &= \max \quad [\, M_1, E(M_2)], \text{ and } E(M_2) = M_1 + E(R_1), \text{ then} \\ M_2 &= \max \quad [\, 800, 890] = 890. \end{aligned}$$

At time point 2,

$$\begin{aligned} M_3 &= \max \quad [\, M_2, E(M_3)], \text{and } E(M_3) = M_2 + E(R_2), \\ M_3 &= \max \quad [\, 890, 990.125] = 990.125. \end{aligned}$$

At time point 3,

$$\begin{aligned} M_4 &= \max \quad [\, M_3, E(M_4)], \text{and } E(M_4) = M_3 + E(R_3), \\ M_4 &= \max \quad [\, 990.125, 1101.514] = 1101.514. \end{aligned}$$

If the physician follows the optimal strategy and treat the patient at the start of all decision points, then the maximum expected return will be 1101.5 units.

Case 2

Now, let $p = 0.05$ then $E(R_j) = -0.135a_j$ and $M_j > E(M_{j+1})$, at all time points. The optimal decision will be to choose M_j, that means he or she has to follow a fixed dynamic strategy (no treatment for all j).

Case 3

If p or θ have different values, the strategy will change at each decision point, e.g., suppose p and θ have the previous example values except $\theta_2 = 0.2$ when $j = 2$ Then $E(R_1) = E(R_3) = 0.45a_j$, but $E(R_2) = -0.3a_j$, and

$$\begin{aligned} M_2 &= \max \; [\,800, 890] = 890, \\ M_3 &= \max \; [\,890, 823.25] = 890, \\ M_3 &= \max \; [\,890, 990.125] = 990.125. \end{aligned}$$

As we see, we have a non fixed dynamic strategy. To maximize the expected return at the end of the final stage, we have to treat the patient at $j = 1, 3$ but not at the second time point.

2.4 Causal Effects

All optimal dynamic treatment strategies considered later, e.g., G-estimation, inverse probability treatment weighted, etc. are approaches to estimating the causal effect of a time-varying treatment on time to some event of interest. Because these approaches are designed for a situation where the treatment may have been repeatedly adapted to patient characteristics, which them selves may also be time-independent. The definition of cause is complex and difficult, but for empirical research, the concept of the causal effect of a treatment seems more straightforward and practically useful (Roderick and Rubin 2000). A key idea is the explanation of causal effects through potential outcomes. Causal effects are comparisons of the potential outcomes that would have been observed under different exposures of patients to treatments. In a studied example, several epidemiological studies showed that women who were taking combined hormone replacement therapy had a lower average incidence of coronary heart disease, leading doctors to propose that hormone replacement therapy was protective against coronary heart disease (Lawlor et al. 2004). But controlled trials showed that hormone replacement therapy caused a significant increase in risk of coronary heart disease. Re-analysis of the data showed that women undertaking hormone replacement therapy were more likely to be from higher socio-economic groups, with better than average diet and exercise regimes. The two were immediate effects of a common cause, rather than cause and effect as had been believed (Roderick and Rubin 2000).

Rubin defines a causal effect of one treatment, t_a, over another, t_b, for a particular patient and an interval of time from j_1 to j_2 as the difference between what would have happened at j_2 if the patient had been exposed to t_a initiated at j_1 and what would have happened at j_2 if the unit had been exposed to t_b initiated at j_1: our definition of the causal effect of the t_a versus t_b treatment will reflect this intuitive meaning.

2.4.1 Longitudinal data

Longitudinal data groups are comprised of repeated measurements of an outcome and a set of covariates for each of many units. One aim of statistical analysis is to model and estimate the marginal expectation of the response variable as a function of the covariates while accounting for the correlation among the repeated observations for a given unit (Scott and Kung 1986). Longitudinal data have the form of repeated observations on the same subject over time. In longitudinal clinical trials, each patient's information, treatment offered, and response to it will be recorded at each decision point in the patient's treatment. For example, we may measure the amount of HIV virus present in the body at three monthly time intervals on patients with HIV infection. Patients are assigned to take different treatments at the start of the study (Davidian 2006). The scientific questions of importance often involve not only the common kinds of questions, such as how the mean response differs across treatment, but also how the change in mean outcome over time differs and other issues regarding the relationship between response and time. Thus, it is necessary to represent the situation in conditions of a statistical model that acknowledges the way in which the data were collected in order to address these questions. Complementing the models, specialized methods of analysis are required. Although the term longitudinal naturally suggests that data are collected over time, the models and methods we will discuss are more generally applicable to any kind of repeated measurement data. That is, although repeated measurement most often takes place over time, this is not the only way that measurements are taken repeatedly on the same unit. For example, units may be human subjects. For each subject, reduction in diastolic blood pressure is measured on several occasions, each occasion involving administration of a different dose of an anti-hypertensive medication. Thus, the subject is measured repeatedly over dose (Davidian 2006).

2.4.2 Counterfactuals

We refer to the outcome of the model as counterfactual (potential outcome), because it is defined under conditions contrary to fact; that is, in reality, not all subjects followed a given exposure history. Models for counterfactual outcomes are known as structural or causal models (Hernán 2005). A counterfactual is a potential outcome, prior to the actual outcome being observed. It is defined as a person's outcome if he had followed a particular treatment regime, which is possibly different from the regime that he was actually observed to follow. The causal effect of a regime may be seen as the difference in outcome if he had followed that regime as compared to a placebo regime or a standard care protocol.

Definition *Let T_i be the causal variable (treatment) of interest for unit i where $i = 1, 2, \cdots, n$. T_i takes a value in a set τ. The potential outcome $\{Y_i(t),\ t \in \tau\}$ represents the outcome that would be observed for unit i if it receives the treatment whose value is t, i.e., $T_i = t$ for $t \in \tau$.*

Hence, for each individual i its set of potential outcomes $\{Y_i(t),\ t \in \tau\}$. From all potential outcomes, only one of them corresponding to the actual treatment T_i can be observed. We use Y_i to denote the observed outcome for unit i. The treatment variable determines which of the potential outcomes will be revealed. This can be seen, for example, from the fact that if the treatment is binary, the observed outcome is given by $Y_i = T_iY_i(1)+(1-T_i)Y_i(0)$.

2.4.3 Observational studies, randomized trials and causal effects

For a long time it has been claimed that observational studies can find stronger treatment effects than randomized trials, (Rosenbaum 2002). In a randomized controlled trial is allocated each subject is randomly assigned to a treatment group or control group before the start of treatment. Estimating and comparing the effects of dynamic treatment regimes from a sample of observed trajectories of treatments and outcomes depends on the assumption that new treatments are assigned independently of potential future responses to treatment, conditional on the history of treatments and response to date, called *sequential ignorability.* In longitudinal observational studies, the assumption of sequential ignorability must be assumed, while randomization of dynamic regimes can guarantee it (Lavori 2001).

In randomized trials, individuals are randomly assigned to a treatment and control group and if the groups are different significantly after treatment the treatment is assumed to cause the difference. In observational studies, under the assumption that the treatment is independent of potential outcomes, it is possible to test whether the treatment has an effect and estimate the mean counterfactual treatment effect.

Definition *Let T_i be a binary treatment variable. The causal effect can be defined for each unit as the difference: $Y_i(1) - Y_i(0)$.*

Definition *Let T_i be a binary (random) treatment variable for unit i where $i = 1, 2, ..., n$. Consider fixed (i.e., non-random) but possibly unknown potential outcomes, $Y_i(1)$ and $Y_i(0)$, for each i. Then, the following sample average causal effects of interest can be defined by $\frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0)$.*

Definition *The treatment is said to be randomized if the treatment variable T_i is independent of all potential outcomes, $Y_i(t)$, for all units, i.e., $Y_i(t) \perp T_i$ for all t and all i* (Robins et al. 1997).

2.4.4 Time-varying treatment

If a variable changes over time, we called it a time-varying variable, for example, sex is a non time-varying variable. A time-varying covariate is a term used commonly in survival analysis. It means that a covariate is not necessarily fixed. If a person is treated at many time points with different doses, then the dose is a time-varying covariate. Or, if one wants to examine the link between area of residence and cancer, this would be complicated by the fact that study subjects might move from one area to another. The area of residency might then be introduced in the statistical model as a time-varying covariate.

Time-varying treatment regimes are treatment policies where the type of treatment or the level of treatment changes over time. On the other hand there are care policies with fixed regimes, where the same dose is maintained. Suppose, for example, there are several ways to treat a patient. A fixed regime would be achieved, for example, by taking the same dose twice a day at the same time for 4 weeks. In contrast, another regime would be to take different combinations of drugs over four weeks and this can be considered a time-varying regime. Note that a time-varying regime can be dynamic or not. The last example was not, but a dynamic treatment regime is one in which the level of treatment received depends on time-varying patient diagnosis.

Now, we consider the following examples.

- Varying doses of warfarin for anticoagulation are needed for the drug to be effective in each patient. Too much warfarin can lead to severe bleeding, and too little can cause risky blood clots. Historically, there has been no guidance for predicting how much of the drug a person will need. Physicians have had to approximately estimate an initial dose of warfarin and then continually monitor the patients International Normalized Ratio (INR) value which is a measure of how fast the blood clots.
- Measures of amount of HIV virus present in body may be taken during a year at three monthly time intervals on patients with HIV infection. Patients are assigned to take different treatments at the start of each time interval.
- Chemotherapy for cancer and control of its side-effects is a dynamic treatment regime.

2.4.5 Assumptions for causal inference

Three assumptions sufficient to identify the average causal effect are *consistency*, *exchangeability* (e. g., no unmeasured confounders) and *positivity*. Generally those hold in a randomized experiment, Rosenbaum and Rubin (1983):

Assumption 1: Consistency

The potential outcome under any particular treatment or action regime identical to the actual outcome if that regime is followed. This states that the results of a subject's treatment allocation are not affected by other subjects' treatment allocation, Formally, let T be an n dimensional vector of treatment assignment, whose i^{th} element represents the treatment value of unit i where $i = 1, 2, ..., n$. Let $Y_i(T)$ be the potential outcome of unit i given the treatment assignment for all units, i.e., T. Then, the assumption implies that $Y_i(T = a) = Y_i(T = b)$ whenever $a_i = b_i$ (Rosenbaum and Rubin 1983).

Assumption 2: No unmeasured confounders

Let M be a set of pre-treatment covariates. Then the assumption says that any regime t of T, received in any interval is conditional on history, but is independent of any future potential outcome. This means $Y(t) \perp T | M = m$ for each possible value t of T and m of M. This assumption is sometimes referred as the conditional exchangeability (sequential randomization). It holds in a randomized experiment in which treatment was randomly assigned (Robins et al. 1997).

Assumption 3: Positivity

The treatment is not deterministically allocated within any level m of the covariates M. That is, not all source population subjects with a given value m of M are assigned to be treated or untreated (Hernán and Robins, 2006). If $P(M = m) \neq 0$ (the population marginal probability that M takes the value m) then $P(T = t|M = m) > 0$ (the conditional probability that T takes the value t among subjects in the population with M equal to m. [The above assumed M and T were discrete variables.]

Without additional assumptions, the optimal regime might be estimated from among the set of feasible regimes (Robins 1994); feasibility requires some subjects to have followed a special regime to make non-parametric inference from this regime, using Robins (1997), the sequential version of no unmeasured assumption at time j,

$$T_j \perp M_{j+1}(\bar{t}_j), \cdots, M_K(\bar{t}_{K-1}), Y(\bar{t}_K) | \bar{M}_j, \bar{M}_{j-1}$$

and at the last time point K,

$$T_K \perp Y(\bar{t}_K) | \bar{M}_K, \bar{M}_{K-1}.$$

2.5 Causal Directed Acyclic Graphs (DAGs):

2.5.1 DAGs for a single time point

Unconditional exchangeability and conditional exchangeability can be translated into the language of causal directed acyclic graphs or DAGs, Pearl (1995); Spirtes, Glymour and Scheines (1993). The causal DAGs have to include all common causes of its variables. Now consider the three causal DAGs of Figure 2.3, Robins, Hernán and Brumback (2000) in which M and U represent measured and unmeasured baseline covariates.

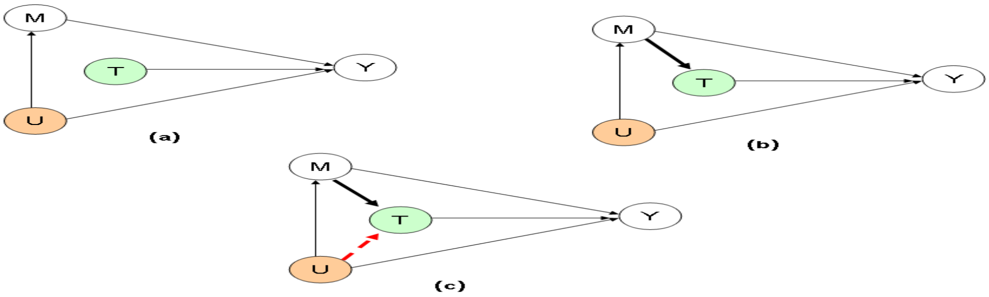

Figure 2.3: Causal directed acyclic graphs.

The causal DAG in Figure 2.3a can represent a randomized experiment for which each subject is randomized to be treated with the same probability $P(T = 1)$. Thus the conditional probability of treatment does not depend on M or U , i.e.,

$$P(T = 1|M = m, U = u) = P(T = 1).$$

We then say that there is no confounding by measured variables M or unmeasured variables U. Equivalently, $T = 1$ and $T = 0$ are unconditionally exchangeable (i.e., $Y(t) \perp T$) because the treatment T and the outcome Y do not share any common causes. When unconditional exchangeability holds, $E[Y(t)]$ equals the mean $E[Y|T = t]$ of the study population actually treated with t.

The causal DAG in Figure 2.3b represents a randomized experiment in which each subject is randomized to be treated with probability $Pr(T = 1|M = m)$ that depends on the subject's value of M but not on U , i.e.,

$$P(T = 1|M = m, U = u) = Pr(T = 1|M = m).$$

For example, we might decide to treat a greater proportion of smokers than non-smokers. We then say that there is confounding but there is not unmeasured confounding. Equivalently, the $T = 1$ and the $T = 0$ are not unconditionally exchangeable. In this setting, $E[Y(t)]$ does not equal the mean $E[Y|T = t]$. However, $E[Y(t)|M = m] = E[Y|T = t, M = m]$. Furthermore, by using data on M, $E[Y(t)]$ can be consistently estimated.

The causal DAG in Figure 2.3c represents a study in which the conditional probability of treatment $P(T = 1|M = m, U = u)$ depends on the unmeasured variables U as well as the measured variables M and thus cannot possibly represent a randomized experiment. We say that there is unmeasured confounding. Equivalently, $T = 1$ and $T = 0$ are not conditionally exchangeable given M because we cannot block all non causal associations between treatment and outcome by conditioning on the measured covariates M. In this setting neither $E[Y(t)|M = m]$ nor $E[Y(t)]$ can be consistently estimated, at least not without further strong assumptions.

2.5.2 DAGs for multi time points (time-varying treatment)

To develop methods for the estimation of the causal effects of a time-varying treatment, we need to generalize the definition of causal effect and the three identifiability conditions of the previous section.

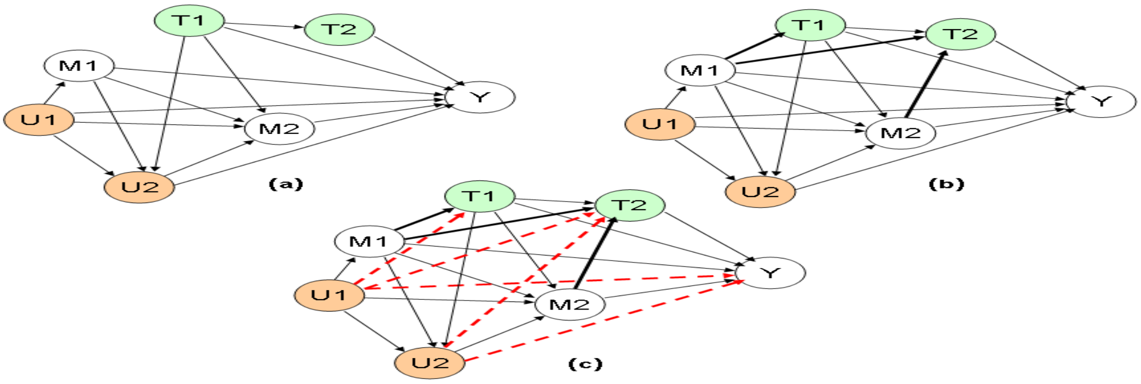

Figure 2.4: Causal directed acyclic graphs with two time point treatments.

Let us assume K time points. Specifically, the generalized identifiability conditions are explained in Section 2.4.5. The three conditions generally hold in ideal sequentially randomized experiments with full compliance. A sequentially randomized experiment is a randomized experiment in which the treatment value at each successive visit k is randomly assigned with known randomization probabilities (bounded away from 0 and 1) that, by design, may depend on a subject's past treatment $\bar{T}_{k-1}$ and covariate history $\bar{M}_k$ through k. As for fixed treatment, exchangeability and conditional exchangeability can be represented by causal DAGs. The DAGs in Figures 2.4a to 2.4c are the time-varying analogs of those in Figures 2.3a to 2.3c (Robins and Greenland 2000). The causal DAG in Figure 2.4a implies unconditional or marginal exchangeability. It represents a sequentially randomized experiment in which the randomization probabilities at each time j depend at most on a subject's past treatment history, which is the proper generalization of no confounding by measured or unmeasured variables to a sequentially randomized experiment. Figure 2.4*b* represents a sequentially randomized experiment in which the randomization probabilities at each time j depend on previous history. There is confounding by measured covariates but no unmeasured confounding. Thus the three identifiability conditions hold. Figure 2.4*c* represents a study case in which the probability of treatment depends on unmeasured confounding variables U that cause Y and cannot possibly represent a sequentially randomized experiment. Thus causal effects cannot be consistently estimated (Hernán and Robins 2006).

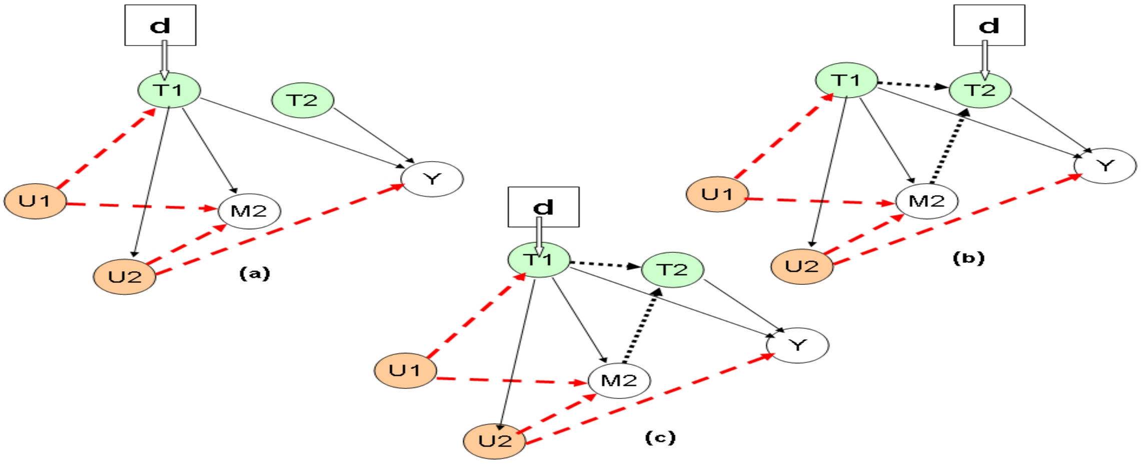

Figure 2.5: Examples of causal directed acyclic graphs with unmeasured variables when stability holds.

Dawid and Didelez (2008) show that *conditional simple stability* is violated if the unmeasured variables U cause decision variables T and the final response Y (as explained in Figure 2.4*c*). Simple stability is closely related to the no unmeasured confounders assumption (Robins, 1997). But they explained that even when there are arrows from U_1 into both of T_1 and Y, stability can be satisfied under specific circumstances such as

- (Figure 2.5*a*) assuming an unconditional intervention in T_2, i.e., no parent set for the action T_2. We can note that $Y \perp d | T_1$. That means the decision rules at $j = 1$, are chosen independently of Y. At $j = 2$, we can easily see that $Y \perp d | T_1, M_2, T_2$, because the decision rules are chosen without taking previous history into account (no arrows into T_2). This strategy is identifiable even though simple stability is violated.

- As shown in Figure 2.5b, if we assume that the intervention strategy in T_2 does depend on its parent set (back dotted lines), i.e. (T_1, M_2), but no arrow exist from U_2 into T_2, thus $Y \perp d | T_2$. In contrast, in the same example, Figure 2.5c shows at $j = 1$, that $Y \perp d | T_1$ is violated and we cannot guarantee that such a conditional strategy is identifiable.

2.5.3 Observational studies and the identifiability assumptions for causal inference

A difference between randomized experiments and observational studies is that the conditional probability of treatment is not known in the latter and thus needs to be estimated from the data.

The major weakness of observational studies is that, unlike in randomized experiments with full compliance, the three identifiability conditions are not guaranteed by design. Conditional exchangeability will not hold if there is unmeasured confounding. Unfortunately, the presence of conditional exchangeability cannot be empirically tested. Even consistency cannot always be taken for granted in observational studies because the counterfactual outcomes themselves are sometimes not well defined, which renders causal inferences ambiguous (Robins and Greenland, 2000; Hernán, 2005). Thus, in observational studies, an investigator who assumes these conditions hold may be mistaken; hence, causal inference from observational data is a risky business.

2.6 Discussion

As shown, dynamic programming (Bellman 1957) is an optimization procedure that is designed to efficiently search for the global optimum of a function; it is an algorithmic technique based on a recurrent formula and some starting states. The problem can usually be divided into stages with a decision required at each stage. Each stage has a number of states associated with it. There exists a recursive relationship that identifies the optimal decision for stage j, given that stage $j + 1$ has already been solved, and the final stage must be solvable by itself. The dynamic programming solutions have a complex polynomial function which assures a much faster running time than other techniques. The most important and difficult issue in dynamic programming is to take a problem and determine stages and states so that all of the above characteristics hold. More details and examples can be found in Cormen et al. (1990).

In the study of dynamic treatment regimes, we need to model the longitudinal distribution of all the covariates and outcomes. However, the information required is not always available. Lack of information in such situation is ascribed either to misspecifying the distributions or the treatment may be mistakenly recommended. As we will see later, methods given by Murphy (2003) and Robins (2004) do not suffer from this problem. Robins (1987) used the so called *theory of causal inference* to assess the direct and indirect effects of time varying treatments based on experimental and observational longitudinal studies. The potential outcome is the value of a status that would happen to a patient under different treatments. In the next chapter we review selected successful applications of treatment decisions in the literature.

Chapter 3

Optimal Dynamic Treatment Regimes

3.1 Introduction

The problem of finding the optimal treatment regime is one of sequential decision-making, where a treatment which appears optimal in the short-term may not be a component of the optimal regime (Lavori 2001). A regime is said to be optimal if it maximizes the mean response at the end of final interval. Dynamic regimes are also called tailored communications, adaptive interventions, or adaptive strategies (Murphy 2003). The problem of estimating treatment effects from observational studies has broad applicability in public health, economics and social sciences. Robins and his colleagues, e.g., Robins et al. (1999), Robins, Hernán and Brumback (2000) and Hernán, Brumback and Robins (2002) have written extensively on the use of a marginal structural models for this purpose with focus primarily on functions of the mean of repeated measures, e.g., Hernán, Brumback and Robins (2002) and on hazard functions for event histories.

3.2 Optimal Dynamic Treatment Regime Approaches

In this chapter, our purpose is to understand in detail the different approaches to optimal dynamic treatment regimes and to investigate the relationships between them (we use and develop some material from Moodie et al. (2007)). We will study two approaches that have been proposed in the literature, namely: Robins G-estimation and Murphy iterative minimization. We hope to show the similarities that are shared between what a first glance might seem like very different approaches. To explain the basics of optimal dynamic treatment regime, let us assume that a physician has to treat n patients dynamically, at K fixed time points, with a binary treatment, either $(T_j = 0)$ with no treatment, so he or she should avoid the side-effect of the drug, or $(T_j = 1)$ with treatment, when the dose toxicity is less than the dose efficacy. M_1 is the baseline covariate and treatments T_j are to be given subsequent to observing status M_j (the level of health). As an example, let the potential final response Y be determined by the equation

$$Y(t_K|\bar{M}_K, \bar{T}_{K-1}) = \beta_K M_1 + \sum_{j=1}^{K} \gamma_j(M_j, T_j, \psi),$$

where $\gamma_j(M_j, T_j, \psi) = (K-j+1)(\psi_{j0}+\psi_{j1}M_j)^{T_j}$. For $j = 2, \cdots, K$, M_{j+1} is a linear function on both M_j and T_j. E.g., $M_{j+1} = \alpha_j + \beta_j M_j + \eta_j T_j$. In this example, for simplicity we will use each of α_j and η_j as a set of zero values. Suppose $K = 4$, $\psi_{j0} = \{75, 320, 60, 21\}$, $\psi_{j1} = \{-1, -2, -0.5, -0.1\}$, $\bar{\beta}_j = \{1.5, 1.6, 1.25, 6\}$ and we have five patients, that their baseline covariates are $M_1 = \{80, 78, 72, 68, 100\}$. How can we find the the optimal strategy which maximises the mean final response?

We can calculate the final response of patient i with always-treat strategy and with never-treat strategy, for all j. The mean difference between these treatment strategies is $E(Y(1) - Y(0))$ which is shown in Table 3.1 to be estimated by $\frac{1}{5}[3144.4 - 2438] = 141.28$.

Subject i	M_1	M_2	M_3	M_4	$Y(0)$	$Y(1)$
1	80	120	192	240	490	625.0
2	78	117	187.2	234	478	644.4
3	72	108	172.8	216	442	702.6
4	68	102	163.2	204	418	741.4
5	100	150	240	300	610	431.0
Total					2438	3144.4

Table 3.1: Treatment effects

Now, to maximise $Y(\bar{T} = \bar{t})$ we should follow the optimal policy,

$$d_j^{opt} = \max_{T_j} \gamma_j(M_j, \psi),$$

that means we should treat the patient only when $\psi_{j0} + \psi_{j1}M_j \geq 1$, equivalently, when $M_{ji} < \frac{1-\psi_{j0}}{\psi_{j1}}$, where M_{ji} is the state of patient i at time j. For this data, optimal treatment strategies are $\{0, 1, 0, 0\}$ for patients one, two and five and $\{1, 1, 0, 0\}$ for other patients. Optimal strategies for all patients are shown in the following table.

Subject i	$\bar{T}_4^{opt}$	$\sum \gamma_j(\bar{M_4}, \bar{T}_4^{opt}, \psi)$	$Y_i(\bar{T}_4^{opt})$
1	$\{0, 1, 0, 0\}$	247	727
2	$\{0, 1, 0, 0\}$	265	733
3	$\{1, 1, 0, 0\}$	327	759
4	$\{1, 1, 0, 0\}$	379	787
5	$\{0, 1, 0, 0\}$	67	667
Total			3673

Table 3.2: Optimal treatment strategies

The mean final response when following optimal treatment strategies is $E(Y_i(T^*) - Y_i(0)) = \frac{1}{5}(3673 - 2438) = 247$, which is better than fixed treatment strategy.

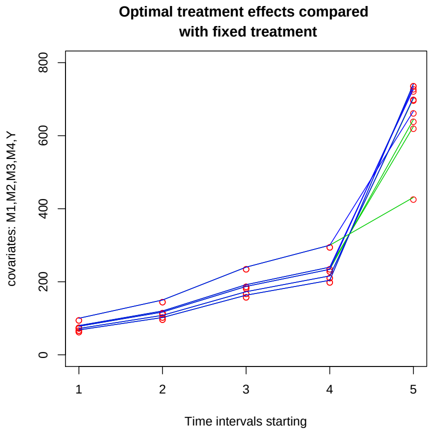

Figure 3.1: States and compared always-treat strategy with optimal strategy on $Y_i(t)$.

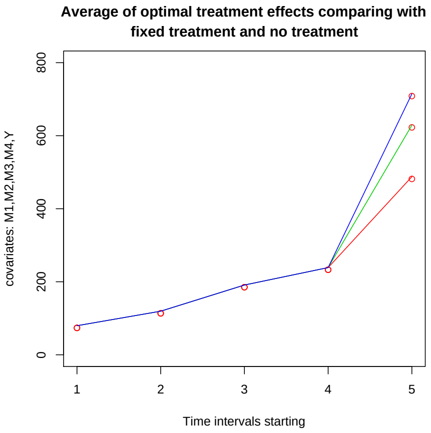

Figure 3.2: Mean states and compared different strategies on the expected mean responses.

Figure 3.1 explains the states and the optimal treatment response for each patient compared with standard regime and Figure 3.2 shows the means of states and improvements on the mean final response when using optimal treatment strategies (blue line) instead of always-treat strategies (green line), and with no-treat strategies (red line). In this example, there are five fixed initial states, binary treatment and four time points. So for each patient i we have 2^4 potential outcomes, e.g., the potential outcomes of patient 3. The following table shows that only one path maximizes the final outcome. Things are often of course more

$\bar{T}_4$	$E[Y(\bar{T}_4)]$	$\bar{T}_4$	$E[Y(\bar{T}_4)]$	$\bar{T}_4$	$E[Y(\bar{T}_4)]$	$\bar{T}_4$	$E[Y(\bar{T}_4)]$
0,0,0,0	442.0	0,1,0,0	751.0	1,0,0,0	450.0	1,1,0,0	759.0
0,0,0,1	440.4	0,1,0,1	749.4	1,0,0,1	448.4	1,1,0,1	757.4
0,0,1,0	387.2	0,1,1,0	696.2	1,0,1,0	395.2	1,1,1,0	704.2
0,0,1,1	385.6	0,1,1,1	694.6	1,0,1,1	393.6	1,1,1,1	702.6

Table 3.3: All possible potential outcomes of patient 3

complicated. For example, we might have random states, e.g., we generate a dataset with $M_j \sim N(\frac{1-\psi_{j0}}{\psi_{j1}}, \sigma^2_{M_j})$. If we divide the states into two parts $M_j < \frac{1-\psi_{j0}}{\psi_{j1}}$ and $M_j \geq \frac{1-\psi_{j0}}{\psi_{j1}}$ then we will face 2^8 different potential outcomes.

The particular problem for optimal dynamic treatment strategies with time varying covariates means that M_j, as intermediate variables depend on earlier treatment and confounders for later treatment decisions. That is why standard regression method gives biased estimators and the alternatives of the regret functions of Murphy (2003) or the blips of G-estimation have been proposed.

3.2.1 The blip function model

This function is used by Robins (2004) to find an optimal regime. It is based on structural nested mean models (SNMM), Robins (1986). Robins defines a SNMM as an expected difference between a person's counterfactual responses on a specific treatment regime from time $j+1$ onwards and on another specific regime from time j conditional on history. There are subclasses of SNMM's, which are called the blip functions. We define an optimal blip-to-reference function as the expected difference in outcome when using a reference regime instead of t_j at time j, in persons who subsequently receive the optimal regime with treatment and covariate history.

Let a reference regime be denoted by $d_j{}^{ref} = d_j{}^{ref}(\bar{m}_j, \bar{t}_{j-1})$. Then let $\underline{d}^{opt}_{j+1}$ be an optimal future regime after time j based on $\bar{t}_j$. The blip to reference at the j^{th} time point will be as follows,

$$\gamma_j^{ref}(t_j|\bar{M}_j, \bar{T}_{j-1}) = E(Y(t_j, \underline{d}^{opt}_{j+1})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(d_j^{ref}, \underline{d}^{opt}_{j+1})|\bar{M}_j, \bar{T}_{j-1})$$

At the K^{th} time point,

$$\gamma_j^{ref}(t_K|\bar{M}_K, \bar{T}_{K-1}) = E(Y(t_K)|\bar{M}_K, \bar{T}_{K-1}) - E(Y(d_K^{ref})|\bar{M}_K, \bar{T}_{K-1}))$$

There is also another subclass of SNMMs, called *optimal blip-to-zero* functions, ${\gamma_j}^{zero}$ defined as the expected difference in outcome when using the zero regime (should be thought of as a standard care), instead of treatment t_j, in people who consequently receive the optimal regime with treatment and covariate history.

$$\gamma_j^{zero}(t_j|\bar{M}_j, \bar{T}_{j-1}) = E(Y(t_j, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(d_j^{zero}, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1})$$

3.2.2 The regret function model

Murphy (2003) modelled the regret function, which is based on structural nested mean models (SNMM) (Robins 1986), and it is the negative of the optimal blip that uses the optimal treatment at time j as the reference regime. Let the regret be

$$\mu_j = \mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}).$$

The regret function at the j^{th} time point will be as follows

$$\mu_j = E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(t_j, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) \tag{3.1}$$

So if we have three time points:

$$\begin{aligned}
\mu_1 &= E(Y(\underline{d}_1^{opt})|M_1) - E(Y(t_1, \underline{d}_2^{opt})|M_1), \\
\mu_2 &= E(Y(\underline{d}_2^{opt})|\bar{M}_2, T_1) - E(Y(t_2, d_3^{opt})|\bar{M}_2, T_1), \\
\mu_3 &= E(Y(d_3^{opt})|\bar{M}_3, \bar{T}_2) - E(Y(t_3)|\bar{M}_3, \bar{T}_2).
\end{aligned}$$

The regret at t_j is the expected difference in the outcome had the optimal treatment been taken at j, instead of treatment t_j, in participants who followed the observed regime up to t_j and the optimal regime from time $j+1$ onwards. Since we hope to maximize Y, by definition the regret is non-negative and equals 0 when the optimal action is chosen. When a non-optimal decision is chosen, the regret quantifies the loss in the mean of Y

assuming that optimal decisions are made in the future, conditional on the history to time j. Murphy suggests defining models for the regrets at each time point, these models being non-parametric, semi-parametric or parametric.

3.2.3 Correspondence between the methods

Optimal blip functions and regrets correspond directly. As we have seen the optimal blip-to-reference function at the j^{th} interval is

$$\begin{aligned}
\gamma_j^{ref} &= E(Y(t_j, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(d_j^{ref}, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) \\
&= E(Y(t_j, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) \\
&+ E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(d_j^{ref}, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1})
\end{aligned}$$

Then

$$\begin{aligned}
\gamma_j^{ref} &= E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(d_j^{ref}, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) \\
&- [E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(t_j, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1})]
\end{aligned}$$

Thus

$$\gamma_j^{ref}(t_j|\bar{M}_j, \bar{T}_{j-1}) = \mu_j(d_j^{ref}|\bar{M}_j, \bar{T}_{j-1}) - \mu_j(t_j|\bar{M}_j, \bar{T}_{j-1})$$

Now, consider the regret function:

$$\mu_j = E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(t_j, \underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1})$$

We can write

$$\begin{aligned}
\mu_j &= E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(d_j^{ref}, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) \\
&- E(Y(t_j, \underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) + E(Y(d_j^{ref}, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1})
\end{aligned}$$

Then

$$\begin{aligned}\mu_j &= E(Y(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y(d_j^{ref}, \underline{d}_{j+1}^{opt})|\bar{M}_j, \bar{T}_{j-1}) \\ &- [E(Y(t_j, \underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1}) - E(Y d_j^{ref}, \underline{d}_{j+1}^{opt}|\bar{M}_j, \bar{T}_{j-1})]\end{aligned}$$

Hence

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = \max_t \gamma_j^{ref}(t|\bar{M}_j, \bar{T}_{j-1}) - \gamma_j^{ref}(t_j|\bar{M}_j, \bar{T}_{j-1})$$

In the next section, we will estimate the parameters ψ of the optimal blip-to-zero function and the regret function and investigate the robustness theoretically and through a simulation study.

3.3 Estimation of Optimal Dynamic Treatment Regimes

3.3.1 G-estimation

G-estimation of structural nested models is a general method that can be further used to estimate counterfactual means under any static or dynamic regime. It is a powerful statistical tool that facilitates the estimation of complex exposures over time in the presence of time-varying confounding and even in the presence of interaction between exposures and covariates that vary over time (Robins 1986, 2004; Robins et al. 2008). In the presence of such covariates, standard approaches for adjustment for confounding are biased (more details can be found in Chapter 5, Section 3). Robins proposes finding the parameters ψ of the optimal blip-to-zero function via G-estimation. He defined H_j as the outcome of a person adjusted by the expected difference between the average outcome if he received t_j instead of the optimal at time j, with given treatment and covariate history to time $j-1$

and who was subsequently treated optimally after time j.

$$H_j(\psi) = Y + \sum_{i=j}^{K}[\gamma_i^{zero}(d_i^{opt}|\bar{M}_i, \bar{T}_{i-1};\psi) - \gamma_i^{zero}(t_i|\bar{M}_i, \bar{T}_{i-1};\psi)].$$

Returning to blip function definition:

$$\begin{aligned}
H_j(\psi) &= Y + \sum_{i=j}^{K}\{E(Y(\underline{d}_i^{opt})|\bar{M}_i, \bar{T}_{i-1}) - E(Y(d_i^{zero}, \underline{d}_{i+1}^{opt})|\bar{M}_i, \bar{T}_{i-1}) \\
&- [E(Y(t_i, \underline{d}_{i+1}^{opt})|\bar{M}_i, \bar{T}_{i-1}) - E(Y(d_i^{zero}, \underline{d}_{i+1}^{opt})|\bar{M}_i, \bar{T}_{i-1})]\} \\
&= Y + \sum_{i=j}^{K} E(Y(\underline{d}_i^{opt})|\bar{M}_i, \bar{T}_{i-1}) - E(Y(t_i, \underline{d}_{i+1}^{opt})|\bar{M}_i, \bar{T}_{i-1}).
\end{aligned}$$

For example, if $K = 3$ then:

$$\begin{aligned}
H_3(\psi) &= Y + E(Y(d_3^{opt})|\bar{M}_3, \bar{T}_2) - E(Y(t_3)|\bar{M}_3, \bar{T}_2) \\
H_2(\psi) &= Y + E(Y(\underline{d}_2^{opt})|\bar{M}_2, T_1) - E(Y(t_2, d_3^{opt})|\bar{M}_2, T_1) \\
&+ E(Y(d_3^{opt})|\bar{M}_3, \bar{T}_2) - E(Y(t_3)|\bar{M}_3, \bar{T}_2) \\
H_1(\psi) &= Y + E(Y(\underline{d}_1^{opt})|M_1) - E(Y(t_1, \underline{d}_2^{opt})|M_1) \\
&+ E(Y(\underline{d}_2^{opt})|\bar{M}_2, T_1) - E(Y(t_2, d_3^{opt})|\bar{M}_2, T_1) \\
&+ E(Y(d_3^{opt})|\bar{M}_3, \bar{T}_2) - E(Y(t_3)|\bar{M}_3, \bar{T}_2).
\end{aligned}$$

We can obtain that

$$H_j(\psi) = Y + \sum_{i=j}^{K}\mu_i(t_i|\bar{M}_i, \bar{T}_{i-1};\psi)$$

where $H_j(\psi)$ is the counterfactual outcome, i.e., $H_j(\psi) = Y[(\underline{d}_j^{opt})|\bar{M}_j, \bar{T}_{j-1};\psi]$ (Robins, 2004, p. 204).

Let $S_j = f_j(\bar{m}_j, \bar{t}_j;\psi)$, be a vector function determined by interactions between the variables and treatment which may affect the outcome. It depends on treatment and history. Further, let $p_j(t_j|\bar{m}_j, \bar{t}_{j-1};\alpha)$, be the probability of receiving treatment t_j or the probability density function if t is continuous. Suppose the blip is linear, as an example,

$\gamma_j = t_j(\psi_{j0} + \psi_{j1} m_j)$. In this case we might take $S_j = \frac{\partial}{\partial \psi}(\gamma_j) = t_j(1, m_j)^T$. Specifying the functions $H_j(\psi)$ and S_j function for the estimation to influence outcome, then we define

$$L(\psi, S) = \sum_{j=1}^{K} H_j(\psi)\{S_j(\bar{M}_j, \bar{T}_j) - E[S_j(\bar{M}_j, \bar{T}_j)]\} \tag{3.2}$$

We will estimate ψ as the value for which $L(\hat{\psi}, S) = 0$.

3.3.2 Unbiasedness

To be sure that $E[L(\psi_{true}, s)] = 0$, let us assume Assumption 1 and Assumption 2 of consistency and no unmeasured confounders that were explained in Section 2.4.5 and assume that M_1 is a random variable considered to be a baseline covariate at the start of the first interval, M_j, where $j = 2, \cdots, K$, are random variables denoting the status at the beginning of the j^{th} interval and Y is a random variable denoting the final response at the end of the K^{th} interval

$$Y = \Phi - \sum_{j=1}^{K} \mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}),$$

where Φ might be depend on any subset of $\bar{M}_K$ but does not depend on any of $\bar{T}_j$, that

$$H_j(\psi_{true}) = Y + \sum_{i=j}^{K} \mu_i(t_i|\bar{M}_i, \bar{T}_{i-1}, \psi_{true}),$$

then

$$\begin{aligned} H_1(\psi_{true}) &= \Phi, \\ H_2(\psi_{true}) &= \Phi - \mu_1(t_1|M_1, \psi_{true}), \end{aligned}$$

and

$$H_j(\psi_{true}) = \Phi - \sum_{i=1}^{j-1} \mu_i(t_i|\bar{M}_i, \bar{T}_{i-1}, \psi_{true}).$$

By the assumption of no unmeasured confounders, that new treatments are assigned independently of potential future responses to treatment, conditional on the history of treatments and response to date. H_j is dependent on $\bar{T}_{j-1}$, but H_j is independent of $\underline{T}_j$. Here the conditional independence justified by sequential ignorability. We know that $E[xy] = E[x]E[y] + [x, y]$, so

$$\begin{aligned} E[L(\psi_{true})] &= \sum_{j}^{K} E[H_j(\psi_{true})\{S_j(\bar{M}_j, \bar{T}_j) - E[S_j(\bar{M}_j, \bar{T}_j)]\}], \\ &= \sum_{j}^{K} E[H_j(\psi_{true})]E\{S_j(\bar{M}_j, \bar{T}_j) - E[S_j(\bar{M}_j, \bar{T}_j)]\} \\ &+ \mathbf{cov}[H_j(\psi_{true}), \{S_j(\bar{M}_j, \bar{T}_j) - E[S_j(\bar{M}_j, \bar{T}_j)]\}]. \end{aligned}$$

Note $E\{S_j(\bar{M}_j, \bar{T}_j) - E[S_j(\bar{M}_j, \bar{T}_j)]\} = 0$. The $\{H_j(\psi_{true}), [S_j - E(S_j|\bar{M}_j, \bar{T}_j)]\}$ is equal to zero. Thus we can conclude that S_j depends on T_j given its earlier history, but H_j is independent of T_j given its earlier history and so the covariance is zero. Now we have that $E[L(\psi_{true})] = 0$. Thus it is an unbiased estimating equation from which consistent provided the treatment allocation probabilities, $p_j(t_j|\bar{T}_{j-1}, \bar{M}_j)$, are known or correctly modelled, so that the values of $E[S_j|\bar{M}_j, \bar{T}_j]$ are correctly specified.

Robins (2004) refined the following equation to gain efficiency,

$$L(\psi, S) = \sum_{j=1}^{k}[H_j(\psi) - E[H_j(\psi)|\bar{M}_j, \bar{T}_{j-1}]]\{S_j(\bar{M}_j, \bar{T}_j) - E[S_j(\bar{M}_j, \bar{T}_j)]\} \tag{3.3}$$

Using $E[H_j(\psi)|\bar{M}_j, \bar{T}_{j-1}])$ gives more efficient estimators than those found using Equation 3.2. Robins proves that estimates found by Equation 3.3 are consistent provided either $E[H_j|\bar{M}_j, \bar{T}_j]$ or $E[S_j|\bar{M}_j, \bar{T}_j]$ is correctly modelled, and thus is said to be doubly-robust.

3.3.3 Modification

When optimal blips are linear in ψ, we can solve for $\hat{\psi}$ explicitly. Use the modification

$$H_{mod,j}(\psi) = Y - \gamma_j(t_j|\bar{M}_j, \bar{T}_{j-1}; \psi) + \sum_{i=j}^{K}[\gamma_i(d_i^{opt}|\bar{M}_i, \bar{T}_{i-1}; \psi) - \gamma_i(t_i|\bar{M}_i, \bar{T}_{i-1}; \psi)].$$

which is a person's outcome adjusted by the expected difference between the average outcome for someone who received t_j and someone who was given the zero regime at time j, who both had the same treatment and covariate history to time j and were treated optimally from time $j+1$.

3.3.4 Iterative minimization for optimal regimes (IMOR)

Murphy (2003) introduced a method to estimate optimal decision rules to produce the maximum final mean response. To estimate the regret function parameters, we can follow estimation procedures which are based on the least squares characterization. Let us suppose that $\mu_j : j = 1, 2, \dots, K$, is the regret function, and that $p_j(t_j|\bar{M}_j, \bar{T}_{j-1})$, the conditional probability of a treatment t_j for given history, is known. Murphy shows that each μ_j in a vector of $\bar{\mu_K}$ satisfies both the constraints,inf $\mu_j(t|\bar{M}_j, \bar{T}_{j-1}) = 0$ and,

$$E[Y + \sum_{i=1}^{K} \mu_i(t_i|\bar{M}_i, \bar{T}_{i-1}) - \sum_t \mu_j(t|\bar{M}_j, \bar{T}_{j-1}) p_j(t|\bar{M}_j, \bar{T}_{j-1})]^2 \leq$$

$$E[Y + \sum_{i=1, i\neq j}^{K} \mu_i(t_i|\bar{M}_i, \bar{T}_{i-1}) + \mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) - \sum_t \mu_j(t|\bar{M}_j, \bar{T}_{j-1}) p_j(t|\bar{M}_j, \bar{T}_{j-1}]^2,$$

for all $\mu_j : j = 1, 2, \dots, K$. Murphy suggests that we replace Y by $Y + c$ where c is an unknown scalar in order to improve the stability of the minimization. Murphy (2003) developed a method that estimates the parameters of the optimal regime, ψ, by searching for $(\hat{\psi}; \hat{c})$ which satisfy

$$\sum_{j=1}^{K} P_n[Y + \hat{c} + \sum_{i=1}^{K} \mu_i(t_i|\bar{M}_i, \bar{T}_{i-1}, \hat{\psi}) - \sum_t \mu_j(t|\bar{M}_j, \bar{T}_{j-1}, \hat{\psi}) p_j(t|\bar{M}_j, \bar{T}_{j-1}; \hat{\alpha})]^2 \leq$$

$$\sum_{j=1}^{K} P_n[Y + c + \sum_{i=1, i\neq j}^{K} \mu_i(t_i|\bar{M}_i, \bar{T}_{i-1}, \hat{\psi}) + \mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}, \psi) - \sum_t \mu_j(t|\bar{M}_j, \bar{T}_{j-1}, \psi) p_j(t|\bar{M}_j, \bar{T}_{j-1}, \hat{\alpha}),]^2$$

for all c and all ψ, where $P_n f = \frac{1}{n}\sum_{i=1}^{n} f(.)$ is the empirical average function. Treatment probabilities, e.g., $p_j(t|\bar{M}_j, \bar{T}_{j-1}, \hat{\alpha})$ can be estimated. Murphy describes an iterative method of finding solutions to the previous estimation equation, which begins by selecting an initial value of $\hat{\psi}$, say $\hat{\psi}^{old}$, then minimizing the right-hand side (RHS) of the equation over $(\hat{\psi}; \hat{c})$ to obtain a new value of ψ, $\hat{\psi}^{new}$, and repeating this until convergence.

3.3.5 Simulation examples

Simulation 1

Moodie, Richardson and Stephens (2007) use a simple two time point example with Normal states and binary actions. We replicate their example that data were generated as $M_1 \sim N(450, 150^2)$, $T_1 \sim Bern(0.5)$, $M_2 \sim N(1.25M_1, 60^2)$ and $T_2 \sim Bern(0.5)$. Blip functions were parameterised, leading to regrets

$$\mu_1(t_1|M_1;\psi) = \begin{cases} I(t_1 = 0)(\psi_{10} + \psi_{11}M_1) & \psi_{10} + \psi_{11}M_1 > 0 \\ -I(t_1 = 1)(\psi_{10} + \psi_{11}M_1) & \psi_{10} + \psi_{11}M_1 < 0 \end{cases}$$

$$\mu_2(t_2|\bar{M}_2, T_1;\psi) = \begin{cases} I(t_2 = 0)(\psi_{20} + \psi_{21}M_1) & \psi_{20} + \psi_{21}M_2 > 0 \\ -I(t_2 = 1)(\psi_{20} + \psi_{21}M_2) & \psi_{20} + \psi_{21}M_2 < 0 \end{cases}$$

and then response $Y \sim N(400 + 1.6M_1 - \mu_1(T_1|M_1;\psi) - \mu_2(T_2|\bar{M}_2, T_1;\psi), 60^2)$.

3.3.6 Results

The goal of this simulation is to compare the performance of Robins and Murphy methods discussed in this chapter, as well to illustrate the double-robustness of G-estimation Equation 3.3.

All results are presented in Table 3.4 for $K = 2$ time points. Using each of G-estimation and IMOR, we estimate the parameter ψ for 1000 data-sets with sample sizes of 500, 1000 patients respectively. In Table 3.4 IMOR results give unbiased estimators with more efficiency than G-estimation using Equation 3.2. Robins (2004) uses G-estimation Equation 3.3, to give estimates which are said to be doubly-robust. In implementing G-estimation

ψ	$\hat{\psi}$	SE	rMSE	Cov.*	$\hat{\psi}$	SE	rMSE	Cov.*
		$n = 500$				$n = 1000$		
g-est.	eqn.(3.2)							
$\psi_{10} = 250$	249.364	309.119	308.965	94.8	240.276	223.771	223.870	95.1
$\psi_{11} = -1$	-1.021	0.736	0.736	94.9	-0.979	0.538	0.538	94.6
$\psi_{20} = 720$	734.252	257.664	257.929	94.8	734.187	182.279	182.740	95.4
$\psi_{21} = -2$	-2.040	0.492	0.493	95.8	-2.033	0.346	0.348	95.3
g-est.**	eqn.(3.3)	incorrect	model					
$\psi_{10} = 250$	247.224	17.312	17.524	93.7	248.437	12.822	12.910	95.2
$\psi_{11} = -1$	-0.993	0.037	0.038	94.1	-0.995	0.027	0.028	95.1
$\psi_{20} = 720$	719.384	36.678	36.665	95.1	719.600	27.184	27.173	94.9
$\psi_{21} = -2$	-1.999	0.077	0.077	94.7	-1.999	0.057	0.057	94.7
g-est.***	eqn.(3.3)	correct	model					
$\psi_{10} = 250$	249.529	17.828	17.825	95.0	249.802	12.676	12.671	95.5
$\psi_{11} = -1$	-1.000	0.037	0.037	94.9	-0.999	0.027	0.027	95.0
$\psi_{20} = 720$	719.529	17.682	17.679	94.6	719.723	12.407	12.404	94.1
$\psi_{21} = -2$	-1.999	0.029	0.029	95.2	-2.000	0.021	0.021	94.1
IMOR								
$\psi_{10} = 250$	251.292	191.348	191.257	95.2	253.198	133.370	133.342	95.1
$\psi_{11} = -1$	-0.999	0.432	0.432	95.3	-1.006	0.302	0.302	95.4
$\psi_{20} = 720$	717.450	193.615	193.535	95.2	718.673	131.550	131.490	94.7
$\psi_{21} = -2$	-1.997	0.299	0.299	94.9	-1.997	0.203	0.203	94.8

$*$ *Coverage of* 95% *Wald-type confidence intervals*

$**$ *$E[H_{mod,j}(\psi)|\bar{m}_j, \bar{t}_j]$ linear in $\bar{m}_j, \bar{t}_j$ (incorrect model)*

$***$ *$E[H_{mod,j}(\psi)|\bar{m}_j, \bar{t}_j]$ piece-wise linear (correct model)*

Table 3.4: Estimation of ψ parameters using G-estimation and IMOR for 1000 data-sets

Equation 3.3, two models were considered for $E[H_{mod,j}(\psi)|\bar{M}_j, \bar{T}_{j-1}]$. The first assumed that $E[H_{mod,j}(\psi)|\bar{M}_j, \bar{T}_{j-1}]$ depends linearly on all of the earlier history (the incorrect model). The second is the correct model which allowed the mean function to be piecewise, discontinuous linear with the optimal rule thresholds.

T_1	M_1	M_2	$E[H_2\|\bar{M}_2, \bar{T}_1]$
0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$400 + 1.6M_1$
1	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$400 + 1.6M_1$
0	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$400 + 1.6M_1 - (\psi_{10} + \psi_{11}M_1)$
1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$400 + 1.6M_1 + (\psi_{10} + \psi_{11}M_1)$
0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{10}}{\psi_{11}}$	$400 + 1.6M_1 - (\psi_{20} + \psi_{21}M_2)$
1	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{10}}{\psi_{11}}$	$400 + 1.6M_1 - (\psi_{20} + \psi_{21}M_2)$
0	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{10}}{\psi_{11}}$	$400 + 1.6M_1 - (\psi_{10} + \psi_{11}M_1) - (\psi_{20} + \psi_{21}M_2)$
1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{10}}{\psi_{11}}$	$400 + 1.6M_1 + (\psi_{10} + \psi_{11}M_1) - (\psi_{20} + \psi_{21}M_2)$

Table 3.5: $E[H_2|\bar{M}_2, \bar{T}_1]$ at the optimal solution using G-estimation Equation 2.2, to make sure that $E(H_2 - E[H_2|\bar{M}_2, \bar{T}_1]) = 0$

The efficiency gained instead of using Equation 3.2 is considerable. However, both of the two models in G-estimation Equation 3.3 give more efficiency comparing with IMOR of Murphy (2003). But using the correct model for $E[H_{mod,2}(\psi)|\bar{M}_2, \bar{T}_1$ also leads to increased efficiency. Table 3.5 explains that in this model the $E[H_{mod,1}(\psi)|\bar{m}_1, \bar{t}_1]$ should equal to $400 + 1.6M_1$if $M_1 \geq -\frac{\psi_{10}}{\psi_{11}}$ and $400 + 1.6M_1 - (\psi_{10} + \psi_{11}M_1)$ otherwise. This is so because the patient should not be treated when $M_1 \geq -\frac{\psi_{10}}{\psi_{11}}$. At the second interval, we use the optimal rule to calculate $E[H_{mod,2}(\psi)|\bar{M}_2, \bar{T}_1]$. Table 3.4 explains that we have 2^3 different possibilities. So if $T_1 = 1$, and $M_1 \geq 250, M_2 < 360$, then $E[H_{mod,2}(\psi)|\bar{M}_2, \bar{T}_1] = 400 +$

$$1.6M_1 - (\psi_{20} + \psi_{21}M_2) + (\psi_{10} + \psi_{11}M_1)I(\psi_{10} + \psi_{11}M_1 < 0) + (\psi_{20} + \psi_{21}M_2)I(\psi_{20} + \psi_{21}M_2 < 0),$$

$$E[H_{mod,2}(\psi)|\bar{M}_2, \bar{T}_1] = 400 + 1.6M_1 + (\psi_{10} + \psi_{11}M_1) - (\psi_{20} + \psi_{21}M_2).$$

That means the optimal decision at $j = 1$ is that the patient should not take treatment when $M_1 \geq -\psi_{10}/\psi_{11}$, and should take treatment that at $j = 2$, because $M_2 < -\psi_{20}/\psi_{21}$.

Simulation 2

In this simulation our goal is to extend Moodie, Richardson and Stephens (2007) by using a three time point example. States were generated as $M_1 \sim N(450, 150^2)$, $M_2 \sim N(1.25M_1, 60^2)$ and $M_2 \sim N(1.6M_1, 60^2)$. Actions were generated as $T_j \sim Bern(0.5)$. The final response was

$$Y \sim N(400 + 2M_1 - \mu_1(T_1|M_1; \psi) - \mu_2(M_2|\bar{M}_2, T_1; \psi) - \mu_3(T_3|\bar{M}_3, \bar{T}_2; \psi), 60^2),$$

where the regret functions were

$$\mu_1(t_1|M_1; \psi) = \begin{cases} I(t_1 = 0)(\psi_{10} + \psi_{11}M_1) & \psi_{10} + \psi_{11}M_1 > 0 \\ -I(t_1 = 1)(\psi_{10} + \psi_{11}M_1) & \psi_{10} + \psi_{11}M_1 < 0 \end{cases}$$

$$\mu_2(t_2|\bar{M}_2, T_1; \psi) = \begin{cases} I(t_2 = 0)(\psi_{20} + \psi_{21}M_2) & \psi_{20} + \psi_{21}M_2 > 0 \\ -I(t_2 = 1)(\psi_{20} + \psi_{21}M_2) & \psi_{20} + \psi_{21}M_2 < 0 \end{cases}$$

$$\mu_3(t_3|\bar{M}_3, \bar{T}_2; \psi) = \begin{cases} I(t_3 = 0)(\psi_{30} + \psi_{31}M_3) & \psi_{30} + \psi_{31}M_3 > 0 \\ -I(t_3 = 1)(\psi_{30} + \psi_{31}M_3) & \psi_{30} + \psi_{31}M_3 < 0 \end{cases}$$

Results are given in Table 3.6. They are similar with results when using two time points. We estimate the parameter ψ for 1000 data-sets with sample sizes of 500, 1000 patients respectively. IMOR gives unbiased estimators with more efficiency than G-estimation using Equation 3.2, but less efficiency comparing with both of the two models in G-estimation

Equation 3.3. G-estimation Equation 3.3, gives estimates which are similar to those when $K = 2$. The efficiency gained by using the G-estimating Equation 3.3 instead of Equation 3.2 is considerable whether we use the incorrect model for $E[H_{mod,3}(\psi)|\bar{M}_3, \bar{T}_2]$ which depends linearly on all of earlier history or use the correct model for $E[H_{mod,3}(\psi)|\bar{M}_3, \bar{T}_2$ which allowed the mean function to be piece-wise discontinuous linear with the optimal thresholds (see a table in Appendix 9.2 which explains that, where we have 2^5 different possibilities of $E(H_{mod,3}|.))$.

3.4 Dynamic Programming for Optimal Regime (DPOR)

As shown in the previous chapter, dynamic programming is able to solve complex dynamic decision problems. In this section we aim to use a function of dynamic programming as an alternative approach for estimating the dynamic treatment regime. We try to show the similarities that are shared between DP and the methods of G-estimation and IMOR and what might seem different.

3.4.1 The dynamic programming model

Let an optimal dynamic programming regime at time point $j = 1, \cdots, K$ be denoted by ${d_j}^{opt}$. T_j represents a treatment (decision) at the j^{th} time point, we suppose T_j is a binary (0,1) decision. The aim is to maximize the final response at the end of the last time point. At the K^{th} decision point

$$\begin{aligned} \eta_K(\bar{M}_K, \bar{T}_K) &= E[Y|\bar{M}_K, \bar{T}_K], \\ {d_K}^{opt}(t_K|\bar{M}_K, \bar{T}_{K-1}) &= \arg \max_{t_K} \ \eta_K(\bar{M}_K, \bar{T}_K). \end{aligned}$$

ψ	$\hat{\psi}$	SE	rMSE	Cov.*	$\hat{\psi}$	SE	rMSE	Cov.*
		$n = 500$				$n = 1000$		
g-est.	eqn.(3.1)							
$\psi_{10} = 250$	256.598	384.230	384.095	95.1	249.124	252.773	252.648	94.7
$\psi_{11} = -1$	-1.020	0.923	0.923	94.1	-0.997	0.615	0.614	94.5
$\psi_{20} = 720$	733.568	317.892	318.023	94.9	733.813	223.947	224.260	95.2
$\psi_{21} = -2$	-2.036	0.600	0.600	94.8	-2.040	0.418	0.420	93.9
$\psi_{30} = 1310$	1338.391	270.646	271.997	94.4	1313.054	182.217	182.152	95.2
$\psi_{31} = -3$	-3.047	0.382	0.385	94.9	-3.006	0.257	0.257	95.0
g-est.**	eqn.(3.2)	incorrect	model					
$\psi_{10} = 250$	247.850	86.108	86.091	95.4	247.011	61.390	61.432	94.8
$\psi_{11} = -1$	-0.991	0.163	0.163	95.1	-0.991	0.115	0.116	94.4
$\psi_{20} = 720$	715.439	38.578	38.828	95.3	715.213	27.901	28.295	94.7
$\psi_{21} = -2$	-1.990	0.080	0.081	94.8	-1.990	0.058	0.059	94.7
$\psi_{30} = 1310$	1312.104	92.291	92.269	95.2	1308.968	65.821	65.796	94.6
$\psi_{31} = -3$	-3.002	0.152	0.152	95.8	-2.998	0.108	0.108	94.7
g-est.***	eqn.(3.2)	correct	model					
$\psi_{10} = 250$	250.879	16.630	16.645	94.6	249.896	12.117	12.111	95.1
$\psi_{11} = -1$	-1.002	0.035	0.035	94.4	-1.000	0.026	0.025	94.2
$\psi_{20} = 720$	719.836	16.442	16.434	95.1	720.495	11.545	11.550	94.7
$\psi_{21} = -2$	-2.000	0.028	0.027	94.7	-2.001	0.019	0.019	95.4
$\psi_{30} = 1310$	1309.638	17.212	12.207	94.8	1309.985	11.862	11.857	95.5
$\psi_{31} = -3$	-2.999	0.023	0.023	93.5	-3.000	0.016	0.016	95.5
IMOR								
$\psi_{10} = 250$	248.671	254.063	253.940	95.2	244.147	183.709	183.711	95.0
$\psi_{11} = -1$	-1.000	0.624	0.624	94.8	-0.986	0.446	0.446	94.6
$\psi_{20} = 720$	732.780	213.515	213.790	94.9	712.411	145.139	145.265	94.4
$\psi_{21} = -2$	-2.020	0.347	0.347	95.3	-1.986	0.240	0.240	95.1
$\psi_{30} = 1310$	1310.822	262.622	262.492	94.9	1309.778	192.021	191.925	94.8
$\psi_{31} = -3$	-3.000	0.340	0.340	95.1	-2.998	0.246	0.246	94.8

$*$ *Coverage of* 95% *Wald-type confidence intervals*

$**$ *$E[H_{mod,j}(\psi)|\bar{m}_j, \bar{t}_j]$ linear in $\bar{m}_j, \bar{t}_j$ (incorrect model)*

$***$ *$E[H_{mod,j}(\psi)|\bar{m}_j, \bar{t}_j]$ piece-wise linear (correct model)*

Table 3.6: Estimation of ψ parameters using G-estimation and IMOR for 3 time points and 1000 data-sets of sample sizes 500 and 1000

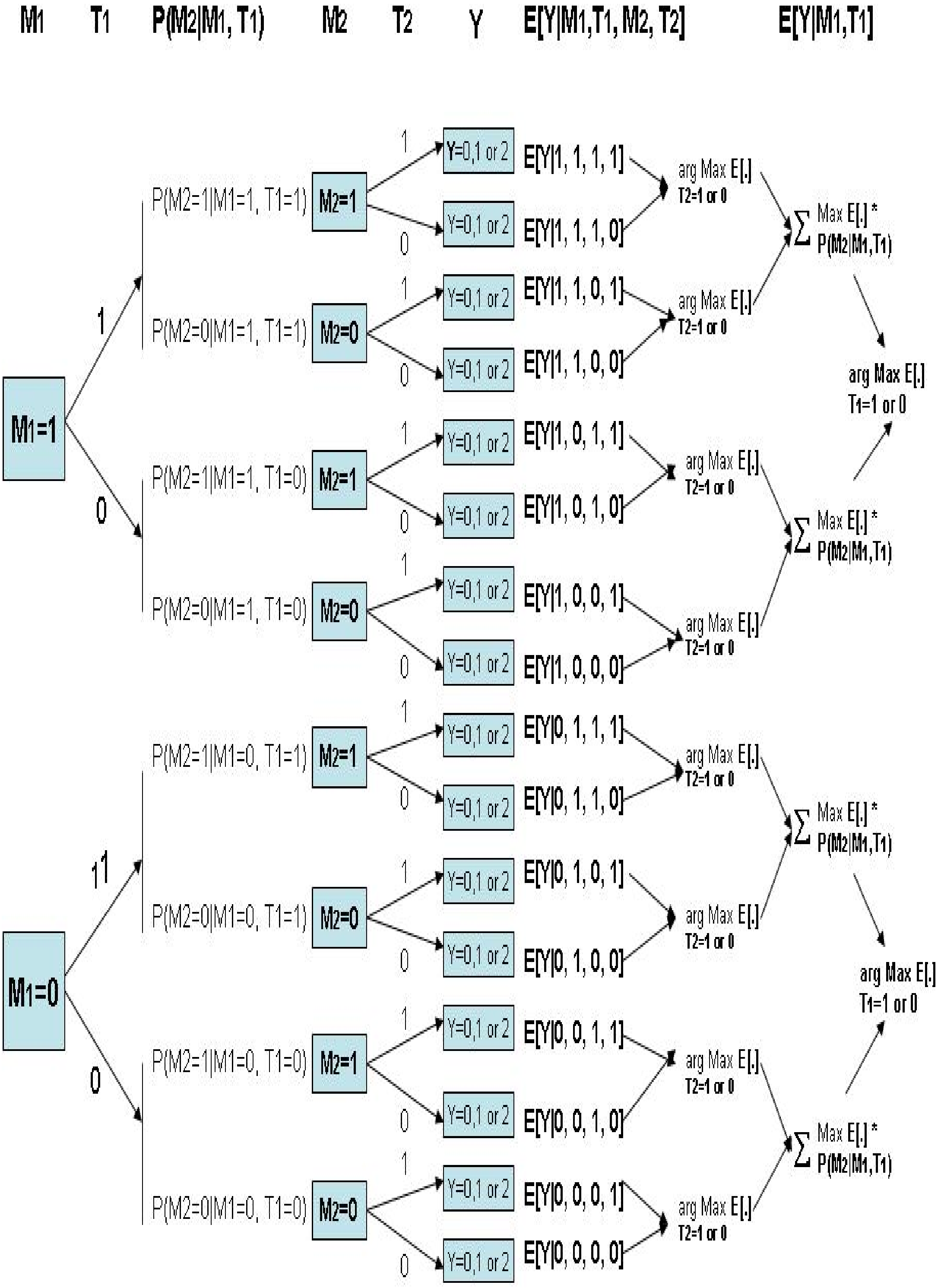

Figure 3.3: The optimal decision using dynamic programming.

Then at the j^{th} decision point

$$\begin{aligned}
\eta_j(\bar{M}_j, \bar{T}_j) &= E[Y(T_j)|\bar{M}_j, \bar{T}_{j-1}], \\
&= \sum_{\bar{M}_{j+1}} \max E[Y(T_j)|\bar{M}_j, \bar{T}_{j-1}] \times Pr(M_{j+1}|\bar{M}_j, \bar{T}_j), \\
d_j{}^{opt}(t_j|\bar{M}_j, \bar{T}_{j-1}) &= \arg \max_{t_j} \ \eta_j(\bar{M}_j, \bar{T}_{j-1}).
\end{aligned}$$

3.4.2 Simulation results using dynamic programming and other methods

We will use dynamic programming methodology to find the optimal treatment regime, for the two time point example, and compare the results with other methods via simulations. Figure 3.3 shows an example for how to use DP policy to decide the optimal decisions. At the second time point

$$\begin{aligned}
\eta_2(\bar{M}_2, \bar{T}_2) &= E[Y(T_2)|\bar{M}_2, \bar{T}_1], \\
d_2{}^{opt}(t_2|\bar{M}_2, T_1) &= \arg \max_{t_2} \ \eta_2(\bar{M}_2, \bar{T}_1).
\end{aligned}$$

Thus through the distribution of $E[Y|\bar{M}_2, \bar{T}_2]$ we can decide which roles can be followed for choosing optimal strategies, e.g., by comparing which T_2 leads to $\max_{t_2} \eta_2(\bar{M}_2, T_1)$. Then at the first decision point

$$\begin{aligned}
\eta_1(M_1, T_1) &= E[Y(T_1)|M_1], \\
&= \sum_{\bar{M}_2} \max E[Y(T_1)|\bar{M}_2] \times E(\bar{M}_2|M_1, T_1), \\
d_1{}^{opt}(t_1|M_1) &= \arg \max_{t_1} \ \eta_1(M_1, T_1).
\end{aligned}$$

Generally, the roles for finding optimal dynamic strategies for all time points $j < K$, are based on all distributions of $E[M_{j+1}|\bar{M}_j, \bar{T}_j]$ and $E[Y(T_j)|\bar{M}_j, \bar{T}_{j-1}]$. Table 3.7 shows the results of generating the data using the following different policies: random decisions, decisions using true values of ψ, decisions using estimated values of ψ by dynamic programming, decisions using estimated values of ψ by G-estimation and decisions using estimated values of ψ by IMOR.

	Dynamic programming	G-estimation	IMOR	Random	True ψ
$n = 100$					
Y	1119.8251	1119.9568	1119.8424	780.2857	1120.1642
SE	30.38086	24.55009	24.64223	33.44929	24.75274
$n = 500$					
Y	1120.1385	1120.0371	1120.1484	780.7488	1120.2901
SE	11.00288	11.04423	10.99385	14.94091	11.02842
$n = 1000$					
Y	1120.0847	1120.1040	1120.1027	780.4983	1120.0963
SE	7.842017	7.857395	7.783616	10.701643	7.858123

Table 3.7: Mean optimal response using dynamic programming, G-estimation, IMOR, Random, True ψ for 2 time points and simulations of 100 data-sets of sample sizes 100, 500 and 1000

The random policy gives small estimated mean values with high standard errors since we did not use any rules for choosing actions. Dynamic programming gives close estimates which are similar to those when using the true values of ψ. But standard errors seem high in cases of modest samples, e.g., $n = 100$ in this example. On the other hand, both G-estimation and IMOR give better estimators even with this small sample size. Thus we can conclude that dynamic programming method gives the optimal actions, but it seems there are many difficulties for implementation such as small sample size, especially when time points are increased, discrete states and responses are needed to calculate earlier states probabilities. However, a large number of expectations are needed to calculate the final response. In total, for this simple example, we need to calculate $2 \times (8 + 2) = 20$ expectations in working out the optimal policy. If the example had $K = 3$ decision stages, still with binary decisions and binary states, we would have needed $2 \times (32 + 8 + 2) = 42$ expectations. For general K we need $2 \times (2^{2K-1} + 2^{2K-3} + \ldots + 2)$ expectations.

3.5 Discussion

This chapter has explained the connections between Robins and Murphy methods that where introduced to make inference about optimal dynamic treatment regimes. The methods are very similar, but not equivalent. The efficiency of IMOR and G-estimation has been shown: G-estimates from equation 3.2 are efficient when the models for the mean and the variance of the counterfactual are correct.

The choice of which method to use in practice is open. G-estimation may be easier to implement particularly if the optimal blip is linear. In order to believe in Assumption 1 'no unmeasured confounders' in an observational study, typically many covariates need to be included which makes models prone to misspecification and dynamic programming impossible, hence double-robustness seems like a strong point. But still a risk if both of $E[H_j|\bar{M}_j, \bar{T}_j]$ or $E[S_j|\bar{M}_j, \bar{T}_j]$ are not correctly modelled. Then the model can lead not only to incorrect estimates of the optimal decision, but to incorrect conclusions as to whether treatment is beneficial. We conclude that, it is important to be aware of the models for observables which are then used in implementation.

Chapter 4

Regret-Regression for Optimal Dynamic Treatment Regimes

4.1 Introduction

As seen there is increasing interest in methods for determining optimal dynamic treatment rules from observational data. Several authors have investigated conditions under which valid causal inference can be obtained while others have concentrated more on estimation for a variety of problems (e.g. Hogan and Lee 2004, Johnson 2008, Lok et al 2004, Moodie et al 2007, Murphy 2003, Petersen et al 2007, Robins 2004).

In samples of modest size there is no realistic alternative to parametric modelling of at least some components of the terms needed to determine an optimal regime. In turn this brings the risk that the chosen model is not suitable for the data. Fundamental statistical practice of model building, checking and comparison has had little attention so far in this literature.

The general problem was described in Chapter 3. Two broad classes of methods are available, which we can think of as *direct* and *indirect*. The direct class can be based on modelling of $E[Y|\bar{M}_j, \bar{T}_{j1}]$ or $E[Y|\bar{T}_j]$. The problem is then how to tease out the causal effects of actions, which may require some form of dynamic programming as well as additional modelling of M_{j+1} given $(\bar{M}_j, \bar{T}_j)$. The computational burden of such an approach scales dramatically with K and soon becomes infeasible. Structural nested mean models (eg Robins, 1994) also fall within the direct class and have an advantage in interpretability. Computational issues remain formidable however. The indirect approach by contrast does not attempt to model the response Y. Instead, causal effects expressed as differences between counterfactuals - outcomes that might have occurred - are parameterised. Examples of these are the regrets of Murphy (2003) and the blips of Robins (2004). Interpretation of estimates is then easier but now model adequacy is less straightforward, since there is no model for the observed response. Computational problems are reduced but not removed. We propose a modelling and estimation strategy which incorporates the regret functions of Murphy (2003) into a regression model for observed responses. Implementation is therefore the focus of this chapter, in which we present and apply a method which:

- is straightforward to implement;
- provides direct estimates of causal parameters;
- allows diagnostic model assessment and model comparisons.

The aim is to combine traditional regression modelling of responses with the described methods for modelling causal effects. Our target is problems of modest sample size in which parametric models are likely to be the only feasible way forward, in which case it is especially important that some form of model assessment procedure be available. The method combines the strengths of the two approaches and is moreover computationally undemanding. The idea is straightforward: we simply include parameterised regret functions in a regression model for Y. Provided the remaining terms in the model are correctly specified, and the usual causal inference assumptions hold, we have the advantage of modelling an observable quantity and hence diagnostic capability, yet from the fitted model we can read off the optimal decisions and also the consequences of suboptimal ones. And the computational challenge is simply to fit the chosen model to the responses, which may require only standard methods and software. Estimation is quick and diagnostics are available, meaning a variety of candidate models can be compared. The method is illustrated using two simulation scenarios of Murphy (2003) and Moodie et al (2007). The approach is described in the next section then the simulations are used in Section 4.3 to investigate the proposal.

4.1.1 Simple example

Let us introduce at this point an extremely simple example which will be used to fix ideas. All subjects are assumed to start in the same state M_1, which can thus be ignored, and there is then a sequence $T_1M_2T_2$ of binary variables followed by some response Y.

Row	T_1	M_2	T_2	N	$E[Y\|T_1, M_2, T_2]$
1	0	0	0	50	3
2	0	0	1	50	8
3	0	1	0	120	2
4	0	1	1	280	7
5	1	0	0	280	6
6	1	0	1	70	5
7	1	1	0	135	4
8	1	1	1	15	1

Table 4.1: The data

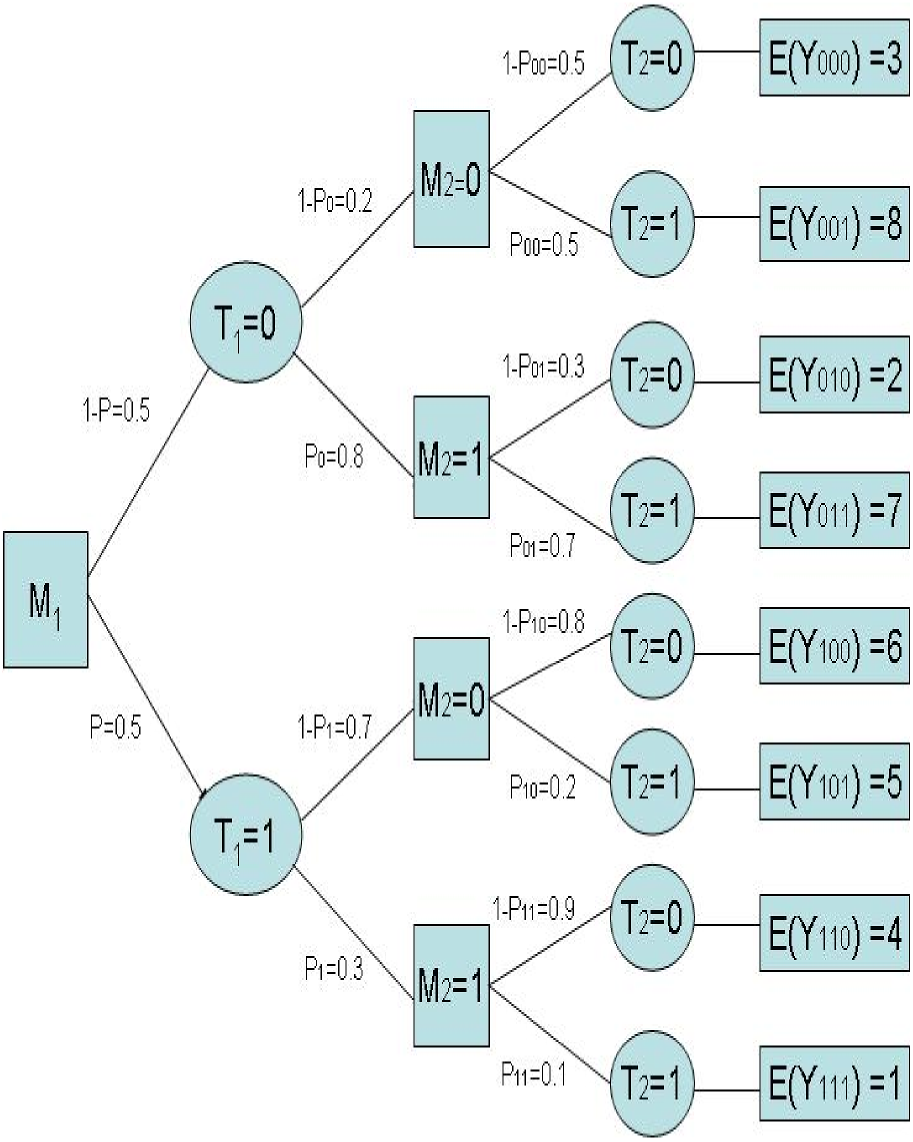

Figure 4.1: The number of subjects and the average value of the outcome $E[Y|T_1, M_2, T_2]$

Table 4.1 and Figure 4.1 show the set-up and the parameter choices we use for illustration. Recall Equation (3.1) that $\underline{d}_j^{opt}$ is understood to mean that the optimal policy regime is followed from time j, and that regret at time j is defined as

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = E(Y \mid \bar{M}_j, \bar{T}_{j-1}, \underline{d}_j^{opt}) - E(Y \mid \bar{M}_j, \bar{T}_{j-1}, t_j, \underline{d}_{j+1}^{opt}).$$

Optimal actions can be taken by working from the final time point and choosing the actions when regrets are equal to zero. To choose optimal actions at the final time point, regrets are directly calculated. At other time point $j = 1, \cdots, k-1$, we might choose optimal actions by calculation of regrets through the expectation of optimal final rewards given the history of previous states and actions. Regrets for making the wrong decision at the second time-point can be read directly from the figure, as 5,5,1 and 3 for (T_1, M_2) equal to (0,0),(0,1),(1,0) and (1,1) respectively. Choice $T_1 = 0$ is optimal for the first time-point and the regret for choosing $T_1 = 1$ can be worked out to be 1.8, which is estimated consistently using Equation (3.2).

We turn now to the direct approach, which involves two stages. The first stage, regression, involves modelling the observable data. The second stage, dynamic programming (DP) or backward induction, uses the models to determine optimal actions, working iteratively from the last time stage. For the simple example there are eight different $T_1M_2T_2$ sequences and hence eight parameters in a saturated model for the response Y. Using the standard main effects and interaction formulation these are

Const	T_1	M_2	T_1M_2	T_2	T_1T_2	M_2T_2	$T_1M_2T_2$
3	3	-1	-1	5	-6	0	-2

From this we can calculate the mean response at each of the eight $T_1M_2T_2$ sequences and hence the regrets due to choices T_2 for each (T_1, M_2). Dealing with the first decision time is trickier: for each of the two values of T_1 we need to calculate

$$\sum_{M_2} E[Y|T_1, M_2, T_2^{opt}]P(M_2|T_1),$$

from which the optimal choice and regret can be found. In total, for general K we need $2^{2K-1} + 2^{2K-3} + \ldots + 2$ expectations. Moreover, we need the multivariate distribution of possible states M as well as our model for Y.

In order to complete our treatment of the simple example of Figure 4.1 we will anticipate a little and apply the regret-regression method to be described in the next section. Let

$I(t_1) = I(T_1 = t_1)$, $I(t_1 m_2) = I(T_1 = t_1, M_2 = m_2)$ and $I(t_1 m_2 t_2) = I(T_1 = t_1, M_2 = m_2, T_2 = t_2)$. Further, let $Z_2(t_1)$ be the residual between M_2 and its expected value given $T_1 = t_1$. So $Z_2(0) = M_2 - p_0$ and $Z_2(1) = M_2 - p_1$ where $p_0 = 0.8$ and $p_1 = 0.3$, both of which would need to be estimated in practice. Instead of the eight-parameter main effects and interaction model summarised in the previous section, we obtain exactly the same saturated fit using a linear model with the eight covariates given below, along with their associated parameter values.

Const	$I(1)$	$Z_2(0)I(0)$	$Z_2(1)I(1)$	$I(000)$	$I(010)$	$I(101)$	$I(111)$
7.2	-1.8	-1	-2	-5	-5	-1	-3

This time we can read off directly that the mean response if the optimal regime is always followed is 7.2. Choosing the wrong action at the first decision time will cost 1.8 in mean. The wrong actions at the second time lead to regrets of 5,5,1 and 3 depending on the earlier sequence (T_1, M_2). The other two terms measure the effects of M_2 *after* allowing for the effect of T_1. For each value of T_1 in this example having $M_1 = 1$ is associated with a decrease in mean: by $1 \times (1 - 0.8)$ if $T_1 = 0$ and by $2 \times (1 - 0.3)$ if $T_1 = 1$, in both cases assuming the optimal T_2 is chosen.

In fitting this model we have chosen the covariates knowing which action is optimal at each time. Reaching this point is trivial: we first take a working optimal for each decision and define the covariates accordingly. For example we might have included $I(001)$ instead of $I(000)$. At the first fit the signs of the final four coefficients determine which actions are indeed optimal: a positive value appears if the working version is wrong. We thus obtain the true optimal decisions at the second time and then re-fit the model. Generally, several iterations of model fitting are necessary. We need to re-fit the model with negative values of regrets at all time points from time K to time 2. This time the sign of the second coefficient - the regret at time 1 - determines the optimal. Thus only two model fits are required. Each requires a linear model and negligible effort.

4.2 Regret-Regression Method

With the regrets defined as in Equation (3.1), Murphy (2003, equation 12) showed that

$$E(Y|\bar{M}_K, \bar{T}_K) = \beta_0(M_1) + \sum_{j=2}^{K} \phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j) - \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}), \tag{4.1}$$

where

$$\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j) = E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j\} - E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}\} \tag{4.2}$$

which compares the expected potential response under the optimal rule *after* M_j is revealed with the corresponding expected value *before* M_j is revealed. Thus the achieved response Y is affected by the initial condition (through β_0), the chosen actions (T_j) (through the regrets μ) and the chance development over time of the states (M_j) (through the ϕ terms). Turning to estimation, both Murphy iterative minimization and Robins G-estimation methods require knowledge of the action probability distribution used in data generation, $P(t_j|\bar{M}_j, \bar{T}_{j-1})$. In a randomized trial this would of course be known, but more generally it needs to be estimated. Rosthoj et al (2006) speculated, with support from simulation experiments, that small misspecifications could lead to convergence difficulties in the estimation algorithms.

For our purposes this is sufficient: the regrets can be modelled directly. Throughout we suppose the three assumptions, which are discussed in Section 3.4.5. A second class to Assumption 3 can be added that the optimal regimes are taken to mean optimal over regimes which have positive probability of occurring in the sample data or which are parametrically identified from observable data. Furthermore we will add the following two assumptions

Assumption 4: *Finite second moment for Y.*

Assumption 5: *Between-subject independence.* The further assumptions of the independence and constant variance are to help ensure that our least squares are well behaved.

Our proposal is that instead of avoiding the $\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j)$ terms in Equation 4.2 we explicitly parameterize them, as $\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j; \beta)$ say, and then simultaneously estimate β and ψ by regressing the observed responses on their associated expectations Equation 4.1. We are not free to parameterize $\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j)$ arbitrarily however. As

$$\begin{aligned}\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j) &= E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j\} - E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}\} \\ &= E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j\} \\ &- E_{M_j|\bar{M}_{j-1},\bar{T}_{j-1}}[E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j\}],\end{aligned}$$

we see that by construction $E_{M_j|\bar{M}_{j-1},\bar{T}_{j-1}}\left\{\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j)\right\} = 0$. Any parameterization needs to respect this condition: the expected value over M_j of each $\phi_j(.)$ term, given the past, needs to be zero.

The proposal is straightforward: model $\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j)$ as a linear combination of residuals between M_j (or functions thereof) and the corresponding conditional expectation given $(\bar{M}_{j-1}, \bar{T}_{j-1})$. Hence we define $Z_j = M_j - E(M_j|\bar{M}_{j-1}, \bar{T}_{j-1})$ and note that the expectation is identified from observational data for $(\bar{M}_{j-1}, \bar{T}_{j-1})$ values of interest. Then assume

$$E(Y|\bar{M}_K, \bar{T}_K) = \beta_0(M_1) + \sum_{j=2}^{K} \beta_j^T(\bar{M}_{j-1}, \bar{T}_{j-1})Z_j - \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}). \qquad (4.3)$$

Here $\beta_j(\bar{M}_{j-1}, \bar{T}_{j-1})$ is a coefficient vector measuring the effect of M_j *after allowing* for $(\bar{M}_{j-1}, \bar{T}_{j-1})$ and assuming optimal actions are chosen from time j onward. Formally there can be a different coefficient for each possible history but in practice we may choose to simplify. Note that the zero mean requirement (given history) is immediate since the $\phi_j(.)$ terms are replaced by linear combinations of residuals.

We motivate the proposal by the following two theorems.

THEOREM 1 Assume $\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1};\psi)$ are correctly parameterized and M_j is made up of indicator variables comparing N_j categories with reference $M_j = 0$, e. g., $M_j = \{0, 1, 2, \cdots, N_j\}$. Then

$$E(Y|\bar{M}_K, \bar{T}_K) = \beta_0(M_1) + \sum_{j=2}^{K}\sum_{m=0}^{N_j} \beta_{jm}(\bar{M}_{j-1}, \bar{T}_{j-1})Z_{jm} - \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{A}_{j-1};\psi).$$

where Z_{jm} is component m of Z_j and

$$\beta_{jm}(\bar{M}_{j-1}, \bar{T}_{j-1}) = E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = m\} - E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 0\},$$

which measures the conditional effect of M_j after allowing for history and assuming optimal rules will be followed for future actions.

Proof

Assume that M_j is binary. Then

$$\begin{aligned}
\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 1) &= E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 1\} \\
&\quad - \big[E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 1\}P(M_j = 1|\bar{M}_{j-1}, \bar{T}_{j-1}) \\
&\quad + \ E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 0\}P(M_j = 0|\bar{M}_{j-1}, \bar{T}_{j-1})\big] \\
&= \big[E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 1\} - E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 0\}\big] \\
&\quad \times \left\{1 - P(M_j = 1|\bar{M}_{j-1}, \bar{T}_{j-1})\right\} \\
&= \beta(\bar{M}_{j-1}, \bar{T}_{j-1})\left\{1 - P(M_j = 1|\bar{M}_{j-1}, \bar{T}_{j-1})\right\}.
\end{aligned}$$

say. Similarly

$$\begin{aligned}
\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 0) &= E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 0\} \\
&\quad - \big[E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 1\}P(M_j = 1|\bar{M}_{j-1}, \bar{T}_{j-1}) \\
&\quad + \ E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}\}P(M_j = 0|\bar{M}_{j-1}, \bar{T}_{j-1})\big] \\
&= \big[E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 0\} - E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j = 1\}\big] \\
&\quad \times P(M_j = 1|\bar{M}_{j-1}, \bar{T}_{j-1}) \\
&= -\beta(\bar{M}_{j-1}, \bar{T}_{j-1})P(M_j = 1|\bar{M}_{j-1}, \bar{T}_{j-1}).
\end{aligned}$$

In both cases $\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j) = \beta(\bar{M}_{j-1}, \bar{T}_{j-1})Z_j$ where $Z_j = M_j - E(M_j|\bar{M}_{j-1}, \bar{T}_{j-1})$. Extension to categorical M_j is straightforward and Theorem 1 follows.

THEOREM 2. Assume $\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}; \psi)$ are correctly parameterized, all moments of each M_j exist and $E\{Y(\underline{d}_j^{opt})|\bar{M}_{j-1}, \bar{T}_{j-1}, M_j\}$ is analytic in M_j. Then for any $\epsilon > 0$ we can find positive integers N_j such that

$$E(Y|\bar{M}_K, \bar{T}_K) = \beta_1(M_1) + \sum_{j=2}^{K} \sum_{m=0}^{N_j} \beta_{jm}(\bar{M}_{j-1}, \bar{T}_{j-1})\tilde{Z}_{jm} - \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}; \psi) + R(\bar{M}_K, \bar{T}_K)$$

where $|R(\bar{M}_K, \bar{T}_K)| < \epsilon$. Here $\tilde{Z}_{jm}$ is element m of $\tilde{M}_j - E(\tilde{M}_j|\bar{M}_{j-1}, \bar{T}_{j-1})$ and $\tilde{M}_j$ is an N_j-vector made up of powers of components of M_j.

Proof

To prove Theorem 2, we need to express $\sum_{j=2}^{K} \phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j)$ defined at Equation 4.1 and Equation 4.3 as $\sum_{j=2}^{K} \sum_{m=0}^{N_j} \beta_{jm}(\bar{M}_{j-1}, \bar{T}_{j-1})\tilde{Z}_{jm} + R(\bar{M}_K, \bar{T}_K)$ as given in the statement of the theorem.

For notational simplicity, suppose M_j is a scalar and let us consider just one j. If $\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j)$ is analytic in M_j then we can expand it about $M_j = 0$ as a power series, say to order N_j. So

$$\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j) = \sum_{m=0}^{N_j} \phi_j^{(m)}(\bar{M}_{j-1}, \bar{T}_{j-1}, 0) \times M_j^m + R_{j,N_j}$$

where $\phi_j^{(m)}(.)$ is the m-derivative of $\phi_j(.)$ and R_{j,N_j} is the residual term, which can be made arbitrarily small by increasing N_j.

Let $\tilde{M}_j = (M_j, M_j^2, \ldots, M_j^{N_j})$ and

$\tilde{Z}_j = \tilde{M}_j - E(\tilde{M}_j|\bar{M}_{j-1}, \bar{T}_{j-1}) = (\tilde{Z}_{j1}, \tilde{Z}_{j2}, \ldots, \tilde{Z}_{j,N_j})$.

Substituting $M_j^m = \tilde{Z}_{jm} + E(M_j^m|\bar{M}_{j-1}, \bar{T}_{j-1})$ we have

$$\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j) = h(\bar{M}_{j-1}, \bar{T}_{j-1}) + \sum_{m=1}^{N_j} \phi_j^{(m)}(\bar{M}_{j-1}, \bar{T}_{j-1}, 0) \times \tilde{Z}_{jm} + R_{j,N_j}$$

where $h(.)$ does not depend upon the state M_j. Since $\phi_j(\bar{M}_{j-1}, \bar{T}_{j-1}, M_j)$ and each of the $\tilde{Z}_{jm}$ terms all have zero expectation over M_j given $(\bar{M}_{j-1}, \bar{T}_{j-1})$, it follows that $h(\bar{M}_{j-1}, \bar{T}_{j-1}) = 0$ necessarily. Redefining $\phi_j^{(m)}(\bar{M}_{j-1}, \bar{T}_{j-1}, 0)$ as the history-dependent coefficient $\beta_{jm}(\bar{M}_{j-1}, \bar{T}_{j-1})$

and repeating the argument for $j = 2, \ldots, K$ leads to the theorem. If M_j is not a scalar then the same arguments apply but with $\tilde{M}_j$ defined to include powers of products of elements of vector M_j.

Theorem 1 shows that Equation 4.3 is always true if M_j is categorical or discrete with bounded support. Theorem 2 shows that the proposal (with M_j redefined as $\tilde{M}_j$) can be an arbitrarily close approximation when the ϕ terms are analytic and moments of M_j exist, such as when there is bounded support. Other justifications of Equation 4.3 for other scenarios are possible for suitably rich $\tilde{M}_j$ but these are not developed here as we do not claim Equation 4.3 will *always* be true: it is an assumed model which we claim can be realistic in many circumstances. Further, gross departures from the assumed model should be detected by careful diagnostics. Our suggestion is thus summarized in the following, which in the sequel we refer to as *regret-regression*.

Proposal.

(i) Regress each M_j (or function of M_j) on history $(\bar{M}_{j-1}, \bar{T}_{j-1})$ and define $Z_j = M_j - E(M_j|\bar{M}_{j-1}, \bar{T}_{j-1})$.

(ii) Assume Equation 4.3 is true. Since $\beta_j(\bar{M}_{j-1}, \bar{T}_{j-1})$ depends on the possibly high-dimensional $(\bar{M}_{j-1}, \bar{T}_{j-1})$, we will need some modelling assumptions in all but the simplest applications unless sample size is huge. For example we might take a Markov rule by which $\beta_j(\bar{M}_{j-1}, \bar{T}_{j-1}) = \beta_j(M_{j-1}, T_{j-1})$ or we might assume stationarity over time. Call the vector of coefficients $\{\beta\}$.

(iii) Estimate the parameters $\{\beta\}$ and ψ by ordinary least squares. With subscript i for subject, we minimize

$$\sum_{i=1}^{n} \left\{ Y_i - \beta_0(M_{1i}) - \sum_{j=2}^{K} \beta_j^T(\bar{M}_{j-1,i}, \bar{T}_{j-1,i}) Z_{ji} - \sum_{j=1}^{K} \mu_j(T_{ji}|\bar{M}_{ji}, \bar{T}_{j-1,i}) \right\}^2 . \qquad (4.4)$$

(iv) Use bootstrap variance estimators, including re-estimation of the residuals (Z_{ji}) at each resample.

(v) Examine residuals between observed and fitted Y for diagnostic assessment.

The conditions which open this section are sufficient for consistent parameter estimation and valid causal inference for actions (T_j) (Henderson et al. 2009). Higher order moments of M_j may need to exist if residuals from non-linear terms are included in Equation 4.3.

4.3 Simulations

Two examples in this section illustrate the procedure. Our choice of ordinary least squares estimation is pragmatic: parameter estimation using standard software is quick and easy, facilitating bootstrap variance estimation. Note that we have modelled the mean response only and have not made any distributional assumptions about Y. If we do, then clearly we can replace the ordinary least squares estimation with maximum likelihood. Similarly, weighted least squares may be preferred if we have knowledge of the variance structure and sandwich-type variance estimators may be constructed as an alternative to bootstrap. See Diggle et al (1994).

4.3.1 Moodie, Richardson and Stephens scenario

Moodie, Richardson and Stephens (2007) use a simple two-time point simulation example (as explained in Chapter 3 Section 3.5). The example used simulations of one thousand repetitions at sample sizes $n = 500$ and $n = 1000$ of G-estimation. The results as described by Moodie et al (2007) can be seen in Table 4.2. Table 4.3 shows a summary of simulation results of one thousand repetitions at different sample sizes estimated by the regret-regression method proposed here. For the latter we used ordinary least squares to fit the correctly specified model $E[Y|\bar{M_2},\bar{T_2}] = \beta_0 + \beta_1 M_1 - \mu_1(T_1|M_1;\psi) - \mu_2(M_1|\bar{M_2},T_1;\psi)$.

G-estimation*				
	$(n = 500)$		$(n = 1000)$	
True ψ	Mean	SE	Mean	SE
250.0	250.01	17.170	249.450	12.160
-1.0	-1.00	0.038	-0.999	0.027
720	720.30	24.050	720.290	10.220
-2.0	-2.00	0.041	-2.001	0.029

Table 4.2: Summary of simulation results of one thousand repetitions based on Moodie et al scenario. $*$ These results are taken from Moodie et al (2007), who used the doubly robust form of g-estimation: their equation (2), which is the most efficient of the methods they considered.

Regret-regression								
		$(n = 100)$				$(n = 250)$		
True ψ	Mean	SE	BSE*	Cov*	Mean	SE	BSE	Cov
250	250.233	29.539	29.492	95	249.876	16.929	16.805	96
-1.0	-0.999	0.065	0.065	96	-0.998	0.0389	0.039	98
720	719.734	28.021	28.094	94	719.840	16.818	16.823	96
-2.0	-1.999	0.051	0.051	95	-2.000	0.030	0.030	94
		$(n = 500)$				$(n = 1000)$		
True ψ	Mean	SE	BSE*	Cov*	Mean	SE	BSE	Cov
250.0	250.066	11.077	11.037	96	249.820	7.639	7.646	96
-1.0	-0.999	0.026	0.026	94	-0.999	0.018	0.018	97
720	720.158	10.821	10.767	94	720.039	7.338	7.317	93
-2.0	-2.001	0.020	0.019	92	-1.999	0.014	0.014	94

Table 4.3: Summary of simulation results of one thousand repetitions using regret-regression method. $*$ BSE and Cov are standard errors and coverage of 95% bootstrap confidence intervals estimated from 100 bootstrap samples.

The `nlm` routine in R was used for parameter estimation. In all simulations the algorithm converged very quickly. Both methods produce apparently unbiased estimators, as they should, with smaller standard errors under the regret-regression method.

Table 4.4 investigates how estimated parameters translate into decision regime performance. After each repetition at sample sizes $n = 500$ a further 10000 observations were generated using each of four different decision rules: the gold standard of always choosing the optimal decision; equally likely randomised decisions; and following the estimated decision rules obtained by G-estimation of the regret functions and by the regret-regression procedure.

	Y_{1000}	SE	Err	Cut1	SE	Cut2	SE
Gold	1120.1	2.4	0.0	250.0		360.0	
Random	780.0	3.5	50.0				
Regrets (G-est)	1119.6	2.8	0.6	249.9	9.9	359.5	12.7
Regret-regression	1120.0	2.5	0.3	250.5	6.3	359.9	2.6

Table 4.4: Summary of simulation results based on Moodie et al scenario.

Column Y_{1000} gives the mean achieved response for each procedure, and column "Err" gives the overall percentage of times a suboptimal decision was made, pooled over both decision times. Columns 'Cut1' and 'Cut2' summarise the estimated cut points at each decision time, with the true values given in the gold standard row. Again both G-estimation and regret-regression perform well, with again less variability when regret-regression is used.

4.3.2 Murphy scenario

Murphy (2003), in her seminal paper introducing regret functions, described a more complex simulation scenario. The simulation was aimed to maximise the final response over 10 time points. In each time point there are two decisions, the first is a binary whether a child should receive special education ($T_{j1} = 1$), or not ($T_{j1} = 0$), where T_{j1} has a uniform distribution on $\{0, 1\}$. If $T_{j1} = 0$ then a second decision T_{j2} is a round of tutoring which is uniform on $(\{0,1,2,3\})$, with $P(T_{j2}) = 1/4$. If $T_{j1} = 1$ then a second decision T_{j2} is a round of special education chosen as uniform $(\{1,2,3\})$ with $P(T_{j2}) = 1/3$. For each interval two subintervals are created, each containing one decision. This gives the effect of considering 20 time intervals. Each individual starts with the initial status M_1 simulated as

$$M_1 \sim N(0.5, 0.01).$$

For $j \geq 2$

$$M_j \sim N(mean_j, 0.01),$$

where $mean_j = 0.5 + 0.2M_{j-1} - 0.07T_{\{j-1\}1}T_{(j-1)2} - 0.01(1 - T_{\{j-1\}1})T_{\{j-1\}2}$. The final response Y is assumed have optimal value 30, which might be reduced by the regrets. The regrets in each timepoint take different forms. For first decision they are

$$6\{t_{j1} - I(M_j > 5/9)\}^2,$$

and for the second they are

$$1.5t_{j1}(t_{j2} - 2m_j)^2 + 1.5(1 - t_{j1})(t_{j2} - 5.5m_j)^2.$$

Then using regrets, the simulated mean final response $E[Y|\bar{M}_{10} = \bar{m}_{10}, \bar{T}_{(10)2} = \bar{t}_{(10)2}]$, will be as follows,

$$\begin{aligned} 30 - 5\sum_{j=1}^{10}(m_j - mean_j) - \sum_{j=1}^{10} 6\{t_{j1} - I(m_j > 5/9)\}^2 \\ - \sum_{j=1}^{10} 1.5t_{j1}(t_{j2} - 2m_j)^2 + 1.5(1 - t_{j1})(t_{j2} - 5.5m_j)^2. \end{aligned}$$

where Y is normally distributed with the simulated mean and variance 0.64.

4.3.3 Estimation

Murphy (2003) estimated the regrets for each time point j by using

$$\mu_{j1}(t_{j1}|\bar{M}_j, \bar{T}_{(j-1)2}) = \psi_1\{t_{j1} - I(M_j > \psi_2)\}^2$$

Then for part 2 of treatment j, the regret takes the correct form

$$\mu_{j2}(t_{j2}|\bar{M}_j, \bar{T}_{j1}) = \psi_4 t_{j1}\{t_{j2} - (\psi_3 + \psi_5 m_j)\}^2 + \psi_7(1 - t_{j1})\{t_{j2}(\psi_6 - \psi_8 m_j)\}^2.$$

Returning to Chapter 3 Section 2.6, the following equation

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}[Y_i + \sum_{l=1,l\neq j}^{k} \mu_l(\bar{M}_l, \bar{T}_l; \hat{\psi}_n) + \mu_j(\bar{M}_j, \bar{T}_j; \psi) - \sum_t \mu_j(\bar{M}_j, \bar{T}_j, t)p_j(t|\bar{M}_j, \bar{T}_j)]^2$$

can be used by the iterative procedure gives the estimated parameters. The assumption is made that the regrets will be zero at the optimal actions. For the first part in action j the optimal decision is

$$d_{j1}^{opt} = I\{m_j > 5/9\},$$

and for the second,

$$d_{j1}^{opt} = 2t_{j1}m_j + 5.5(1 - t_{j1})m_j.$$

Note that this is already in the form Equation 4.3, and that once estimates of ψ are available the estimated optimal regime is to choose actions which lead to zero estimated regrets.

	Murphy*	
True ψ	Mean	SE
6.00	6.89	0.210
0.56	0.56	0.002
0.00	0.05	0.184
1.50	1.50	0.125
2.00	2.01	0.255
0.00	0.06	0.128
1.50	1.48	0.078
5.50	5.54	0.358

Table 4.5: Summary of simulation results based on Murphy scenario. One thousand repetitions at sample size $n = 1000$. Results $*$ are taken from Murphy (2003).

Regret-regression								
	$(n = 250)$				$(n = 500)$			
True	Mean	SE	BSE	Cov	Mean	SE	BSE	Cov
ψ								
6.00	5.967	0.062	0.061	94	5.970	0.043	0.043	95
0.56	0.555	0.0005	0.0005	95	0.555	0.0005	0.0005	95
0.00	-0.020	0.073	0.072	98	-0.015	0.053	0.053	96
1.50	1.489	0.060	0.060	96	1.493	0.041	0.041	95
2.00	2.021	0.093	0.091	96	2.018	0.069	0.069	96
0.00	-0.009	0.035	0.034	94	-0.006	0.022	0.022	95
1.50	1.493	0.030	0.028	95	1.495	0.016	0.016	96
5.50	5.531	0.074	0.071	95	5.519	0.052	0.052	96
β								
30.00	29.928	0.587	0.534	96	29.913	0.251	0.251	96
-5.00	-4.981	0.353	0.342	95	-4.981	0.213	0.213	96
	$(n = 750)$				$(n = 1000)$			
True	Mean	SE	BSE	Cov	Mean	SE	BSE	Cov
ψ								
6.00	5.968	0.032	0.032	94	5.971	0.029	0.029	95
0.56	0.555	0.0004	0.0004	97	0.555	0.0004	0.0004	94
0.00	-0.019	0.039	0.039	98	-0.017	0.036	0.036	95
1.50	1.491	0.030	0.030	95	1.491	0.029	0.029	95
2.00	2.024	0.054	0.053	96	2.021	0.047	0.047	6
0.00	-0.005	0.016	0.016	97	-0.004	0.015	0.015	95
1.50	1.498	0.014	0.0141	95	1.496	0.011	0.011	95
5.50	5.517	0.039	0.0387	93	5.515	0.037	0.037	95
β								
30.00	29.940	0.186	0.1848	95	29.909	0.167	0.167	96
-5.00	-5.010	0.172	0.1684	96	-4.995	0.157	0.157	95

Table 4.6: Simulation results of one thousand repetitions based on Murphy scenario using regret-regression method.

Tables 4.5 and 4.6 compares results obtained using Murphy's iterative estimation method with those from regret-regression, which took a correctly specified model for Y and also required fitting the state evolution model. Murphy approximated the indicator function in μ_1 by a very steep logistic curve in order to have a smooth function. There was no need for us to do this: for fixed ψ_2 we used Newton-Raphson to obtain, very quickly, least squares estimates of the other parameters and then simply searched for ψ_2. This was not feasible using Murphy's estimation technique, which converges only slowly.

	Gold	Random	Murphy*	Regret-regression
Mean	30	-39.56	29.27	29.88
Standard deviation	0.018	0.203	0.19	0.024

∗ *These results are taken from Murphy (2003).*

Table 4.7: Summary of simulated performance of decision rules based on Murphy.

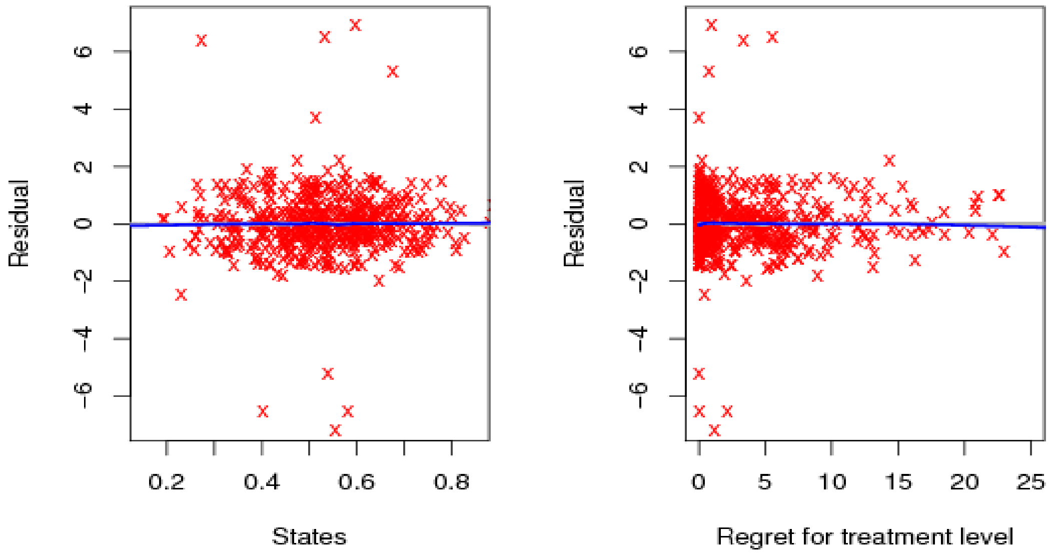

Figure 4.2: Residual plot for Murphy simulation with correct model fit, $n = 1000$.

Results in Table 4.6 show the regret-regression method works very well in parameter estimation. Table 4.7 shows performance of decision rules. Parameter estimates were obtained as for Table 4.5 and after each repetition a further 10000 observations were generated using each of four different decision rules: the gold standard of always choosing the optimal decision; randomised decisions as described by Murphy for the original data generation; and following the estimated decision rules obtained by regret models fitted by the Murphy iterative technique or the regret-regression method.

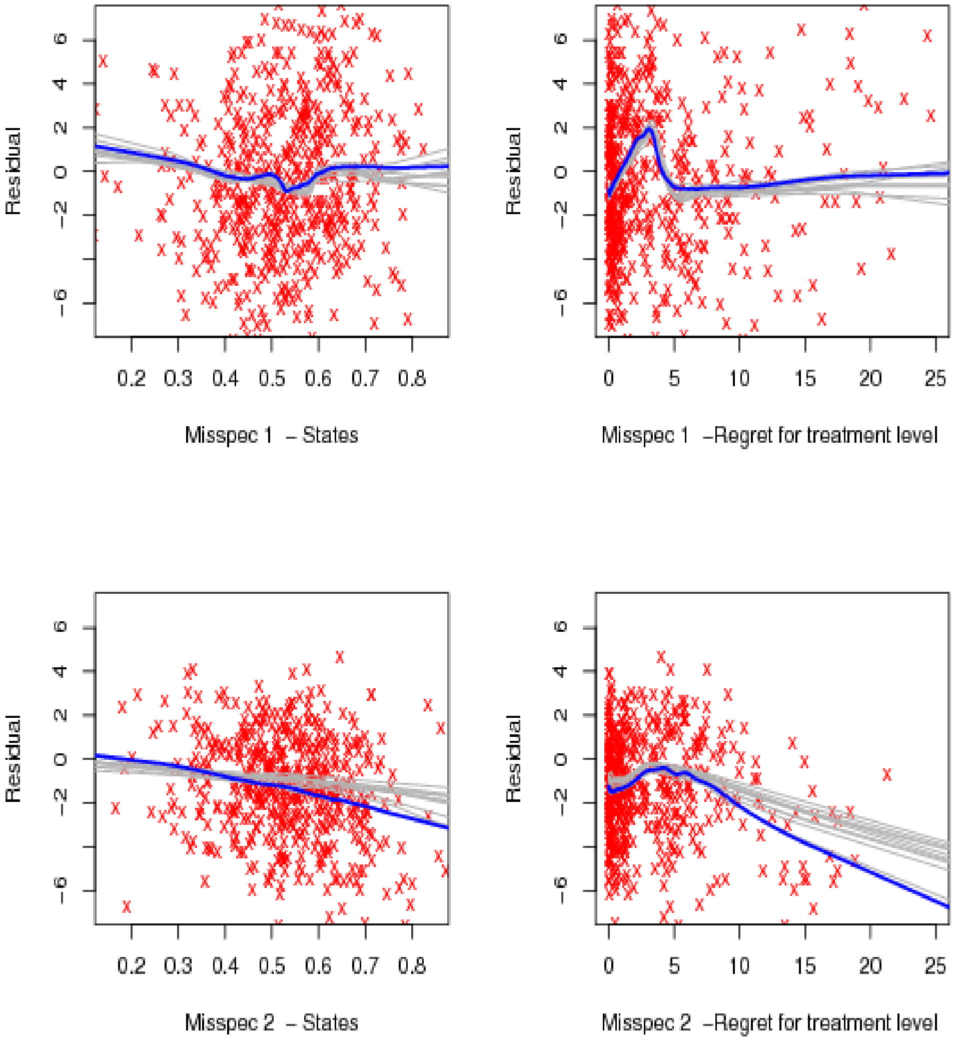

Figure 4.3: Residual plots for Murphy simulation with her misspecified fits, $n = 1000$.

Turning now to diagnostics, Figure 4.2 shows, for a typical simulation, residuals from the model fit plotted against state (left plot) and regret for the second-time decision on amount of treatment (right plot). There are 10 states and 10 second-time actions for each subject but for presentation purposes we have pooled into single plots. Also, the plot shows just a 10% random sample of points. Superimposed on each plot is a smooth (using R `supsmu`) trend through the complete data, which in each case looks like a horizontal straight line at value zero. There is thus, as expected, no evidence against the fitted model. For reference, the figure also includes as grey lines similar smooth trends for ten other simulations. These are almost all obscured by the original line as in all cases the mean was constant at zero. Murphy (2003) briefly discussed model misspecification and described two alternative parametric forms for the regret functions. These misspecified models will be described well in Chapter 7, Section 7.1.2. So we generate data using the original model but we fit the other misspecified models. Residuals are replicated to account for 10 states and 10 even regrets for each value of Y. Also shown (as grey lines) are smooth fits to 10 further samples to give an indication of between-sample variability. In Figure 4.3 shows residuals when each of the alternative models are true but the model described above is fitted. In both cases it is clear that the diagnostics would pick up the misspecification. The crosses in Figure 4.3 show a 10% sample of residuals and the solid line is a smooth fit to the complete data. The left plots compare residuals with the state observations and the right plots with the fitted even regret values. The top row is based on Murphy's first misspecification and the bottom row on her second misspecification. Also shown (as grey lines) are smooth fits to 10 further samples to give an indication of between-sample variability.

4.4 Application: Anticoagulation Dosage

Our application is on the Warfarin data. Rosthøj et al (2006) attempted to fit regret models to data from 303 patients given warfarin treatment for anticoagulation. The data are described in Chapter 1,Section 3.2, but in brief:

- M_j is the standardised difference between INR, a measure of blood clotting speed, and the target range. About half the values were $M_j = 0$, meaning within range. Positive M_j means clotting time was too long and anticoagulation dose might be decreased, negative the opposite.

- T_j is the change in dose at visit j. There was no change in dose on about 60% of occasions.

- Fourteen measurements per person were used, but the first four were treated as a stabilisation period and the final action had no effect, meaning $K = 9$ in the analyses.

- Response Y was the overall percentage of time INR was within target range.

In this application we suppose the assumptions A1-A5, which are discussed in Section 3.4.5 and Section 4.2 that we believe non of them are violated. Rosthøj at al (2007) were able successfully to fit just one simple regret model:

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = \begin{cases} I(t_j \neq 0)(5.84 + 1.59t_j^2) & M_j = 0 \\ 0.24(t_j + 2.01M_j)^2 & M_j \neq 0. \end{cases}$$

We refer to the previous model as Model M1. In order to assess the suitability of the model, we re-estimate ψ's using the regret-regression method

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = \begin{cases} I(t_j \neq 0)(\psi_1 + \psi_2 t_j^2) & M_j = 0 \\ \psi_3(t_j + \psi_4 M_j)^2 & M_j \neq 0 \end{cases}$$

and re-name it as Model M2. We need residuals $Z_j = M_j - E[M_j|\bar{M}_{j-1}, \bar{T}_{j-1}]$. These were obtained from a mixture model for M_j with a logistic component for $P(M_j = 0)$ and a linear component for $|M_j|$ given $(M_j \neq 0)$.

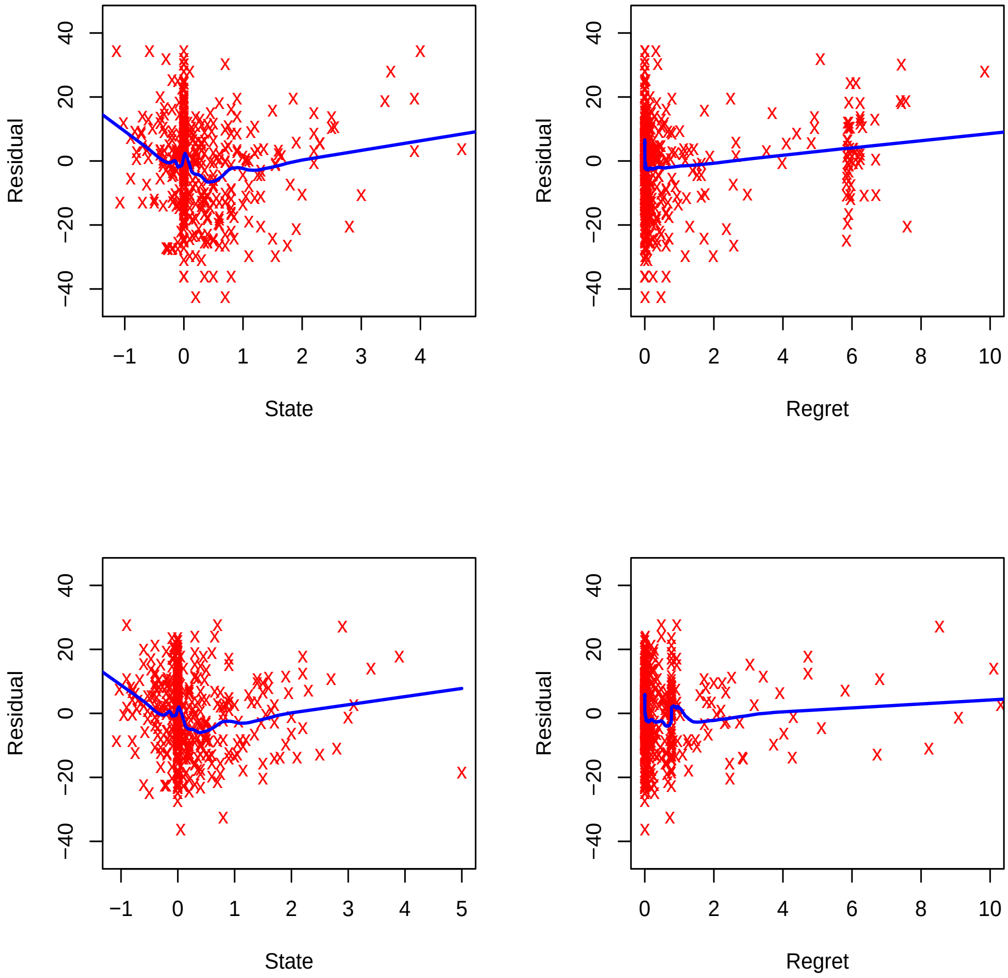

Figure 4.4: Residual plots for models M1 and M2.

In both cases we included as covariates 17 functions of previous states and actions up to lag 4, including all main effects and some pairwise interaction and quadratic terms, and both actual values and indicators of value zero. We used the same model at all time points but estimated the coefficients separately. All terms were not always needed in the model but since there is no interest in the model for M_j per se there is no loss in modest over fitting. We assumed the non-regret residual contribution to response Y was linear in Z_j for all models. In Figure 4.4 the crosses show a 20% sample of residuals and the solid lines show a smooth fit to the complete data. The left plots compare residuals with the state observations and the right plots with the fitted regret values. The top row is based on the Rosthøj et al (2007) regret fit (**Model M1**) and the bottom row of Figure 4.4 gives residual plots for the regret-regression fit based on the Rosthøj et al regret function, with new parameter values (**Model M2**). Neither model seems adequate given there residuals. A variety of models were also considered. Some, not all, are listed below.

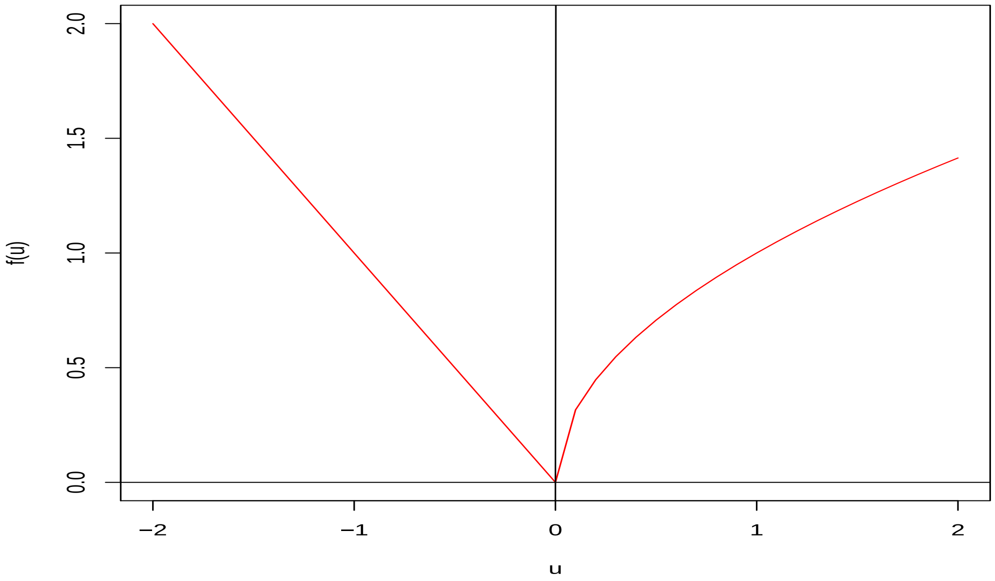

Figure 4.5: Regret function in Model M5, , where $f(u) = |u|$ if $u < 0$ and $f(u) = \sqrt{u}$ otherwise.

Model M3: This uses regrets as proportional to M2 with time-varying coefficients

$$\tilde{\mu}_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = \begin{cases} I(t_j \neq 0)(\psi_1 + \psi_2 t_j^2) & M_j = 0 \\ \psi_3(t_j + \psi_4 M_j)^2 & M_j \neq 0 \end{cases}$$

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = \psi_{3+j}\tilde{\mu}_j(t_j|\bar{M}_j, \bar{T}_{j-1}) \qquad j > 1$$

Model M4: This model suggests to use lagged effects of a subset of previous covariates and actions

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = \psi_1 \left(t_j + \psi_2 M_j + \psi_3 M_j^2 + \psi_4 M_{j-1} + \psi_5 T_{j-1} + \psi_6 T_j T_{j-1}\right)^2.$$

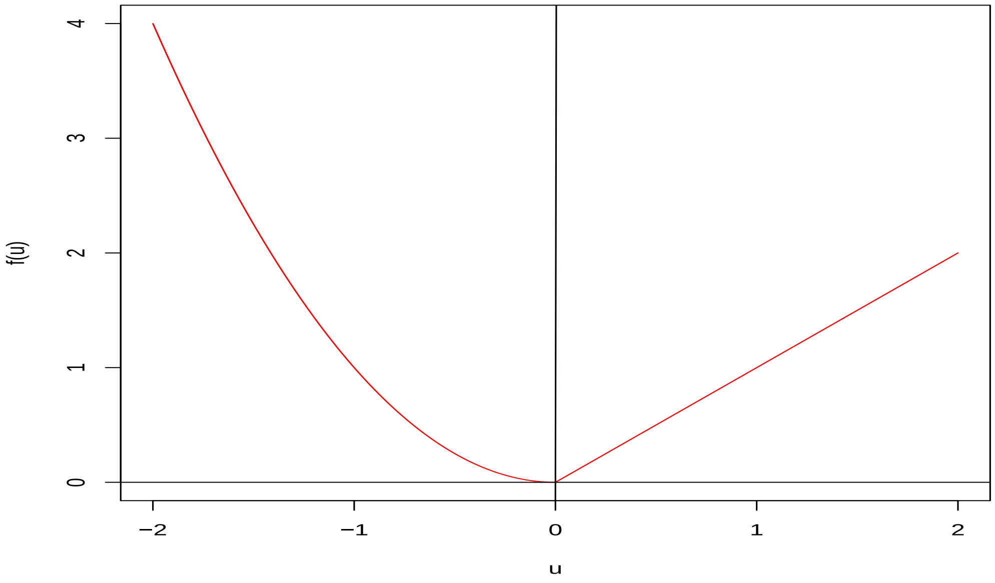

Figure 4.6: Regrets function in both models M6 and M7, where $f(u) = u^2$ if $u < 0$ and $f(u) = u$ otherwise.

Model M5: The model uses asymmetric regrets with categorised M_j,

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}, M_j^* = m; \psi) = \psi_{m1} f(T_j - \psi_{m2}),$$

where the optimal decision rule is $d_j^{opt} = \psi_{m2}$. Here we categorised M_j into 5 states. $M_j^* = \{-2, -1, 0, 1, 2\}$. If the INR was in range then $M_j^* = 0$ otherwise M_j^* is defined according to whether M_j is above or below the positive or negative median. The model uses the link function $f(u) = \sqrt{u}$ if $u \geq 0$ and $f(u) = |u|$ otherwise. Figure 4.5 describes shape of regrets function in Model M5.

Model M6: The model uses asymmetric regrets with categorised M_j,

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}, M_j^* = m; \psi) = \psi_{m1} f(T_j - \psi_{m2})$$

which is exactly the same with Model M5, but using a different link function $f(u) = u$ if $u \geq 0$ and $f(u) = u^2$ otherwise.

Model M7: The model uses asymmetric regrets with categorised M_j and lagged M_{j-1}^* effect,

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}, M_j^* = m; \psi) = \psi_{m1} f(T_j - \psi_{m2} - \psi_{m3} M_{j-1}^*)$$

where the optimal decision rule is $d_j^{opt} = \psi_{m2} + \psi_{m3} M_{j-1}^*$. For simplicity, we use $M_{j-1}^* = \{-1, 0, 1\}$. The model uses the same link function as Model M6. The shape of the link function is shown in Figure 4.6.

Figure 4.7 shows residuals for some of the other regret-regression fits. The top row is based on Model M4 then the middle row is based on Model M5 and the bottom row on the final fit Model M7. Table 4.8 shows that the residual sum of squares for response Y is reduced by nearly 60% for this regret-regression model (M0 v M1), where Model M0 is a null model with $SSR = \sum(Y - \bar{Y})^2$. With the same model but re-estimating parameters there is a reduction of about 68%.

Fitting was quick and easy in all cases, based on a combination of a least-squares fit for linear terms embedded in a numerical search (using R function `nlm`) for non-linear terms.

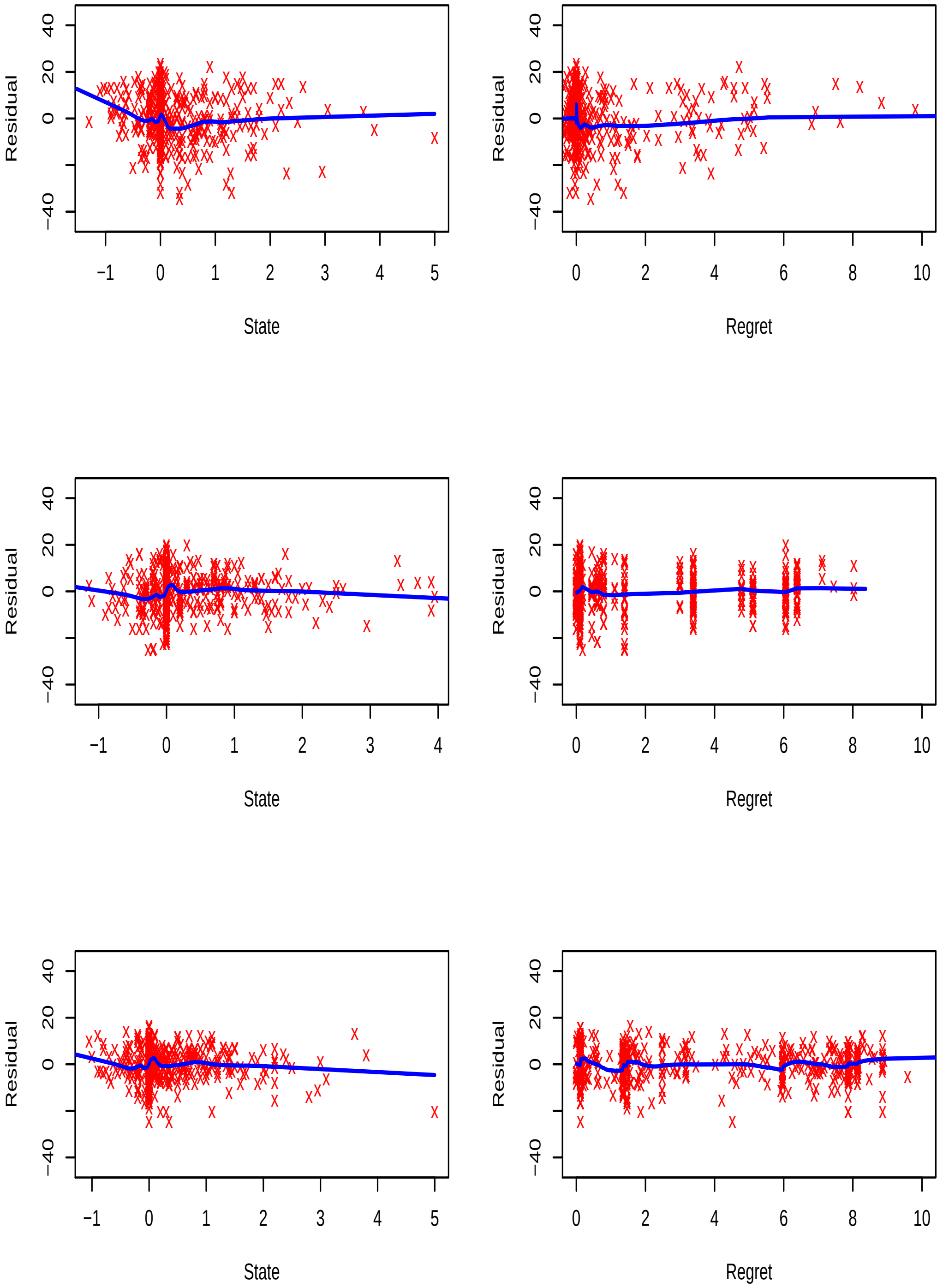

Figure 4.7: Residual plots for models M4, M5 and M7.

We bounded optimal actions at ±3 (only three from 2727 observed actions were outside this range). A separate asymmetric regret function was fitted for each category of state.

Model	dim(ψ)	RSS	Adjusted R^2
			(%)
M0: Null	-	125234.5	-
M1: Rosthøj	4	50373.1	57.6
M2: Rosthøj, re-estimated	4	39829.9	68.2
M3: Rosthøj, time varying	12	35817.9	69.0
M4: Lag in states and actions	6	31007.9	73.7
M5: Discrete M_j and T_j, asymmetric	10	20208.9	82.7
M6: Discrete M_j, asymmetric	10	20997.7	81.9
M7: Discrete M_j, asymmetric, lag term	15	16347.4	85.7

Table 4.8: Residual sums of squares under various regret-regression models for anticoagulation data. See Section 5 and Appendix for model descriptions

This allows in a simple way the magnitude of regret, not just the optimal decision, to depend on M_j as well as distance of T_j from the optimum. Not having this flexibility was given by Robins (2004) as an argument against a regret class proposed by Murphy (2003).

$M_j^* = m$	ψ_{m1}	SE	ψ_{m2}	SE	ψ_{m2}	SE
-2	0.67	0.30	2.15	0.23	-1.11	0.35
-1	0.38	0.10	2.74	0.16	-1.57	0.69
0	0.97	0.37	-0.14	0.28	-1.12	0.68
1	2.38	0.27	-2.33	0.27	-0.98	0.24
2	2.83	0.94	-3.00	0.48	0.25	0.21

Table 4.9: Parameter estimates and bootstrap standard errors for anticoagulation model M7

For category m of state M_j the assumed regret function was

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}, M_j^* = m; \psi) = \psi_{m1} f(T_j - \psi_{m2} - \psi_{m3} M_{j-1}^*)$$

where $f(u) = u$ if $u \geq 0$ and $f(u) = u^2$ otherwise. Parameter estimates with bootstrap standard error are given in Table 4.9.

It took under 20 minutes to obtain estimates from the 100 bootstrap samples (using R rather than a faster language). The illustrative regrets shown in Figure 4.8 suggest together with Table 4.9 that the results are intuitively reasonable. Each subplot corresponds to different value of the current categorized state variable. The three lines indicate whether the previous state was in range (green solid lines), above range (blue dotted lines), or below range (red dashed lines). From $\{\psi_{m2}\}$ we see that if the current state M_j is low (clotting time too long) then an increase in dose is suggested; if the current state M_j is high then a decrease is suggested; and if the patient is in range ($M_j = 0$) then the effective optimum is not to change. Inspection of $\{\psi_{m3}\}$ indicates that for the first four categories of M_j^* the optimal actions are moderated by the previous state M_{j-1} also in the expected way: more drastic dose changes are indicated if two successive states have the same sign; attenuation toward no-change if they have opposite signs. For the fifth category, M_{j-1} seems unimportant for optimal dose. In terms of regret for suboptimal action, we see from the ψ_{m1} column that the consequences of the poor choice are more severe for patients whose M_j values are high, and hence have clotting time which is too long. This model explained 87% of the sum of squared response residuals (Table 4.8) and the diagnostic plots shown in the bottom row of Figure 4.4 and Figure 4.7 show what we consider to be a very reasonable fit.

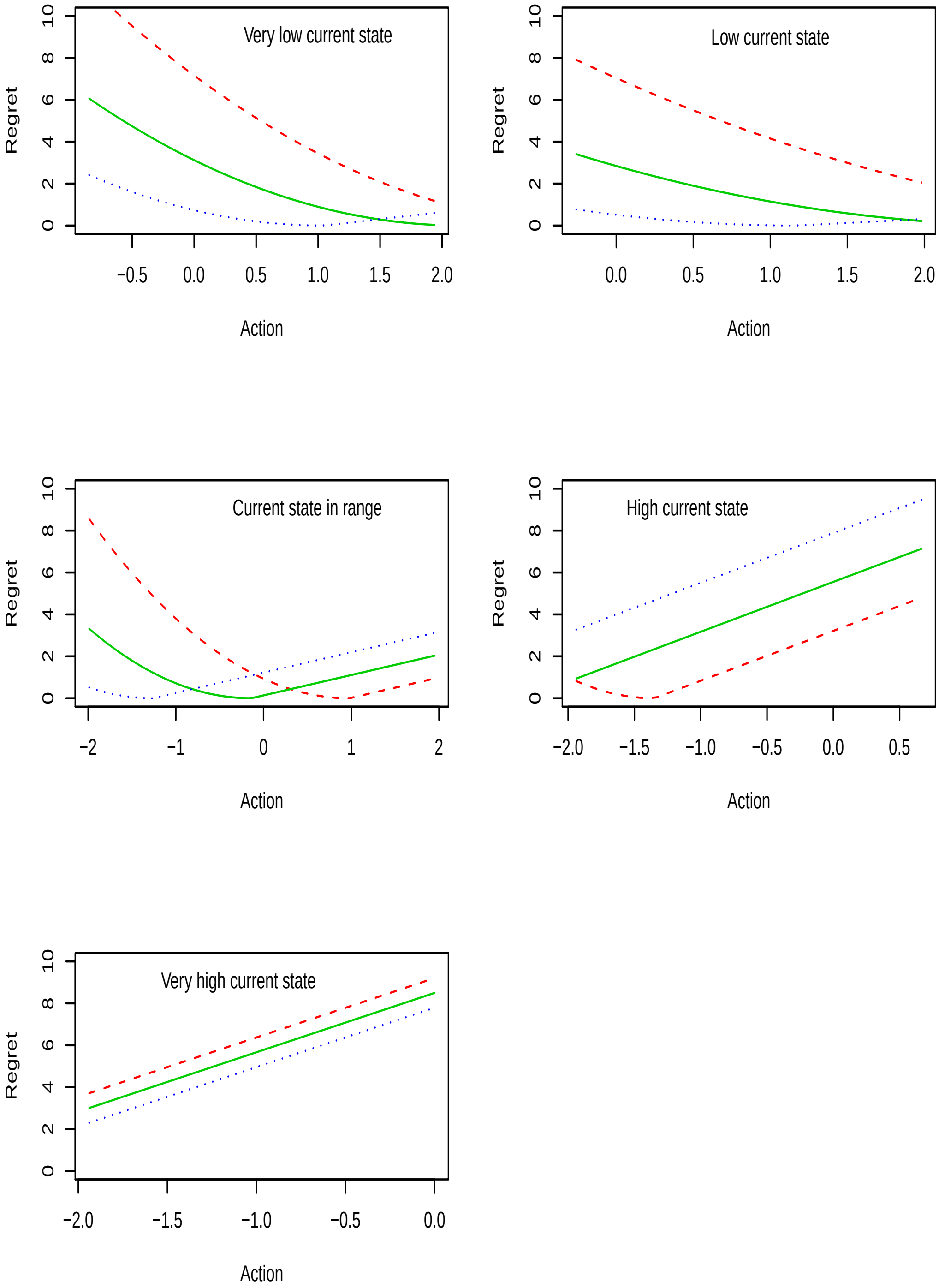

Figure 4.8: Estimated regrets functions under Model M7.

4.5 Discussion

We have proposed a method for finding optimal dynamic treatment regimes. The method has not required $E[T_j|\bar{M}_j, \bar{T}_{j-1}]$, which is needed for the other methods. Inclusion of the linear combination of residuals in Equation 4.3 is reminiscent of a path analysis method for dealing with time-dependent confounders, as exemplified by Borgan et al (2006) for instance. This brings additional modelling assumptions, not needed by Murphy or Robins, and hence the possibility of misspecification. If the model is correct however, or close to correct, then we expect gains in efficiency.

Murphy (2003) had primary interest in the parameters of the regrets, and considered other unknown functions involved in data generation as being nuisance parameters. We agree in part only: unless sample size is enormous we see no alternative to assuming parametric regret models. In that case some form of diagnostic is essential for good statistical practice, and development of diagnostics based on models for observables is an obvious way forward. The use of residual plots to detect misspecified regret functions was illustrated in Section 4.2. Chapter 7 on diagnostic methods has further investigation to investigate power of proposed tests and to understand more how residual means react to different types of misspecification.

Chapter 5

Regret-Regression and Inverse Probability of Treatment Weighting

5.1 Introduction

As seen in the previous chapter, we have more efficiency by choosing a fully parameterised model for the mean. The regret-regression method for estimating the causal effect of the actions on the final response allows diagnostic model assessment and model comparisons. It requires modelling of $E(M_j|\bar{M}_{j-1}, \bar{T}_{j-1})$, to remove the effect of the time-dependent confounders on the counterfactuals, which is not used for the Murphy (2003) iterative estimation or the Robins (2004) G-estimation methods. On the other hand regret-regression has not required $E[T_j|\bar{M}_j, \bar{T}_{j-1}]$, which is needed for the others. In this chapter a comparison of the regret-regression and inverse probability of treatment weighting is presented. Let us assume that each of N study subjects, are either treated $T = 1$ or not $T = 0$. We observe an outcome Y measured at the end of follow-up, and a vector M of baseline covariates. Let Y_t denote the counterfactual or potential outcome for a subject under treatment level $T = t$. We have two counterfactual variables $Y(t = 1)$ and $Y(t = 0)$. For example, if a subject's outcome would be 8 under treatment and would be 3 under non treatment, then we can write $Y(t = 1) = 8$, $Y(t = 0) = 3$ and $Y(t = 1) - Y(t = 0) = 5$. For the actual study, if this subject was treated, then his observed Y will be 8. Furthermore an observed outcome Y is the counterfactual outcome $Y(t)$ for a subject who is treated with level t. As discussed in Chapter 2, Section 4, there are three assumptions sufficient for causal inference. When those three identifiability conditions hold, one can use any of the two methods discussed below *G-formula* or *Inverse Probability of Treatment Weighted* (IPTW) to consistently estimate $E(Y_t)$. They are used to evaluate a fixed treatment regimes from observational data with time varying covariates as explained in secions 5.2 and 5.3. In this Chapter, our aims are to compare results of these two methods with regression analysis (Secion 5.4) and with the regret-regression method. Then in Section 5.6 we show how to use the G-formula or the IPTW for finding optimal dynamic treatment strategies. In this case we change our worked example (from Chapter 4), to a similar one that optimal decisions T_2 depend on the value of M_2.

5.2 The G-Formula

The traditional approach to estimate causal parameters uses the G-computation formula from Robins (1986) and Robins (1987). The G-computation formula is a general conditional expectation of a counterfactual given earlier history that can be further used to estimate counterfactual means $E[Y_d]$, under any static or dynamic regime d (Robins uses (g) instead of using d). This methodology relies on the model used for $E(Y|M, T)$ (Neugebauer and van der Laan 2002). For a given value t of T and vector M of baseline covariates, the G-formula (based on covariates M) for $E[Y(t)]$ is defined in this simple case as

$$E[Y(t)] = \sum E_m(Y|M = m, T = t)P(M = m),$$

where the sum is for all values m of M in the population. This equality is a result from the Assumption 2 of no unmeasured confounders. The G-formula for $E[Y(t=1)]$ is the standardized mean of Y in the group of patients with $T=1$. Note the G-formula depends on the distribution in the population of the observed variables T, M and Y. In practice, this distribution will be estimated from the study data. When M takes values on a continuous scale, then the sum $\sum$ is replaced by an integral, and the G-formula for this simple case

$$E[Y(t)] = \int E(Y|M=m, T=t)\ dF(M=m).$$

5.3 The IPTW Formula

As alternative approach to estimate causal parameters is to use the Inverse Probability of Treatment Weighted (IPTW) method. The IPTW formulas for $E[Y(t=1)]$ and $E[Y(t=0)]$ based on M are the means of Y using $T=1$ and $T=0$ respectively in a pseudo-population constructed by weighting each subject in the population by their inverse probability of treatment weight (IPTW)

$$SW = \frac{P(T=t)}{P(T=t|M=m)},$$

When M takes continuous values, then

$$SW = \frac{f(T)}{f(T|M)}.$$

where $f(T)$ and $f(T|M)$ are the probability density functions evaluated at the subject's data T, and T given M, respectively. In a randomized experiment, $f(T|M)$ is known by design, but in an observational study we have to estimate it from the data. We refer to SW as stabilized weights and to the pseudo-population created by weights as a stabilized pseudo-population. In fact, we could alternatively create an unstabilized pseudo-population by weighting each subject by their unstabilized weight

$$W = \frac{1}{f(T|M)}$$

However, we can use stabilized or unstabilized weights. The IPTW formula for $E[Y(t)]$ in the stabilized and unstabilized populations is

$$E[Y(t)] = E\left[\frac{I(T=t)f(T)}{f(T|M)}Y\right]/E\left[\frac{I(T=t)f(T)}{f(T|M)}\right]$$

and

$$E[Y(t)] = E\left[\frac{I(T=t)}{f(T|M)}Y\right]/E\left[\frac{I(T=t)}{f(T|M)}\right]$$

respectively. Here

(i) $E[I(T=t)f(T)N/f(T|M)]$ and $E[I(T=t)N/f(T|M)]$ are the number of subjects in the stabilized and unstabilized pseudo-populations with $T=t$, and

(ii) $E[I(T=t)f(T)NY/f(T|M)]$ and $E[I(T=t)NY/f(T|M)]$ are the sum of their Y values.

Hernán and Robins (2006) discuss the mathematical equivalence between the g-formula, standardization and IPTW under positivity. The equivalence is based on the mathematical identities

$$\begin{aligned} E[Y(t)] &= E\left[\frac{I(T=t)f(T)}{f(T|M)}Y\right]/E\left[\frac{I(T=t)f(T)}{f(T|M)}\right], \\ &= E\left[\frac{I(T=t)}{f(T|M)}Y\right]/E\left[\frac{I(T=t)}{f(T|M)}\right], \\ &= \int E(Y|M=m, T=t)\; dF(M=m). \end{aligned}$$

When treatment is unconditionally randomized (see Figure 2.2*a* in Chapter 2, Section 5) both the G-formula and the IPTW estimate for $E[Y(t)]$ are equal to the mean $E(Y|T=t)$

of Y among those with treatment level t in the population because T and M are independent. When the randomization is conditional on M (Figure 2.2*b*), then the average causal effect differs from the $E(Y|T=1)-E(Y|T=0)$ and data on M are needed to consistently estimate $E(Y_t)$. The G-formula estimates $E[Y(t)]$ by the joint distribution of the variables M, T, and Y that would have been observed in a study in which every subject received t.

The IPTW method effectively mimics data where the treatment T is independent of M so that, if the causal graph in Figure 2.2*b* holds in the actual population, the causal graph in Figure 2.2*a* with no arrow from M to T will hold in the pseudo-population. The only difference between stabilized and unstabilized IPTW is that in the unstabilized pseudo-population $P(T=1)=0.5$ while in the stabilized pseudo-population $P(T=1)$ is as in the actual population. Thus $E[Y_t]$ in the actual population is $E_{ps}(Y|T=t)$ where the subscript *ps* is to remind us that we are taking the average of Y among subjects with $T=t$ in either pseudo-population.

In summary, when the three identifiability conditions hold, the average causal effect $E[Y(t=1)]-E[Y(t=0)]$ in the population is the difference $E_{ps}(Y|T=1)-E_{ps}(Y|T=0)$ in the pseudo-population.

What about observational studies? Let us assume that the three identifiability conditions consistency, conditional exchangeability, positivity are met in a particular observational study. Then there is no conceptual difference between such an observational study and a randomized experiment. The three conditions imply that the observational study can be the same as a randomized experiment and hence that the G-formula, or IPTW can also be used to estimate counterfactual $E(Y_t)$ from the observational data. When the consistency and conditional exchangeability conditions fail to hold, the IPTW and G-formula for $E(Y_t)$

based on M are still well defined and can be estimated from the observed data; however the formulas no longer equal $E(Y_t)$ and thus do not have the causal interpretation as the mean of Y had all subjects received treatment t. When positivity fails to hold for treatment level t, the IPTW formula remains well defined but fails to equal $E(Y_t)$, while the G-formula is undefined (Hernán and Robins, 2006).

The G-formula, and IPTW can provide consistent estimates of counterfactual quantities like $E[Y_{\bar{t}}]$ under generalizations of our previous definitions of consistency, conditional exchangeability, and positivity.

The IPTW formula based on $\bar{M}$ for the counterfactual mean $E[Y(t)]$ is the average of Y among subjects with $\bar{T} = \bar{t}$ in a stabilized or unstabilized pseudo-population constructed by weighting each subject by their subject-specific stabilized IPTW

$$SW = \prod_{j=1}^{K} \frac{f(T_j|\bar{T}_{j-1})}{f(T_j|\bar{T}_{j-1}, \bar{M}_j)},$$

or their unstabilized IPTW

$$W = \prod_{j=1}^{K} \frac{1}{f(T_j|\bar{T}_{j-1}, \bar{M}_j)}$$

When the three identifiability conditions hold, either IPTW creates a pseudo-population in which the mean of Y_t is identical to that in the actual population but like on DAG (Figure 2.2a), the randomization probabilities at each time j depend at most on past treatment history. The only difference is that in the unstabilized pseudo-population $P_{ps} = (T_K = 1|\bar{T}_{K-1}, \bar{M}_K) = 0.5$.

The following is an extremely simple example described in Chapter 4, Section 1.1. All subjects are assumed to start in the same state M_1, which can thus be ignored. We consider a sequentially randomized trial in which $N = 1000$. Patients are randomly assigned at time $K = 1$ to treatment $T_1 = 1$ with probability 0.5 and to placebo $T_1 = 0$ otherwise. Patients continue on treatment or placebo until their next visit to clinic at time $K = 2$, where they are again randomly assigned to take treatment with probabilities

$$P00 = P(T_2 = 1|T_1 = 0, M_2 = 0) = 0.5,$$

$$P01 = P(T_2 = 1|T_1 = 0, M_2 = 1) = 0.7,$$

$$P10 = P(T_2 = 1|T_1 = 1, M_2 = 0) = 0.2,$$

$$P11 = P(T_2 = 1|T_1 = 1, M_2 = 1) = 0.1.$$

Table 4.1 and Figure 4.1 provide the number of subjects and the average value of the outcome $E[Y|T_1, M_2, T_2]$. Note we ignore M_1, because we assume that all subjects start in the same baseline M_1. We suppose consistency. Further, we can conclude that the positivity condition is satisfied, because otherwise one or more of the eight rows would have no subjects.

G-formula

If we can estimate the counterfactual means $E[Y(\bar{t} = \{0, 0\})]$, $E[Y(\bar{t} = \{1, 0\})]$, $E[Y(\bar{t} = \{0, 1\})]$ and $E[Y(\bar{t} = \{1, 1\})]$ under the 4 possible static regimes. All four means can be consistently estimated by the G-formula, because the three identifiability conditions hold in a sequentially randomized trial. Because the confounder M_2 is a binary variable the G-formula can be written as

$$\begin{aligned} E[Y(\bar{t})] = E(Y|T_1 = t_1, M_2 = 0, T_2 = t_2)P(M_2 = 0|T_1 = t_1) \\ +E(Y|T_1 = t_1, M_2 = 1, T_2 = t_2)P(M_2 = 1|T_1 = t_1). \end{aligned}$$

Let us define that $E(Y|T_1 = t_1, M_2 = m_2, T_2 = t_2) = Y_{t_1,m_2,t_2}$, $I(T_1 = t_1, M_2 = m_2) = I(t_1, m_2)$ and $I(T_1 = t_1) = I(t_1)$. Then using the G-formula, we can calculate the four means under each of the regimes are

$$\begin{aligned}
E[Y(\bar{t}=\{0,0\})] &= Y(000)\times\sum I(0,0)/\sum I(0)+Y(010)\times\sum I(0,1)/\sum I(0),\\
&= 3\times 100/500+2\times 400/500=2.2,\\
E[Y(\bar{t}=\{0,1\})] &= Y(001)\times\sum I(0,0)/\sum I(0)+Y(011)\times\sum I(0,1)/\sum I(0),\\
&= 8\times 100/500+7\times 400/500=7.2,\\
E[Y(\bar{t}=\{1,0\})] &= Y(100)\times\sum I(1,0)/\sum I(1)+Y(110)\times\sum I(1,1)/\sum I(1),\\
&= 6\times 350/500+4\times 150/500=5.4,\\
E[Y(\bar{t}=\{1,1\})] &= Y(101)\times\sum I(1,1)/\sum I(1)+Y(111)\times\sum I(1,1)/\sum I(1),\\
&= 5\times 350/500+1\times 150/500=3.8.
\end{aligned}$$

We conclude that, if one did not know about G-methods, a natural attempt to estimate, for example; $E[Y(\bar{t}=\{1,1\})]$ from the data in Table 5.1 would be to calculate $E(Y|T_1=1,T_2=1)$. This gives

$$E(Y|T_1=1,T_2=1)=\frac{1}{85}(5\times 70+1\times 15)=4.29.$$

Because this analysis fails to adjust for the confounder M of T's effect on Y, the mean 4.29 is non-causal and biased as an estimate of the mean contrast 3.8.

We can note directly the maximum mean response at the second time decisions to choose the optimal ($T_2=0$ or $T_2=1$) for each one of the four possible paths $(T_1,M_2)=\{(0,0),(0,1),(1,0),(1,1)\}$, then the optimal regimes of T_2 for the previous set (T_1,M_2) are equal respectively to $\{1,1,0,0\}$. Then choose the optimal action at the first time point $T_1^{opt}=0$. Note the wrong actions at the second time lead to regrets of 5,5,1 and 3 depending on the earlier sequence (T_1,M_2). The mean response if the optimal regime is always followed is $E(Y(\bar{t}=\{0,1\})=7.2$. Choosing the wrong action at the first decision time costs $E[Y(\bar{t} = \{0, 1\})] - E[Y(\bar{t} = \{1, 0\})] = 1.8$.

IPTW

We now describe how to use IPTW for estimating the counterfactual means $E[Y(\bar{t})]$ under the four regimes $\bar{t} = \{t_1, t_2\}$. The first step is to create a stabilized pseudo population by weighting the subjects in each row in Table 5.1 by the stabilized weights using the IPTW that

$$SW = \frac{P(T_1 = t_1)P(T_2 = t_2|T_1 = t_1)}{P(T_1 = t_1|M_1 = m_1)P(T_2 = t_2|M_2 = m_2, T_1 = t_1)}.$$

Because all subjects are assumed to start in the same state M_1, the factor $P(T_1 = t_1)$ cancels, because in our study the potential confounder M_1 is absent. So the formula will be as follows

$$SW = \frac{P(T_2 = t_2|T_1 = t_1)}{P(T_2 = t_2|M_2 = m_2, T_1 = t_1)}$$

or by unstabilized weights, the formula is

$$W = \frac{1}{P(T_1 = t_1|M_1 = m_1)P(T_2 = t_2|M_2 = m_2, T_1 = t_1)}.$$

Then using one of these formulas we can obtain the pseudo-population, then estimate directly the means under each of the regimes.

For example, for the first row:

$$f(T_2|T_1) = P(T_2 = 0|T_1 = 0) = 170/500 = 0.34$$

and

$$f(T_2|T_1, M_2) = P(T_2 = 0|T_1 = 0, M_2 = 0) = 50/100 = 0.5$$

Each of the 50 subjects in the first row therefore receives the weight $SW_{000} = 0.34/0.5 = 0.68$. Hence, the row has $0.68 \times 50 = 34$ subjects in the stabilized pseudo-population. This is in column $N_{ps(SW)}$ in Table 5.1. The other terms in the column are calculated similarly.

T_1	M_2	T_2	N	$E(Y\|M_2,\bar{T}_2)$	$f(T_2\|T_1)$	$f(T_2\|M_2,T_1)$	SW	$N_{ps(SW)}$	W	$N_{ps(W)}$
0	0	0	50	3	0.34	0.5	0.68	34	4	200
0	0	1	50	8	0.66	0.5	1.32	66	4	200
0	1	0	120	2	0.34	0.3	1.13	136	100/15	800
0	1	1	280	7	0.66	0.7	0.94	264	100/15	800
1	0	0	280	6	0.83	0.8	1.04	290.5	2.5	700
1	0	1	70	5	0.17	0.2	0.85	59.5	10	700
1	1	0	135	4	0.83	0.9	0.92	124.5	100/45	300
1	1	1	15	1	0.17	0.1	1.7	25.5	20	300

Table 5.1: Data corresponding to the example to explain how to use IPTW for estimating $E[Y_{\bar{t}}]$.

Also each of the subjects in the first row can be received the weight $W_{000} = \frac{1}{(0.5)(0.5)} = 4$. Hence, the row has $4 \times 50 = 200$ subjects in the unstabilized pseudo-population. The IPTW weights cancel the arrow between M_2 and T_1 in the pseudo-population as shown in Figure 2.2a. The absence of the arrow can be easily confirmed by checking whether $T_2 \perp M_2 | T1$, where $\perp$ represents independence in the pseudo-population. For example, this conditional independence holds in the stabilized pseudo-population of our example because

$$P_{ps}(T_2 = 1|T_1 = 0, M_2 = 0) = \frac{66}{100} = P_{ps}(T_2 = 1|T_1 = 0, M_2 = 1) = \frac{264}{400} = 0.66,$$

and

$$P_{ps}(T_2 = 1|T_1 = 1, M_2 = 0) = \frac{59.5}{350} = P_{ps}(T_2 = 1|T_1 = 1, M_2 = 1) = \frac{25.5}{150} = 0.17.$$

So the causal DAG corresponding to the pseudo-population does not have the arrow M_2 to T_2.

Using the IPTW stabilized or unstabilized pseudo-population formula, the four means under each of the regimes are

$$
\begin{aligned}
E_{ps(SW)}[Y(\bar{t}=\{0,0\})] &= \frac{3\times 34 + 2\times 136}{170} = 2.2,\\
E_{ps(SW)}[Y(\bar{t}=\{0,1\})] &= \frac{8\times 66 + 7\times 264}{330} = 7.2,\\
E_{ps(SW)}[Y(\bar{t}=\{1,0\})] &= \frac{6\times 290.5 + 5\times 124.5}{415} = 5.4,\\
E_{ps(SW)}[Y(\bar{t}=\{1,1\})] &= \frac{4\times 59.5 + 1\times 25.5}{85} = 3.8,
\end{aligned}
$$

or

$$
\begin{aligned}
E_{ps(W)}[Y(\bar{t}=\{0,0\})] &= \frac{3\times 200 + 2\times 800}{1000} = 2.2,\\
E_{ps(W)}[Y(\bar{t}=\{0,1\})] &= \frac{8\times 200 + 7\times 800}{1000} = 7.2,\\
E_{ps(W)}[Y(\bar{t}=\{1,0\})] &= \frac{6\times 700 + 5\times 300}{1000} = 5.4,\\
E_{ps(W)}[Y(\bar{t}=\{1,1\})] &= \frac{4\times 700 + 1\times 300}{1000} = 3.8.
\end{aligned}
$$

The following table shows the average values of the outcome $E_{ps}(Y|T_1, M_2, T_2)$ of the four static regimes, using the both of the stabilized and the unstabilized pseudo-population As

T_1	T_2	$N_{ps(SW)}$	$N_{ps(W)}$	$E_{ps}(Y\|T_1, T_2)$
0	0	170	1000	2.2
0	1	330	1000	7.2
1	0	415	1000	5.4
1	1	85	1000	3.8

Table 5.2: The stabilized and the unstabilized pseudo-population for the average value of the outcome $E_{ps}(Y|T_1, M_2, T_2)$.

expected, the values of $E_{ps}(Y|T_1, T_2)$ obtained by IPTW, in the pseudo-population, are equal to those obtained by the G-formula. In this example, we do not need to use models to estimate the inverse probability weights because we can be easily calculated by hand from the data. Also, we do not need models for the counterfactual means $E[Y(\bar{t})]$ because these means can also be calculated by hand. Let us consider the marginal structural mean model,

$$E[Y(\bar{t})] = \gamma_0 + \gamma_1 t_1 + \gamma_2 t_2 + \gamma_3 t_1 t_2$$

The model is referred to as a marginal structural mean model (MSM) because it models the marginal mean of the counterfactuals Y_t and models for counterfactuals are often referred to as structural models, Hernán at al (2000). If they simply fit a model for $E(Y|T_1, T_2)$ instead of using a marginal structural model for potential outcomes, to calculate the parameters $\gamma_0, \gamma_1, \gamma_2$ and γ_3. We obtain,

$$E(Y|T_1, T_2) = 2.2 + 2.5T_1 + 5T_2 - 7T_1T_2,$$

this gives biased estimates of $\gamma's$ because of confounding by M_2. Now we can use the pseudo-population data because

$$\begin{aligned} E_{ps(W)}[Y(\bar{t} = \{0,0\})] &= \gamma_0, \\ E_{ps(W)}[Y(\bar{t} = \{1,0\})] &= \gamma_0 + \gamma_1, \\ E_{ps(W)}[Y(\bar{t} = \{0,1\})] &= \gamma_0 + \gamma_2, \\ E_{ps(W)}[Y(\bar{t} = \{1,1\})] &= \gamma_0 + \gamma_1 + \gamma_2 + \gamma_3. \end{aligned}$$

and therefore, using the estimates for $E_{ps}(Y|T_1, T_2)$ in Table 5.2, $\gamma_0 = 2.2$, $\gamma_1 = 5.4 - 2.2 = 3.2$, $\gamma_2 = 7.2 - 2.2 = 5$ and $\gamma_3 = 3.8 - 2.2 - 5 - 3.2 = -6.6$.

This estimation procedure is equivalent to fitting a linear model with each subject weighted by SW as follows

$$E_{ps}(Y|T_1, T_2) = 2.2 + 3.2T_1 + 5T_2 - 6.6T_1T_2$$

5.4 Marginal Structural Models for Optimal Static Regimes

Robins, (1994) describe how to use marginal structural models (MSMs) for estimating the effect of treatment on the counterfactual outcomes. In this section we aim to show how to estimate the parameters using the marginal structural models comparing with the IPTW and the regret-regression methods.

$$\begin{aligned} Y(t_1, 0) &= Y(0,0) + \beta_1 t_1 \\ Y(t_1, t_2) &= Y(t_1, 0) + \beta_{21} t_2 + \beta_{22} t_2 M_2(t_1) + \beta_{23} t_1 t_2 + \beta_{24} t_1 M_2(t_1) t_2. \end{aligned}$$

T_1	M_2	T_2	N	$E(Y\|\bar{M}_2, \bar{T}_2)$	$Y(t_1, 0)$	$Y(0,0)$
0	0	0	50	3	3	3
0	0	1	50	8	$8 - \beta_{21}$	$8 - \beta_{21}$
0	1	0	120	2	2	2
0	1	1	280	7	$7 - \beta_{21} - \beta_{22}$	$7 - \beta_{21} - \beta_{22}$
1	0	0	280	6	6	$6 - \beta_1$
1	0	1	70	5	$5 - \beta_{21} - \beta_{23}$	$5 - \beta_1 - \beta_{21} - \beta_{23}$
1	1	0	135	4	4	$4 - \beta_1$
1	1	1	15	1	$1 - \beta_{21} - \beta_{22}$ $-\beta_{23} - \beta_{24}$	$1 - \beta_1 - \beta_{21}$ $-\beta_{22} - \beta_{23} - \beta_{24}$

Table 5.3: Data corresponding to the example to explain how to use IPTW for estimating $E[Y(\bar{t})]$.

Let us assume a structural model for our example. But we will now include M_2 in the model.

The model has one equation for each time point with one unknown parameter β_1 in the time point 1 equation and a vector β_2 of 4 unknown parameters in the second time point equation. By solving the first equation at $t_1 = 1$, then the parameter $\beta_1 = Y(1,0) - Y(0,0)$ represents the subject the effect of treatment t_1 on the outcome when treatment t_2 is set to zero.

The 4 parameters β_2 in the second equation parameterize the effect of t_2 on Y within the 4 possible levels of past treatment and covariate history. For example $\beta_{21}, \beta_{22}, \beta_{23}, \beta_{24}$ are respectively the effect of t_2, the interactions between each of $(t_2, M_2(t_1))$, (t_1, t_2) and $(t_1, M_2(t_1), t_2)$ on Y. Also β_{21} and $\beta_{21} + \beta_{22}$ are, respectively, the effect of t_2 on Y when $t_1 = 0$ with $M_2(t_1 = 0) = 0$ and the effect of t_2 on Y when $t_1 = 0$ with $M_2(t_1 = 0) = 1$. Similarly $\beta_{21} + \beta_{23}$ and $\beta_{21} + \beta_{22} + \beta_{23} + \beta_{24}$ are, the effect of t_2 on Y when $t_1 = 1$ with $M_2(t_1 = 1) = 0$ and with $M_2(t_1 = 1) = 1$, respectively.

Returning to our example, we begin by estimating the parameter vector β_2. To do so, in Table 5.3, we first use the structural model

$$Y(t_1, t_2) = Y(t_1, 0) + \beta_{21} t_2 + \beta_{22} t_2 M_2(t_1) + \beta_{23} t_1 t_2 + \beta_{24} t_1 M_2(t_1) t_2,$$

to calculate the mean of $Y(t_1, 0)$ in terms of the unknown parameter vector β_2. To help understand these calculations, we see in the second data row of Table 5.3, the expression $8 - \beta_{21}$ for the mean of $Y(t_1, 0) = Y(0,0)$ among subjects with $t_1 = 0$, $M_2 = 0$, $t_2 = 1$. By solving the structural model equation we find the other expressions of $Y(t_1, 0)$ and $Y(0,0)$. To estimate β_2 we consider first the stratum $(T_1, M_2(t_1)) = (0,0)$. From data rows 1 and 2 in the Table 5.3, we find that the mean when $T_2 = 0$ is 3 and $8 - \beta_{21}$ when $T_2 = 1$. Hence $8 - \beta_{21} = 3$.

Next we equate the means of $Y(t_1, 0)$ in data rows 3 and 4 corresponding to stratum $(T_1, M_2) = (0,1)$ to $7 - \beta_{21} - \beta_{22}$. Since $\beta_{21} = 5$, we conclude $\beta_{22} = 0$. Continuing we equate the means of $Y(t_1, 0)$ in data rows 5 and 6 to obtain $5 - \beta_{21} - \beta_{23}$. Since $\beta_{21} = 5$, we conclude $\beta_{23} = -6$. Finally, equating the means of $Y(t_1, 0)$ in data rows 7 and 8 to obtain $1 - \beta_{21} - \beta_{22} - \beta_{23} - \beta_{24}$. Since $\beta_{21} = 5$, $\beta_{22} = 0$, $\beta_{23} = -6$, we conclude $\beta_{24} = -2$. To estimate β_1, as shown above $\beta_1 = Y(1, 0) - Y(0, 0)$, so $\beta_1 = E[Y(1, 0)] - E[Y(0, 0)]$. Thus

$$\beta_1 = [6\frac{350}{500} + 4\frac{150}{500}] - [3\frac{100}{500} + 2\frac{400}{500}] = 5.4 - 2.2 = 3.2.$$

Now we can estimate the causal effect of t_1 and t_2 on Y,

$$Y(t_1, 0) = Y(0, 0) + \beta_1 t_1,$$

Thus $E[Y(1, 0)] = E[Y(0, 0)] + \beta_1 = 2.2 + 3.2 = 5.4$. Then using the second model,

$$Y(t_1, t_2) = Y(t_1, 0) + \beta_{21}t_2 + \beta_{22}t_2M_2(t_1) + \beta_{23}t_1t_2 + \beta_{24}t_1M_2(t_1)t_2,$$

we can estimate $E[Y(0, 1)]$, and $E[Y(1, 1)]$
$E[Y(0, 1)] = E[Y(0, 0) + \beta_{21} + \beta_{22}M_2(0)] = 2.2 + 5 + 0E[M_2(0)] = 7.2$,
$E[Y(1, 1)] = E[Y(0, 0)+\beta_1+\beta_{21}+\beta_{23}+(\beta_{22}\beta_{24})M_2(1)] = 2.2+3.2+5-6+(0-2)(0.3) = 3.8$.
Recall the fitting model using IPTW

$$E_{ps}(Y|T_1, T_2] = 2.2 + 3.2T_1 + 5T_2 - 6.6T_1T_2,$$

All of these estimated values of the model parameters agree with those obtained by the G-formula and by IPTW. The value -6.6 is the causal effect of the interaction between t_1 and t_2 on Y. We obtain this value using the previous results by subtracting $E[Y(1, 0)]-E[Y(0, 0)]$ which is the effect of t_1 out of $E[Y(1, 1)] - E[Y(0, 0)]$, the total treatments effect as follows

$$(3.8 - 2.2) - 5 - 3.2 = -6.6$$

.

5.5 Comparing the Regret-regression Method with the Inverse Probability of Treatment Weighting Method

As shown in Chapter 4, the regret-regression method has the ability to estimate an optimal dynamic treatment regime by choosing a fully parameterised model for the mean. The method estimates the effect of treatments on the potential final responses. Basically it requires modelling of $E[M_j|\bar{M}_{j-1}, \bar{T}_{j-1}]$, to remove the effect of the confounders (the intermediate states $M_2, \cdots, M_K$) on those outcomes. Note that G-formula (see page) does not need $E[T_j|\bar{M}_j, \bar{T}_{j-1}]$ but instead also uses $E[M_j|\bar{M}_{j-1}, \bar{T}_{j-1}]$, that is why it is not considered in this section. Instead of that the inverse probability of treatment weighting needs modelling of $E[T_j|\bar{M}_j, \bar{T}_{j-1}]$, which is needed as well for Murphy and Robins methods.

Recall equation 4.3 in Chapter 4 Section 2

$$E[Y|\bar{M}_K, \bar{T}_K] = \beta_0(M_1) + \sum_{j=2}^{K} \beta_j(\bar{M}_{j-1}, \bar{T}_{j-1}) Z_j - \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}).$$

Inclusion of the linear combination of residuals Z_j in the formula is reminiscent of a path analysis method for dealing with time-dependent confounders. In this section we will compare the regret-regression method with the inverse probability of treatment weighting to observe any difference or similarity between them. The regret-regression coefficients below for a regime $\{T_1 = t_1, T_2 = t_2\}$ are classified as

- Const: denotes the expected final response $E[Y(t_1, t_2)]$ when following a specific regime $d = \{t_1, t_2\}$.
- $\mu_1 I(t_1)$: denotes a regret value if we follow a different regime of t_1 then a regime of t_2.

- $Z_2(T_1)I(T_1)$: be the residual between M_2 and its expected value given T_1. Suppose $Z_2(T_1)I(T_1) = a$ then $a \times (1 - p(M_2 = 1))$ is the loss for going (randomly) to a wrong path of M_2 which does not contain $\max Y(T_1)$.

- $\mu_2 I(T_1, M_2, T_2 \neq t_2)$ are regret values when following the other regime of T_2 (e.g., if $t_2 = 0$ then the other regime is $t_2 = 1$).

The regret-regression estimates of the counterfactual means $E[Y_{\bar{t}=\{0,0\}}]$, $E[Y_{\bar{t}=\{1,0\}}]$, $E[Y_{\bar{t}=\{0,1\}}]$ and $E[Y_{\bar{t}=\{1,1\}}]$ under the four possible static regimes are

First regime d=$\{$0, 0$\}$

Const	$\mu_1 I(1)$	$Z_2(0)I(0)$	$Z_2(1)I(1)$	$\mu_2 I(001)$	$\mu_2 I(011)$	$\mu_2 I(101)$	$\mu_2 I(111)$
2.2	3.2	-1	-2	5	5	-1	-3

Second regime d=$\{$0, 1$\}$

Const	$\mu_1 I(1)$	$Z_2(0)I(0)$	$Z_2(1)I(1)$	$\mu_2 I(000)$	$\mu_2 I(010)$	$\mu_2 I(101)$	$\mu_2 I(111)$
7.2	-3.4	-1	-2	-5	-5	1	3

Third regime d=$\{$1, 0$\}$

Const	$\mu_1 I(0)$	$Z_2(0)I(0)$	$Z_2(1)I(1)$	$\mu_2 I(001)$	$\mu_2 I(011)$	$\mu_2 I(101)$	$\mu_2 I(111)$
5.4	-3.2	-1	-2	5	5	-1	-3

Forth regime d=$\{$1, 1$\}$

Const	$\mu_1 I(0)$	$Z_2(0)I(0)$	$Z_2(1)I(1)$	$\mu_2 I(000)$	$\mu_2 I(010)$	$\mu_2 I(100)$	$\mu_2 I(110)$
3.8	3.4	-1	-4	-5	-5	1	3

According to these results, we can read off directly the mean response and different regrets. For example, if regime $d = \{0, 0\}$ is always followed the mean is 2.2. Choosing the other action at the first decision time (which is the optimal action) increases 3.2 in mean. The other actions at the second time lead to regrets of -5,-5,1 and 3 depending on the earlier sequence (T_1, M_2). They increase 5 in mean when $T_2(t_1 = 0) = \{1, 1\}$ and decrease $\{1, 3\}$ in mean when $T_2(t_1 = 1) = \{1, 1\}$ because the first set of T_2 are the optimal actions and the second are not. The other two terms measure the effects of M_2 after allowing for the effect of T_1. For each value of T_1 in this example having $M_1 = 1$ is associated with a decrease in mean: by $1 \times (1 - 0.8)$ if $T_1 = 0$ (2 instead of 2.2) and by $2 \times (1 - 0.3)$ if $T_1 = 1$ (4 instead of 5.4), in both cases assuming the $T_2 = 0$ is chosen. Recall the optimal regimes

Const	$\mu_1 I(1)$	$Z_2(0)I(0)$	$Z_2(1)I(1)$	$\mu_2 I(000)$	$\mu_2 I(010)$	$\mu_2 I(101)$	$\mu_2 I(111)$
7.2	-1.8	-1	-2	-5	-5	-1	-3

which are $T_1^{opt} = 0$, $T_2^{opt}(T_1 = 0, M_2) = 1$ and $T_2^{opt}(T_1 = 1, M_2) = 0$. As seen when following optimal regimes all regrets coefficients must be negative in the regret-regression model. We can use regret-regression method for the other four different regimes. The only difference is regret coefficients might be positive in the regret-regression model (when one or more regimes are not optimal). The second regime $d = \{0, 1\}$ gives the value of the maximum final response because at second time point it follows the optimal action, which is $t_2(t_1 = 0, M_2) = 1$ but follows the action $t_2(t_1 = 1, M_2) = 1$ which is not the optimal action. That is why its regret at time one of 3.4 $(7.2 - 3.8)$ is bigger than 1.8 $(7.2 - 5.4)$ when using optimal regimes. The previous results for the four regimes are exactly the same with those we got in Table 5.2 by using inverse probability of treatment weighting which were 2.2, 7.2, 5.4 and 3.8. The following are simulations results on estimation of the optimal final response. For each simulation we use 100 datasets of samples size 100 and 1000.

As we see inverse probability of treatment weighting and regret-regression give exactly the same results. Both of the regret-regression method and the inverse probability of treatment weighting remove the effect of the confounder's covariates. However, an important

Sample size	SD	Regret-regression		IPTW	
		Mean optimal response	SE	Mean optimal response	SE
100	0	7.2	0.06	7.2	0.06
	0.25	7.2	0.11	7.2	0.11
	0.5	7.19	0.13	7.19	0.13
	1	7.18	0.22	7.18	0.22
1000	0	7.2	0.02	7.2	0.02
	0.25	7.2	0.02	7.2	0.02
	0.5	7.2	0.04	7.2	0.04
	1	7.2	0.06	7.2	0.06

Table 5.4: Comparing the regret-regression and the inverse probability of treatment weighting methods for estimating optimal regimes using 2 time points

advantage of the regret-regression is to use a unique model. Furthermore, problems arise when using inverse probability of treatment weighting in cases of sample of modest size or when there is need to estimate high dimensional parameters. Appendix 9.2 shows that $\hat{\mu}_1$ (the estimated regrets at time 1) as obtained from the regret-regression versus the inverse probability of treatment weighting are identical in the previous example of two time point situation and binary states and actions.

5.6 The G-formula and the IPTW for Finding Optimal Dynamic Strategies

As seen we are able to use the regret-regression method for finding optimal dynamic regimes using the regret idea by choosing the optimal actions of T_2 which are depend on the value of M_2. In this section we aim to use the G-formula or the IPTW for finding optimal dynamic treatment strategies. We need to change our worked example, to a similar one that optimal decisions T_2 depend on the value of M_2. Let us assume that

$$\{E(Y_{000}), E(Y_{001}), E(Y_{100}), E(Y_{101})\} = \{8, 3, 5, 6\},$$

instead of the values $\{3, 8, 6, 5\}$. Hence choosing optimal actions of T_2 depend on the values of M_2 as follows

T_1	M_2	T_2	N	$E(Y\|\bar{M_2}, \bar{T_2})$	$f(T_1)$	$f(T_2\|M_2, T_1)$	W	$N_{ps(W)}$
0	0	0	50	8	0.5	0.5	4	200
0	0	1	50	3	0.5	0.5	4	200
0	1	0	120	2	0.5	0.3	100/15	800
0	1	1	280	7	0.5	0.7	100/15	800
1	0	0	280	5	0.5	0.8	2.5	700
1	0	1	70	6	0.5	0.2	10	700
1	1	0	135	4	0.5	0.9	100/45	300
1	1	1	15	1	0.5	0.1	20	300

Table 5.5: Data corresponding to the example to explain how to use G-formula and IPTW for estimating optimal dynamic actions.

Table 5.4 shows that optimal strategy depends on M_2, e.g., when $T_1 = 0$ if $M_2 = 0$ then $T_2^{opt} = 0$ and if $M_2 = 1$ then $T_2^{opt} = 1$. As time 2 is the final time point, we can choose optimal T_2 directly. As discussed before the regret-regression method can be used to find optimal dynamic treatment regimes at time 2 by try and error based on the regrets idea.

In this case the regret-regression coefficients are

Const	$\mu_1 I(1)$	$Z_2(0)I(0)$	$Z_2(1)I(1)$	$\mu_2 I(001)$	$\mu_2 I(010)$	$\mu_2 I(100)$	$\mu_2 I(111)$
7.2	-1.8	-1	-2	-5	-5	-1	-3

When following optimal regimes all regrets coefficients must be negative in the regret-regression model. Hence optimal roles using the regret-regression method are $T_1^{opt} = 0$, $T_2^{opt}(T_1 = 0, M_2 = 0) = 0$, $T_2^{opt}(T_1 = 0, M_2 = 1) = 1$, $T_2^{opt}(T_1 = 1, M_2 = 0) = 1$, $T_2^{opt}(T_1 = 1, M_2 = 1) = 0$.

After determine the optimal dynamic regimes directly at the second time point, which maximize $E(Y|T_1, M_2, T_2)$. The optimal dynamic policies for the first time point can be achieved by comparing the G-formula of each $E[Y(T_2^{opt})|T_1, M_2]$

$$\begin{aligned} E[Y(T_2^{opt})|T_1 = 0, M_2] &= E(Y_{000}) \times \sum I(0,0)/\sum I(0) + Y_{011} \times \sum I(0,1)/\sum I(0), \\ &= 8 \times 100/500 + 7 \times 400/500 = 7.2, \\ E[Y(T_2^{opt})|T_1 = 1, M_2] &= Y_{101} \times \sum I(1,0)/\sum I(1) + Y_{110} \times \sum I(1,1)/\sum I(1), \\ &= 6 \times 350/500 + 4 \times 150/500 = 5.4. \end{aligned}$$

To find IPTW policy, we need to compare $E_{ps(W)}[Y(T_2^{opt})|T_1, M_2]$ of unstabilized pseudo-population formula. From table 5.3

$$\begin{aligned} E_{ps(W)}[Y(T_2^{opt})|T_1 = 0, M_2] &= (8 \times 200 + 7 \times 800)/1000 = 7.2, \\ E_{ps(W)}[Y(T_2^{opt})|T_1 = 1, M_2] &= (6 \times 700 + 5 \times 300)/1000 = 5.4. \end{aligned}$$

Thus $T_1^{opt} = 0$. Hence the optimal dynamic roles using all the three methods are the same. But the regret-regression method avoids problems which arise when using samples of modest size or when there is need to estimate high-dimensional parameters.

Chapter 6

Regret-regression and Multi-armed Bandit Problem

6.1 Introduction

The multi-armed bandit problem, described by Robbins (1952), is a simple machine learning problem based on the idea of a traditional slot machine (one-armed bandit) but with more than one arm. When pulled, each arm provides a reward drawn from a distribution corresponding to that specific arm. The problem is a classical example of the trade-off between exploration and exploitation. On the one hand, if the gambler plays exclusively on the machine that he thinks is best (exploitation), he may fail to discover that one of the other arms actually has a higher average return. On the other hand, if he spends too much time trying out all the machines and gathering statistics (exploration), he may fail to play the best arm often enough to get a high total return.

The classical example of a bandit problem is deciding what treatment to give each patient in a clinical trial when the effectiveness of the treatments are initially unknown and the patients arrive sequentially (Thompson 1933). The decision-maker is an experimenter who allocates one of K experimental treatments sequentially to a sample of patients. The patient attends at the clinic at the j^{th} time point, and only one of the K treatments could be used on the patient. Outcomes accumulate as the trial continues. The objective is to treat patients to maximize their final expected response. One example concerns patients with prostate cancer where they need to balance the possibility of a long life expectation against possible stressful treatment side-effects (patient trade-off). Each choice of an arm returns in an immediate random return, but the process determining these returns evolves during the play of the bandit. The characterizing feature of bandit problems is that the distribution of returns from one arm only changes when that arm is chosen. Hence the rewards from an arm do not depend on the rewards obtained from other arms. This feature also assumes that the distributions of returns do not depend clearly on calendar time.

Dynamic programming or backward induction is the standard method for finding the optimal strategies. Other methods, as the myopic rule or stay with a winner (sometimes called play the winner) rule are used more commonly, although these two methods do not always give the optimal result. The myopic rule is to choose the arm with greater expected gain at each stage. In the two-armed Bernoulli bandit, Feldman (1962) showed that myopic strategies are optimal when the number of patients is known and the success probability of the two treatments has a two-point prior distribution. However, myopic strategies are not optimal, or even good, generally. The play the winner rule is to use the same arm if the last observation from this arm yielded a success at the last stage. The play the winner rule is optimal if two arms are independent and if the arm used at the last stage is optimal. For more details of strategies, see Berry and Fristodt (1985).

Other than sequential selection for every observation, separating a medical trial into several stages is a more realistic way of solving the two-armed bandit problem. Since the data can be collected at intervals throughout the trial, there is no need to know the result of previous patients before giving the next patient treatment, and the calculations can thus be simplified. Canner (1970) discussed the binary two-armed bandit, in two-stage setting with a sampling cost for the first stage patient, but the same number of patients are assigned to both treatments. Witmer (1986) also discussed the Bernoulli two-armed bandit in a multi-stage setting but with one treatment known. Clayton and Witmer (1988) considered the Bernoulli one-armed bandit in a two-stage setting, with the successes in the first stage discounted by a factor. Gittins and Wang (1992) showed that if two arms have the same prior mean and only one arm is to be used for all patients, the arm with greater prior variance shall yield greater outcomes.

6.2 The Multi-Armed Bandit Problem

The multi-armed bandit problem remained unsolved for many years. Only through the use of dynamic programming could a solution be found and then only for small problems. The disadvantage of this method is that such formulations are often too general to exploit the special structure of the problem at hand and they are extremely computationally expensive for problems of reasonable size. However in some cases structural results have been found to lead to efficient solution procedures.

In series of papers, Gittins proposed a method of solution. This method gives a calibrating index to each of the competing arms, namely the Gittins index. Each index depends only upon the current state of the corresponding arm. At each decision time $j = 0, 1, 2, \cdots, K$ a decision must be taken as to which of the T arms will be selected for processing and the optimal policy selects that arm with the current largest index. If arm $t \in T$ is chosen at time j then a discounted reward of

$$\lambda^j R\{M_j(t)\},$$

is gained, where $\lambda \in [0, 1)$ is a discount rate, R is a reward function defined on the state space of arm t and $M_j(t)$ is the value of a Markov chain modelling the evolution of arm t at time j. After a unit of time dedicated to arm t, it changes state according to a Markov law of motion P_t. The states of the other arms remain unchanged. The objective is to find a policy for allocating arms for processing that maximizes the total expected discounted reward over an infinite horizon.

The multi-armed bandit problem is a discounted Markov decision process denoted

$$\{(\Omega_t, P_t, R[M(t)], \lambda) : 1 \leq t \leq T\}$$

with the following special features:

1- At each decision time point $j = 1, 2, \cdots, K$ an action from the set T is taken. Here action $T_j = t_j$ taken at time j is interpreted as selecting arm t during the interval $[j, j + 1)$.

2- A time homogeneous Markov chain $\{M_j(t) : j = 1, 2, \cdots, K\}$ with a countable state space Ω_t is used to model the evolution of each arm. Note that, the state of the system at any time j may be written as,

$$M_j(t) = \{M_j(1), \cdots, M_j(T)\}.$$

3- A transition matrix P_t is such that if action t is taken then arm t in state m subsequently enters state $\widetilde{m}$ with probability

$P_t[(m, \widetilde{m})], (m, \widetilde{m}) \in \Omega_t \times \Omega_t$, where Ω_t is the set of possible states. Further, arms which are not selected for processing remain unchanged in state. That is, if arm t is chosen at time j then

$$p\{M_q(j + 1) = M_q(j)\} = 1 \quad \text{for} \quad q \neq t,$$

and

$$p[M_{j+1}(t) = \widetilde{m} | M_j(t) = m\} = P_t(m, \widetilde{m}).$$

4- A bounded n-dimensional reward function $R :\mapsto R^+$; for n states; $M_j(t) = \{M_{1j}(1), \cdots, M_{nj}(T)\}$ and a discount factor $\lambda \in [0, 1)$ are such that if arm t is taken at time j then an immediate reward of $\lambda^j R[M_j(t)]$ is earned. Hence, the total expected reward corresponding to policy d, may be expressed as,

$$R(d, M) = E_d \left[\sum_{j \geq 0} \lambda^j R\{M_j(d)\} | M_0(d) \right].$$

Here E_d represents the expectation taken over all realizations of the process under policy d, where $d \in T$ and d_j represents the choice policy d makes at time j and $M_0(d)$ is an initial state of arm t under policy d.

5- The policy d is any rule for choosing actions at each decision time point in terms of the history of the process to date. An optimal policy d^{opt} is any rule that maximises the total expected reward, i.e.

$$R(d^{opt}, M) = \max_d R(d, M).$$

Here the maximum is over all policies. It is known that there exists an optimal policy for a discounted Markov decision process that is deterministic, stationary and Markov (Ross (1970)).

In Chapter 2, Section 1, we described principles of dynamic programming. For particular problems the use of dynamic programming and the application of Bellman's principle of optimality (Bellman, 1957) would allow these classical problems to be solved. However, as the size of the problem increases, serious computational difficulties arise, as discussed earlier. Additionally, no insight into the structure of the optimal policy is obtained.

In the next sections we wish to maximize the expected value of the sum of non-discounted rewards ($\lambda = 1$) in sections from 6.3 up to 6.6 or discounted rewards ($0 < \lambda < 1$) in all sections from 6.3 up to the end of this chapter. In both, rewards achieved during situations of finite states; e.g., for arm t at time j, it might there are n states; $M_j(t) = \{M_{1j}(t), M_{2j}(t), \cdots, M_{nj}(t)\}$. Infinite states, e.g. for arm t at time j which have unlimited states; $M_j(t) = \{M_{1j}(t), M_{2j}(t), \cdots\}$, are not considered here. Also, we will use both situations of a finite horizon and an infinite horizon where in the first we face K time points, $j = \{1, \cdots, K\}$ but for the infinite horizon there are unlimited time points, $j = \{1, 2, \cdots\}$.

6.3 Regrets and Optimal Dynamic Regime for the Multi-Armed Bandit Problem

In this section, we will look at the multi-armed bandit problem as a dynamic problem. The aim is to solve this problem using different approaches for finding optimal dynamic regimes. For example, consider a multi-armed bandit problem example with two arms and two states. At time j the state value $M_j(t)$ is a 2-vector $\{M_{1j}(t), M_{2j}(t)\}$, where $M_j(1) \in \{1, 2\}$ is the value of arm one and $M_j(2) \in \{1, 2\}$ of arm two. The action T_j is to choose one of the arms. Response Y is then incremented by total non-discounted rewards which depend on the values of the chosen arms. In our example the rewards are 6 or 4 for the two values of arm one, and 8 or 3 for the two values of arm two. If arm one is selected then $M_j(1)$ is updated for time $j+1$ according to a Markov chain but $M_j(2)$ remains at its previous value. The opposite happens if arm two is selected: $M_j(2)$ is updated but $M_j(1)$

is unchanged. As explained Figure 6.1 the transition matrices are

$$P_1 = \begin{pmatrix} 0.2 & 0.8 \\ \\ 0.3 & 0.7 \end{pmatrix} \qquad \text{and} \qquad P_2 = \begin{pmatrix} 0.4 & 0.6 \\ \\ 0.5 & 0.5 \end{pmatrix}.$$

This is a special case of the multi-armed bandit problem.

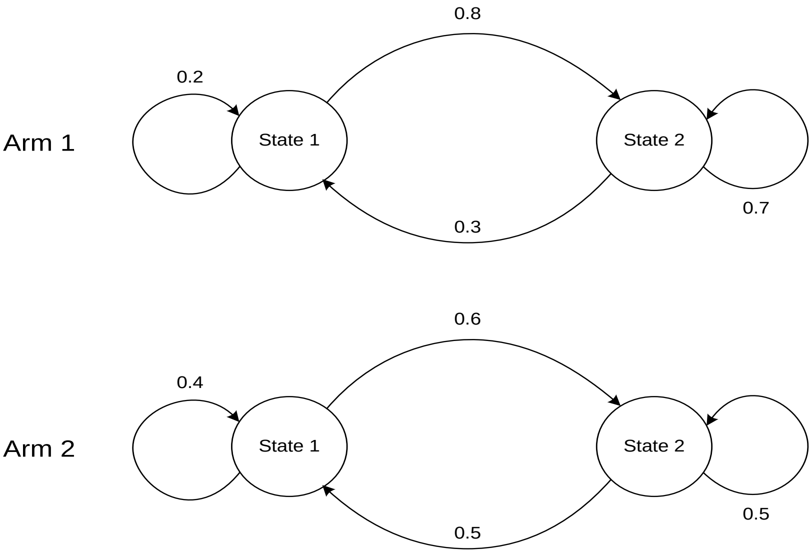

Figure 6.1: A two-armed bandit example.

For simplicity let us suppose $K = 2$ time points. Figure 6.2 shows all possible states, actions and final expected total rewards. The possible four states are $M_j = \{(1,1),(1,2),(2,1),(2,2)\}$. We use a binary action $t_j = 0$ for choosing arm 1 and $t_j = 1$ for choosing arm 2 and denote the four states of M_j as 1, 2, 3 and 4 respectively. In produce the direct approach to analysis the problem involves two stages. The first stage, regression, involves modelling the observable data. The second stage, dynamic programming (DP) or backward induction, uses the models to determine optimal actions, working iteratively from the last time stage.

In this example, for each initial state, there are eight different T_1, M_2, T_2 sequences and hence eight parameters in a saturated model for the expected total final rewards Y. Using the standard main effects and interaction formulation these are

Initial state	Const	T_1	M_2	T_2	T_1M_2	T_1T_2	M_2T_2	$T_1M_2T_2$
1	13	1	-1	1	1	6	1	-6
2	14	-5	-1	-5	1	12	1	-6
3	11	1	-1	1	1	-22	1	4
4	12	-5	-1	-5	1	24	1	-6

From this we can calculate the mean response at each of the eight $T_1M_2T_2$ sequences (for each initial state) and hence the regrets due to choices T_2 for each (T_1, M_2). Dealing with the first decision time is trickier: for each of the two values of T_1 we need to calculate

$$\sum_{M_2} E[Y|T_1, M_2, T_2^{opt}]P(M_2|T_1),$$

from which the optimal choice and regret can be found. Recall Equation 3.1, the regret function at time j is defined as

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = E(Y \mid \bar{M}_j, \bar{T}_{j-1}, \underline{d}_j^{opt}) - E(Y \mid \bar{M}_j, \bar{T}_{j-1}, t_j, \underline{d}_{j+1}^{opt}).$$

Optimal actions can be taken by working from the final time point and choosing the actions when regrets are equal to zero. To choose optimal actions at the final time point, regrets are directly calculated. At other time points $j = 1, \cdots, K-1$, we might choose optimal actions by calculation of regrets through the expectation of optimal final rewards given the history of previous states and actions. Table 6.1 shows the optimal action for each state M_j, along with the regrets for choosing a suboptimal action.

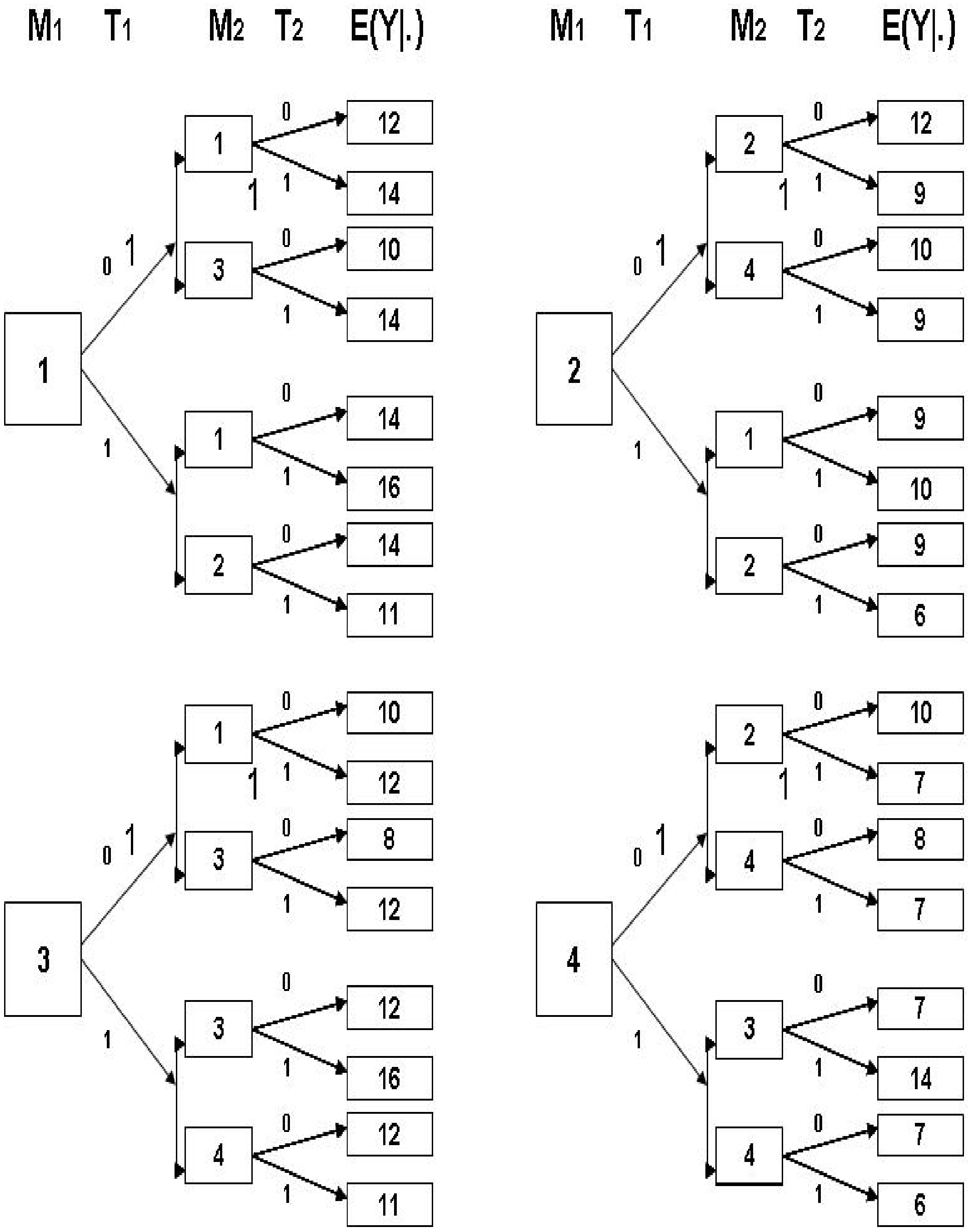

Figure 6.2: States, actions and the mean total rewards ($K = 2$)

M_1	R_1	T_1	$\max E(R_2)$	μ_1	M_2	$P(M_2\|M_1,T_1)$	R_2	T_2^{opt}	$\max E(Y\|M_1,T_1,M_2,T_2^{opt})$
1	(6,8)	0	14	0.8	1	0.2	(6,8)	1	14
					3	0.8	(4,8)	0	14
		1*	14.8	0	1	0.4	(6,8)	1	16
					2	0.6	(6,3)	0	14
2	(6,3)	0*	10.4	0	2	0.2	(6,3)	0	12
					4	0.8	(4,3)	0	10
		1	10	0.4	1	0.5	(6,8)	1	11
					2	0.5	(6,3)	0	9
3	(4,8)	0	12	1.6	1	0.3	(6,8)	1	12
					3	0.7	(4,8)	1	12
		1*	13.6	0	3	0.4	(4,8)	1	16
					4	0.6	(4,3)	0	12
4	(4,3)	0	8.6	0.4	2	0.3	(6,3)	0	10
					4	0.7	(4,3)	0	8
		1*	9	0	3	0.5	(4,8)	1	11
					4	0.5	(4,3)	0	7

$*$ *for optimal actions at first time point.*

Table 6.1: Optimal actions and total final response

6.4 Regret-Regression for the Multi-Armed Bandit Problem

In the regret-regression method, the observed Y is parameterised as a function of the regrets and of a linear combination of residuals between states M_j and their associated conditional expectations given earlier history. Recall the regret-regression formula

$$E[Y|\bar{M}_K,\bar{T}_K] = \beta_0(M_1) + \sum_{j=2}^{K}\beta_j(\bar{M}_{j-1},\bar{T}_{j-1})Z_j - \sum_{j=1}^{K}\mu_j(T_j|\bar{M}_j,\bar{T}_{j-1}), \tag{6.1}$$

where $\beta_j(\bar{M}_{j-1},\bar{T}_{j-1})$ is a vector measuring the effect of M_j for $(\bar{M}_{j-1},\bar{T}_{j-1})$ and assuming optimal actions are chosen from time j onward. Let $\mu_1 I(M_1 T_1) = \mu_1 I(M_1 = m_1, T_1 =$

$1 - t_1^{opt}$) and $\mu_2 I(M_1, T_1, M_2, T_2) = \mu_1 I(M_1, T_1, M_2, T_2 = 1 - t_2^{opt})$ are regrets of choosing the wrong decision at first and second time point, respectively. Further, let $Z_2(m_1 t_1)$ be the residual between M_2 and its expected value given $M_1 = m_1, T_1 = t_1$. So $Z_2(10) = M_2 - p_1(1,2)$, $Z_2(11) = M_2 - p_2(1,2)$, $Z_2(20) = M_2 - p_1(1,2)$, $Z_2(21) = M_2 - p_2(2,2)$, $Z_2(30) = M_2 - p_1(1,2)$, $Z_2(31) = M_2 - p_2(1,2)$, $Z_2(40) = M_2 - p_1(2,2)$ and $Z_2(41) = M_2 - p_2(2,2)$ where $p_1(1,2) = 0.8$, $p_2(1,2) = 0.6$, $p_1(2,2) = 0.7$ and $p_2(2,2) = 0.5$ all of which would need to be estimated in practice. Instead of the standard model by using regression with thirty two-parameter main effects and interaction summarised before (eight-parameter for each of four initial states), the regret-regression obtain exactly the same saturated fit using a linear model with the twenty covariates given below, along with their associated parameter values.

Const1	Const2	Const3	Const4	$\mu_1 I(10)$	$\mu_1 I(21)$	$\mu_1 I(30)$
14.8	10.4	13.6	9	-0.8	-0.4	-1.6

$\mu_1 I(40)$	$Z_2 I(10)$	$Z_2 I(11)$	$Z_2 I(20)$	$Z_2 I(21)$	$Z_2 I(30)$	$Z_2 I(31)$
-0.4	0	-2	-2	-2	0	-4

$Z_2 I(40)$	$Z_2 I(41)$	$\mu_2 I(1)$	$\mu_2 I(2)$	$\mu_2 I(3)$	$\mu_2 I(4)$
-2	-4	-2	-3	-4	-1

The expected mean responses for each initial state (1,1),(1,2),(2,1) and (2,2) when following optimal regime are 14.8, 10.4, 13.6 and 9. Costs in the mean if we choose wrong actions at the first decision time point are 0.8, 0.4, 1.6 and 0.4 in the mean. If we always choose the initial state by a random coin toss and follow the optimal regime, the overall mean will be $E(Y^{opt}) = 0.25 \times (14.8 + 10.4 + 13.6 + 9) = 11.95$. The wrong actions at the second time of states (1,1),(1,2),(2,1) and (2,2), lead to regrets of 2, 3, 4 and 1 respectively. The effects of M_2 *after* allowing for the effect of M_1 and T_1 can be measured by the other eight terms. At the state (1,1) having $t_1 = 0$ decreases the mean expected final response: by $0 \times (1 - 0.8)$ and by $2 \times (1 - 0.6)$ if $T_1 = 1$, in both cases assuming the optimal action T_2 later.

6.5 Comparing the Regret-regression with Other Methods

In the literature, the multi-armed bandit problem is posed as a statistical decision model of an agent trying to optimize his decisions while improving his information at the same time. There are several different algorithms and models which can be applied for this problem, e.g. *Linear Programming*, *Gittins Index*, Q-learning etc. These various algorithms are approximate in a situation of discount rewards and infinite horizon (time of converge when discounted reward approximates to zero), otherwise they are exact. As in standard regression models, the form of the regression model in Q-learning need to assume the same conditions (assumptions A1 to A5) of the regret-regression method which were discussed in Section 2.4.5 and Section 4.2. While in other method we assume only the assumptions A1 to A3 in Section 2.4.5 as sufficient assumptions for causal inference.

In this section we compare regret-regression with other methods. For simplicity, we only consider a two-armed bandit problem with a finite-horizon $K = 2$, a finite-state $n = 2$, and binary actions at each stage, coded (0,1), for choosing arm 1 and arm 2 respectively assigned randomly in the training data depends on probabilities of states in each arm, optimal strategies are dynamic (depend on previous states and actions) as described in Section 6.3.

First, we will have a quick view of the Q-learning and the inverse probability treatment weighting method using the chosen example in Figure 6.2 with a non-discounted reward $\lambda = 1$.

6.5.1 Q-learning method

Q-learning, is a simple incremental algorithm developed from the theory of dynamic programming (Ross,1983). In Q-learning, policies and the value function are represented by a two-dimensional lookup table indexed by state-action pairs. There are many variants of *Q-learning* (Watkins (1989), Sutton and Barto (1998), Ormoneit and Sen (2002), Lagoudakis and Parr (2003), Ernst et al. (2005), Murphy (2005)). Q-learning can be viewed as a generalization of regression to the multi-stage decision problem to compute solutions to bandit problems. In this section, we use linear regression to fit the Q-functions, and then we compare the regret-regression estimator with the Q-learning estimator via simulated experiments. Here is a simple introduction using least squares. We consider the fitted Q-learning algorithm with linear regression using least squares. Let the stage-j ($j = 1, 2$) Q-function be modeled as

$$Q_j(\bar{M}_j, \bar{T}_j; \alpha_j, \beta_j) = \alpha_j^{f}{}_0(\bar{M}_j, \bar{T}_{j-1}) + \beta_j^{f}{}_1(\bar{M}_j, \bar{T}_j)T_j$$

where $f_0(\bar{M}_j)$ and $f_1(\bar{M}_j)$ are two basis functions of the history of states and actions, with $f_1(\bar{M}_j, \bar{T}_j)$ denoting the the history information of states and actions that interact with the action T_j. We have separated these two parts because only the second term features in the policy. Thus even though we estimate all the parameters from the training data, our main interest lies in the policy parameters β_j's only (α_j's are nuisance parameters). For the 2 time point's case, define

$$\begin{aligned} Q_2(\bar{M}_2, \bar{T}_2) &= E[Y|\bar{M}_2, \bar{T}_2)] \\ Q_1(M_1, T_1) &= E[Y|M_1, T_1], \end{aligned}$$

where $Q_2(\bar{M}_2, \bar{T}_2)$ and $Q_1(M_1, T_1)$ are the Q-function at time 1 and 2, and Y is the final total reward. If the two Q-functions were known then using backwards induction (e.g. as in dynamic programming) we see that the optimal decision rules are

$$\begin{aligned}
d_2^{opt}(\bar{M}_2, \bar{T}_2) &= \max_{T_2} Q_2(\bar{M}_2, T_1) \\
d_1^{opt}(M_1, T_1) &= \max_{T_1} Q_1(M_1).
\end{aligned}$$

For Q-learning with function approximation, the time j optimal Q-function is approximated by $Q_j(\bar{M}_j, \bar{T}_j; \alpha_j, \beta_j)$. For example we might use a linear approximation to the Q-functions, use regression to estimate the parameters (α_j, β_j), and then choose the decision rules so as to maximize the estimated Q-functions. We now describe the Q-learning algorithm with function approximation as in (Murphy 2005). To start, define the algorithm follows.

$$\begin{aligned}
(\hat{\alpha}_2, \hat{\beta}_2) &= \arg \min_{\alpha_2, \beta_2} \frac{1}{n} \sum_{i=1}^{n} (Y_i - Q_2(\bar{M}_{2i}, \bar{T}_{2i}; \alpha_2, \beta_2))^2 \\
\hat{d}_2^{opt}(\bar{M}_2, T_1) &= \arg \max_{T_2} Q_2(\bar{M}_2, \bar{T}_2, \hat{\alpha}_2, \hat{\beta}_2) \\
\hat{Y}_{1i} &\leftarrow \max_{T_2} Q_2(M_{2i}, \bar{T}_2; \hat{\alpha}_2, \hat{\beta}_2), \quad i = 1, 2, \cdots, n.
\end{aligned}$$

$$\begin{aligned}
(\hat{\alpha}_1, \hat{\beta}_1) &= \arg \min_{\alpha_1, \beta_1} \frac{1}{n} \sum_{i=1}^{n} (\hat{Y}_{1i} - Q_1(M_{1i}, T_{1i}; \alpha_1, \beta_1))^2 \\
\hat{d}_1^{opt}(M_1) &= \arg \max_{T_1} Q_1(M_1, T_1, \hat{\alpha}_1, \hat{\beta}_1) \\
\text{Output } \hat{\pi} &= (\hat{d}_1, \hat{d}_2) \text{ as the estimated optimal policy}
\end{aligned}$$

More precisely, Q-learning is a form of approximate dynamic programming where the conditional mean responses are estimated from the data since they cannot be computed explicitly (Laber et al. 2010). In order to use data to construct decision rules, a data analysis model that relates response to the pretreatment variables is employed. A particularly simple but useful model is,

$$Y = \alpha_0 + \alpha_1 V_1 + \cdots + \alpha_p V_p + (\beta_0 + \beta_1 W_1 + \cdots + \beta_q W_q)T + \epsilon$$

where V's and W's are selected state and action variables or summaries of the selected state and action variables and ϵ is the error term. The coefficients, that is the α's and β's, might be estimated from data using regression analysis. Suppose a maximum value of Y corresponds to a good response, then the decision rule is determined as follows: The first step in constructing the decision rule is to maximize

$$(\beta_0 + \beta_1 W_1 + \cdots + \beta_q W_q)T.$$

Let T be binary $\{0, 1\}$. If this algebraic maximization is performed, the decision rule is given a patient with the variables $(W_1, \cdots, W_q)$, choose $T = 1$ if the sum $(\beta_0 + \beta_1 W_1 + \cdots + \beta_q W_q) \geq 0$ and choose $T = 0$ otherwise. In the example of $K = 2$, we can consider each initial state so that,

$$Q_2(\bar{M}_2, \bar{T}_2) = \alpha_0 + \alpha_1 T_1 + \alpha_2 M_2 + \alpha_3 T_1 M_2 + (\beta_0 + \beta_1 T_1 + \beta_2 M_2 + \beta_3 T_1 M_2)T_2$$

For simplicity, this equation can be written as,

$$Q_2 = \gamma_{20} + \gamma_{21} T_2,$$

then we can decide the optimal policy by choosing $T_2^{opt} = 0$ when $\gamma_{21} < 0$ and $T_2^{opt} = 1$ when $\gamma_{21} > 0$. For the example, the fitted Q_2 models of the four initial states are

$$\begin{aligned}
Q_2(M_1 = 1) &= 13 + T_1 - M_2 + T_1 M_2 + (1 + 6T_1 + M_2 - 6T_1 M_2)T_2 \\
Q_2(M_1 = 2) &= 14 - 5T_1 - M_2 + T_1 M_2 - (5 - 12T_1 - M_2 + 6T_1 M_2)T_2 \\
Q_2(M_1 = 3) &= 11 + T_1 - M_2 + T_1 M_2 + (1 + 18T_1 + M_2 - 6T_1 M_2)T_2 \\
Q_2(M_1 = 4) &= 12 - 5T_1 - M_2 + T_1 M_2 - (5 - 24T_1 - M_2 + 6T_1 M_2)T_2
\end{aligned}$$

and following the dynamic programming (DP) or backward induction, we determine optimal actions, working iteratively from the last time stage, as explained in the following table.

States	Rewards	T_1	M_2	γ_{20}	γ_{21}	T_2^{opt}	$Q_2^{opt} = \max E(Y\|T_1, M_2, T_2^{opt})$
1	(6,8)	0	1	12	2	1	14
			3	10	4	1	14
		1	1	14	2	1	16
			2	14	-3	0	14
2	(6,3)	0	2	12	-3	0	12
			4	10	-1	0	10
		1	1	09	2	1	11
			2	09	-3	0	09
3	(4,8)	0	1	10	2	1	12
			3	08	4	1	12
		1	3	12	4	1	16
			4	12	-1	0	12
4	(4,3)	0	2	10	-3	0	10
			4	08	-1	0	08
		1	3	07	4	1	11
			4	07	-1	0	07

Table 6.2: Optimal actions and maximum expected rewards using Q-learning

To obtain the optimal policies at $j = 1$, we fit the model of Q_1 which is a function of Q_2^{opt}, M_1 and T_1 (according to the equation on page 123, line 12 that we replace Y_{1i}, which equals to $\max_{T_2} Q_2$ instead of Y_i), then we can consider the model: $Q_1 = \gamma_{10} + \gamma_{11}T_1$, then we can decide the optimal policy by choosing $T_1^{opt} = 0$ when $\gamma_{11} < 0$ and $T_1^{opt} = 1$ when $\gamma_{11} > 0$. Using the Q_1 models,

$$\begin{aligned} Q_1(M_1 = 1) &= 14 + 0.8T_1, \qquad Q_1(M_1 = 2) = 10.4 - 0.4T_1 \\ Q_1(M_1 = 3) &= 12 + 1.6T_1 \text{ and } Q_1(M_1 = 4) = 8.6 + 0.4T_1 \end{aligned}$$

We determine optimal actions, as explained in the following table.

M_1	Rewards	γ_{10}	γ_{11}	T_1^{opt}	$Q_1^{opt} = \max E(Y\|T_1^{opt})$
1	(6,8)	14	0.8	1	14.8
2	(6,3)	10.4	-0.4	0	10.4
3	(4,8)	12	1.6	1	13.6
4	(4,3)	8.6	0.4	1	9

In general we can use least squares iteration to estimate the Q-function parameters. If $j = K$, then these will be equal to the true values. But for $j = 1, \cdots, K-1$ they will be different but close to true values in case of large samples. The next table shows the Q_1 estimated parameters using 100 simulations

		$n = 100$	$n = 1000$	$n = 10000$
$M_1 = 1$	γ_{10}	14.0000	14.0000	14.0000
	γ_{11}	0.6667	0.8209	0.8008
$M_1 = 2$	γ_{10}	10.6154	10.3837	10.3973
	γ_{11}	-0.4904	-0.5526	-0.38933
$M_1 = 3$	γ_{10}	12.0000	12.000	12.0000
	γ_{11}	1.8980	1.8461	1.6013
$M_1 = 4$	γ_{10}	8.6667	8.6129	8.5990
	γ_{11}	0.7333	0.5595	0.4037

6.5.2 Inverse probability treatment weighted method

As explained in Chapter 5, the IPTW formula based on $\bar{M}$ for the counterfactual mean $E[Y(\bar{t})]$ is the average of Y among subjects with $\bar{T} = \bar{t}$ in a pseudo-population constructed by weighting each subject by their subject-specific IPTW

$$W = \prod_{j=1}^{K} \frac{1}{P[T_j|\bar{T}_{j-1}, \bar{M}_j]},$$

Recall the previous bandit example with a sequence M_1, T_1, M_2, T_2 of binary states changing after playing according to the transition matrices, actions and mean total rewards Y measured at the end of follow-up. All subjects are assumed to start in a randomly chosen selection of the four different states 1, 2, 3 and

M_1	$N(M_1)$	T_1	$N(T_1)$	$R(T_1)$	M_2	T_2	$R(T_2)$	$E(Y\|\bar{M}_2,\bar{T}_2)$	N	W	$N_{ps(W)}$	$E[Y(t_2^{opt})\|\bar{M}_2,T_1]$
1	250	0	75	6	1	0	6	12	9	100/18	50	
						1*	8	14	6	100/12	50	14
					3	0	4	10	12	100/6	200	
						1*	8	14	48	100/24	200	14
		1	175	8	1	0	6	14	42	100/42	100	
						1*	8	16	28	100/28	100	16
					2	0*	6	14	84	100/56	150	14
						1	3	11	21	100/14	150	
2	250	0	175	6	2	0*	6	12	28	100/56	50	12
						1	3	9	7	100/14	50	
					4	0*	4	10	56	100/28	200	10
						1	3	9	84	100/42	200	
		1	75	3	1	0	6	9	22.5	100/18	125	
						1*	8	11	15	100/12	125	11
					2	0*	6	9	30	100/24	125	9
						1	3	6	7.5	100/6	125	
3	250	0	100	4	1	0	6	10	18	100/24	75	
						1*	8	12	12	100/16	75	12
					3	0	4	8	14	100/8	175	
						1*	8	12	56	100/32	175	12
		1	150	8	3	0	4	12	12	100/12	100	
						1*	8	16	48	100/48	100	16
					4	0*	4	12	36	100/24	150	12
						1	3	11	54	100/36	150	
4	250	0	150	4	2	0*	6	10	36	100/48	75	10
						1	3	7	9	100/12	75	
					4	0	4	8	42	100/24	175	
						1*	3	7	63	100/36	175	8
		1	50	3	3	0	4	7	10	100/8	125	
						1*	8	11	40	100/32	125	11
					4	0*	4	7	20	100/16	125	7
						1	3	6	30	100/24	125	

$*$*For optimal actions at second time point.*

Table 6.3: States, actions, unstabilized pseudo-populations and optimal outcomes.

4. Then $N = 1000$ subjects are randomly assigned at two time points to play arm 1 or play arm 2 with any assuming probabilities, e.g., $P(T_1 = 0|M_1 = 1) = P(T_1 = 1|M_1 = 2) = 0.3$, $P(T_1 = 0|M_1 = 3) = P(T_1 = 1|M_1 = 4) = 0.4$ at time 1 and $P(T_2 = 0|M_2 = 1, M_1, T_1) = P(T_2 = 1|M_2 = 4, M_1, T_1) = 0.6$, $P(T_2 = 0|M_2 = 2, M_1, T_1) = P(T_2 = 1|M_2 = 3, M_1, T_1) = 0.8$ at time 2. In a real study all terms would be different and need to be estimated in practice. Table 6.3 describes how to use IPTW for estimating the means $E[Y(t)]$ under the different regimes $\bar{t} = \{t_1, t_2\}$. We can calculate the mean expected total rewards for each initial state under the different regimes $\{T_1 = t_1, T_2 = t_2^{opt}\}$. Columns 10 and 12 are the example assumptions of data population for each path and the pseudo-population which was calculated based on the inverse probability weighted method, where the actions are randomly assigned with the with the probabilities; $P(T_1|M_1)$ and $P(T_2|\bar{M}_2, T_1)$.

To decide the optimal strategies at time 1, we should compare $E(Y|\bar{M}_2, T_1 = 0, t_2^{opt})$ and $E(Y|\bar{M}_2, T_1 = 1, t_2^{opt})$ for each state. E.g., when the initial state is 1 then the optimal decision is to play arm 2, because $E(Y|\bar{M}_2, T_1 = 1, t_2^{opt}) = [16(100) + 14(150)]/250 = 14.8$ is greater than $E(Y|\bar{M}_2, T_1 = 0, t_2^{opt}) = [14(50) + 14(200)]/250 = 14.0$. In a random study all probabilities of each action given history would be different and need to be estimated. Then we weight each subject by their subject-specific IPTW

$$\hat{W} = \prod_{j=1}^{K} \frac{1}{\hat{P}[T_j|\bar{T}_{j-1}, \bar{M}_j]}.$$

The following calculations are done by estimated values using 100 simulations of data set size 1000

State	T_1	$E_{ps}(Y\|M_1, T_1)$
1	0	$14(048.21) + 14(201.72)/249.93 = 13.96$
	1*	$16(098.01) + 14(149.52)/247.53 = 14.79$
2	0*	$12(049.34) + 10(203.47)/252.81 = 10.39$
	1	$11(124.53) + 09(124.95)/249.48 = 09.98$
3	0	$12(076.11) + 12(174.72)/250.83 = 12.04$
	1*	$16(099.03) + 12(151.19)/250.22 = 13.59$
4	0	$10(074.02) + 08(175.88)/249.90 = 08.62$
	1*	$11(125.36) + 07(126.61)/251.02 = 08.99$

$*$*For optimal actions at first time point.*

The previous table shows how to choose optimal actions at time 1 by using $E(Y|\bar{M}_2, T_1, t_2^{opt})$ and the pseudo-population of IPTW for each path.

6.6 Simulation Results

As we have seen, regret-regression, Q-learning and IPTW methods in our example, lead to the same optimal policies. These policies for states $\{1,2,3,4\}$ are respectively $\{1,0,1,1\}$ at time 1 and $\{1,0,1,0\}$ at time 2. To be sure we use simulations of 1000 data sets with different sample sizes 100, 1000 and 10000 for estimating $\max E(Y|M_1,T_1)$, which equals 11.95 if using true values. The results are very close each other, as shown below,

	$n = 100$		$n = 1000$		$n = 100000$	
Method	Mean	SE	Mean	SE	Mean	SE
Regret-regression	11.944	0.081	11.954	0.056	11.950	0
Inverse probability	11.947	0.096	11.951	0.061	11.955	0.006
Q-learning	11.947	0.096	11.951	0.061	11.948	0.007

Although the three methods give the same policies and the same maximum rewards, problems arise when using IPTW method in samples of modest size (observing many paths without data). On the other hand the Q-learning method needs to estimate high-dimensional parameters which are different in each time point. But using regret-regression, only two model fits are required.

6.7 The Gittins Index

An alternative method of solution, based around *forwards induction* was introduced by Gittins and Jones (1974). Forward induction policies are constructed by choosing actions to maximize a measure of the current reward rate. Robinson (1982) define the dynamic allocation index (DAI) as

$$G[i,\tau] \;=\; \sup_{\tau>0} G[M_j(t),\tau], \quad \textit{where } j = 1,2,\cdots,K$$

$$= \sup_{\tau>0} \frac{E\left[\sum_{j=0}^{\tau-1} \lambda^j R\{M_j(t)\}|M_0(t) = i\right]}{1 - E\left(\lambda^\tau\right)},$$
$$= \sup_{\tau>0} \frac{E\left[\sum_{j=0}^{\tau-1} \lambda^j R\{M_j(t)\}|M_0(t) = i\right]}{E\left(\sum_{j=0}^{\tau-1} \lambda^j |M_0(t) = i\right)},$$

where $G[M_j(t)]$ is the expected discounted reward per expected unit of discounted time when arm t is operated from initial state $m_0(t)$ and τ is a positively valued stopping time defined on the process. The Gittins index policy is the one that selects the arm with the current largest DAI. Such policies, since Whittle (1980), are now referred to as *Gittins Index policies.* Stationary or non-stationary policies can be obtain using the Gittins Index. This is because whether the unselected arms remain still (frozen in their current states) or transition to another state with a different reward is inconsequential; in either case the player does not obtain the reward from arms he does not play (Whittle, 1980). There are a number of methods for calculating the Gittins index including direct calculation, calibration methods, linear programming and special purpose algorithms.

Theorem 5.1 (Whittle 1980)

A policy d^{opt} *is optimal for multi-armed bandit problem* $\{(\Omega_t, P_t, R[M_j(t)], \lambda) : 1 \leq t \leq T\}$, *if and only if at each decision time point*

$$d^{opt}(j) = i \Rightarrow G\{M_j(i)\} = \max_{1\leq t\leq T} G[M_j(t)]\}$$

Any such will be referred to as a Gittins index policy. Further, the maximum $G[M_j(t)]$ equation is achieved at a time

$$\tau^*[i] = \min_{j>0}\{j : G[M_j(t)] \leq G[M_j(t) = i]\}$$

That is, the first time that the processed arm enters a state that has a lower Gittins index than its initial Gittins index. The term Gittins index was adopted by Whittle (1980) in recognition of Gittins contribution. The advantage of this result is that the Gittins index for each arm is independent of the other arms. Hence, a multi-armed bandit problem is computationally equivalent to one-armed bandit problem. This means major computational savings can be made, allowing larger problems to be solved. Gittins index policies are rules to choose the arm with highest priority which maximizes the total expected discounted reward.

6.8 Calculating the Gittins Index

The methods for calculating the Gittins index are numerous and include direct calculation, calibration methods, linear programming and special purpose algorithms. A fuller account of some of these algorithms can be found in Gittins (1989). Direct computation of the Gittins index requires an exhaustive search over all possible times τ in the defining equation, to locate the stopping time τ^* which yields the index. As one would expect, this approach is computationally expensive and hence would rarely be recommended. Calibration techniques are equivalent to finding a retirement reward for which one would be indifferent between continuing the processing of the arm and accepting the retirement reward. Note that the retirement reward means the best reward form other arms at time τ when leaving arm t.

Linear programming algorithms have been proposed which are again based on the idea of a retirement reward and they can give the exact Gittins index. These use results derived from dynamic programming. The simple algorithm can, in principle, be used to extract value functions which are related to the Gittins index (Chen and Katehakis 1986). Special purpose algorithms have also been developed for certain models for which optimal policies can be found through the use of Gittins index. These include target problems, where the aim is to identify the arm with the highest expected reward, and sampling models for which some of the parameters are unknown and must be discovered through experimentation. Robinson (1982) describes in detail some of the most efficient approaches to an exact Gittins index index computation. In the following algorithm B_k is a $k \times k$ matrix, S_k and F_k are k-vectors and $k^{'}$ indicates the state whose DAI-ordering is $k(k = 1, 2, \cdots, n)$, i.e. $1^{'}$ is the state with the largest DAI, $2^{'}$ is the state with the second largest DAI, and so on; ties are separated arbitrarily.

Algorithm. State $1^{'}$ is some state m satisfying

$$R[m(t)] = \max_{\widetilde{m(t)}} R[\widetilde{m(t)}]$$

and $B_1(1,1) = (1 - \lambda P(1^{'}, 1^{'}))^{-1}$. For $k = 2, 3, \cdots, n$ state $k^{'}$ is some state m, not in the set $\{1^{'}, 2^{'}, ..., (k-1)^{'}\}$, satisfying

$$\frac{R(m) + \lambda \sum_{l<k} P(m, l^{'}) S_{k-1}(l)}{1 + \lambda \sum_{l<k} P(m, l^{'}) F_{k-1}(l)} = \max_{j} \left[\frac{R(\widetilde{m}) + \lambda \sum_{l<k} P(\widetilde{m}, l^{'}) S_{k-1}(l)}{1 + \lambda \sum_{l<k} P(\widetilde{m}, l^{'}) F_{k-1}(l)} \right]$$

where

$$S_{k-1}(l) = \sum_{w=1}^{k-1} B_{k-1}(l, w) R(w^{'}) \quad \text{and} \quad F_{k-1}(l) = \sum_{w=1}^{k-1} B_{k-1}(l, w).$$

The matrix B_k is given by

$$\begin{aligned}
B_k(k,k) &= \{1 - \lambda P(k^{'}, k^{'}) - \lambda^2 \sum_{l<k} P(k^{'}, l^{'}) Z_{k-1}(l)\}^{-1}, \\
B_k(m,k) &= \lambda B_k(k,k) Z_{k-1}(m) \qquad (m = 1, 2, \cdots, k-1), \\
B_k(k,\widetilde{m}) &= \lambda B_k(k,k) \sum_{l<k} P(k^{'}, l^{'}) B_{k-1}(l, \widetilde{m}) \qquad (\widetilde{m} = 1, 2, \cdots, k-1), \\
B_k(m,\widetilde{m}) &= B_{k-1}(m, \widetilde{m}) + \lambda Z_{k-1}(m) B_k(k, \widetilde{m}) \qquad (m \text{ and } \widetilde{m} = 1, 2, \cdots, k-1),
\end{aligned}$$

where

$$Z_{k-1}(m) = \sum_{l<k} B_{k-1}(m, l) P(l^{'}, k^{'}) \qquad (m = 1, 2, \cdots, k-1).$$

We turn now to the simple example, we assume a discounted reward situation that $\lambda = 1$. The following shows how to find a policy to maximize the total rewards using the Gittins index.

Arm 1

$$R(m) = \max_{\widetilde{m}} R(\widetilde{m}) = \max(6,4) = 6.$$

Suppose $\lambda = 0.9999$. Thus the Gittins Index for the first element in arm one reward vector is 6. Now we will calculate the Gittins Index for the second element,

$$\frac{R(m) + \lambda \sum_{l<k} P(m,l^{'})S_{k-1}(l)}{1 + \lambda \sum_{l<k} P(m,l^{'})F_{k-1}(l)} = \max_{j} \left[\frac{R(\widetilde{m}) + \lambda \sum_{l<k} P(\widetilde{m},l^{'})S_{k-1}(l)}{1 + \lambda \sum_{l<k} P(\widetilde{m},l^{'})F_{k-1}(l)} \right]$$

where

$$\begin{aligned} B_1(1,1) &= (1-\lambda P(1^{'},1^{'}))^{-1} = (1-0.9999P_1(1,1))^{-1} = (1-.9999(0.2))^{-1} = 1.24 \\ S_{k-1} &= \sum_{w=1}^{k-1} B_{k-1}(l,w)R(m^{'}) = B_1(1,1)R(1) = 1.24 \times 6 = 7.49 \quad \text{and} \\ F_{k-1} &= \sum_{w=1}^{k-1} B_{k-1}(l,w) = B_1(1,1) = 1.24 \end{aligned}$$

Thus the Gittins index for the second state is

$$DAI_{12} = \frac{R(\widetilde{m}) + \lambda P_1(2,l)S_1(l)}{1 + \lambda P_1(2,l)F_1(l)} = \frac{4 + 0.9999(0.3)(7.49)}{1 + 0.9999(0.3)(1.24)} = 4.54$$

Arm 2

$$R(m) = \max_{\widetilde{m}} R(\widetilde{m}) = \max(8,3) = 8$$

and the Gittins Index for the first element in arm two rewards vector is 8. Now we will calculate the Gittins Index for the second element,

$$\begin{aligned}
B_1(1,1) &= (1-\lambda P(1^{'},1^{'}))^{-1} = (1-0.9999P_2(1,1))^{-1} = (1-.9999(0.4))^{-1} = 1.66 \\
S_{k-1} &= \sum_{w=1}^{k-1} B_{k-1}(l,w)R(w^{'}) = B_1(1,1)R(1) = 1.66 \times 8 = 13.33 \quad \text{and} \\
F_{k-1} &= \sum_{w=1}^{k-1} B_{k-1}(l,w) = B_1(1,1) = 1.66
\end{aligned}$$

Thus the Gittins index for the second state is

$$DAI_{22} = \frac{R(\widetilde{m}) + \lambda P_2(2,l)S_1(l)}{1+\lambda P_2(2,l)F_1(l)} = \frac{3+0.9999(0.5)(13.33)}{1+0.9999(0.5)(1.66)} = 5.27.$$

So for this problem we have

Arm	DAI index (state 1)	DAI index (state 2)
1	6	4.54
2	8	5.27

Then our different policies are as follows

States	Rewards	Optimal action T_j using Gittins index	Optimal action T_j using play-the-winner policy
(1,1)	(6,8)	1	1
(1,2)	(6,3)	0	0
(2,1)	(4,8)	1	1
(2,2)	(4,3)	1	0

Under almost no discounting ($\lambda = 0.9999$), the difference between Gittins Index policy and play-the-winner rule is at state (2,2) where the rewards on offer are (4,3). The Gittins policy of choosing arm one acknowledges future expectation - the possibility that the arm one reward value could change from 3 to 8 - whereas the play-the-winner rule is myopic and takes the higher immediate reward of 4 on offer from arm one. Both of them are derived under an assumption that the process continues indefinitely and the optimal policy is stationary.

6.9 Linear Programming

The previous section explained how it is difficult to calculate Gittins index, even in the simple case of a multi-armed bandit problem. This section introduces an easy alternative approach to calculate Gittins index. The linear programming approach for the multi-armed bandit problem can be found in Chen and Katehakis (1986). They considered a finite state bandit process and were able to demonstrate the Gittins index for state M can be obtained by solving a linear program. The problem to be considered involves T variables $M_1, M_2, \cdots, M_T$, linear functions

$$U(M, R) = \max \sum_{j \geq 0} \sum_{t=1}^{T} R_j I(M_j(t)),$$

and T constraint equations

$$\sum_{t=1}^{T} I(M_j(t)) = K, \quad \text{for } j = 1, 2, \cdots, K$$

where $I(M_j(t))$ is an indicator and R_j is a reward of $M_j(t)$. The following conventions will be observed throughout the remainder of this section:

> (i) The constraint equations have at least one non-negative real solution and are such that they are linearly independent.
>
> (ii) The objective function can not be expressed as a linear combination of the constraint function.

6.9.1 Linear programming with a finite horizon

Suppose $\{V^*, W_i^*, i \in \Omega\}$ are the optimal solution to linear program,

$$\min_U U = V + \sum_{i \in \Omega} W_i$$

subject to

$$\begin{aligned} V + W_i &\geq R(i) + \lambda \sum_j P(i,j) W_j \quad i, j \in \Omega \ - \ \{m\} \\ V &\geq R(m) + \lambda \sum_j P(m,j) W_j \quad j \in \Omega \ - \ \{m\} \\ W_i &\geq 0, \ V \in R, \qquad\qquad\qquad i \in \Omega, \end{aligned}$$

where m is a specific state, i for other states and the sign $(- \ \{m\})$ denotes except m. Then it follows that $G(m) = V^*$ and $W_m^* = 0$. Therefore, if the above linear program can be solved to obtain $G(m)$, it is possible to construct an efficient procedure to calculate the Gittins indexes $G(m)$ for $m = 1, 2, \cdots, \Omega$. However, Kallenberg (1986) observed that the number of pivot steps (the pivot step is to choose an element corresponding the location of the largest rewards and the smallest costs) will be highly dependent upon the chosen permutation of the states in Ω. When $\Omega = 2$, then to calculate $G(m_1)$ and $G(m_2)$ only two constraints in the linear program have to replaced. Also in the objective function we will have only two variables, then the optimal solution can be obtained by a simple graphical solution. If the problem contains more than two variables, then we need to use other solution methods such as the *simplex method* (explained later). In our previous example, we found the Gittins index for arm 1, state 1 is 6 and 8 for arm 2, state 1. To calculate the Gittins index for arm 1, state 2 the linear program formula will be as

$$\min_U U = V + W_1$$

subject to

$$V + W_1 \geq R(1) + \lambda P(1,1) W_1$$

$$V \geq R(2) + \lambda P(2,1)W_1$$

$$W_1 \geq 0,$$

then using the specific values in the transition matrix and the rewards vector, these two constraints will be $V + 0.8W_1 \geq 6$ and $V - 0.3W_1 \geq 4$. Then the possible graphical solution points are $(\{W, V\} = \{(7.5, 0), (1.82, 4.54)\})$. Hence $V^* = 4.54$. Similarly, the Gittins index for arm 2, state 2 can be found by solving the following linear program formula

$$\min_U U = V + W_1$$

subject to

$$V + 0.6 \geq 8 \quad \text{and} \quad V - 0.5 \geq 3$$

$$W_1 \geq 0,$$

The possible solution points are $(\{W, V\} = \{(13.33, 0), (4.54, 5.27)\})$. Thus $V^* = 5.27$. For state 2 in arm 1. These are the same results as seen earlier.

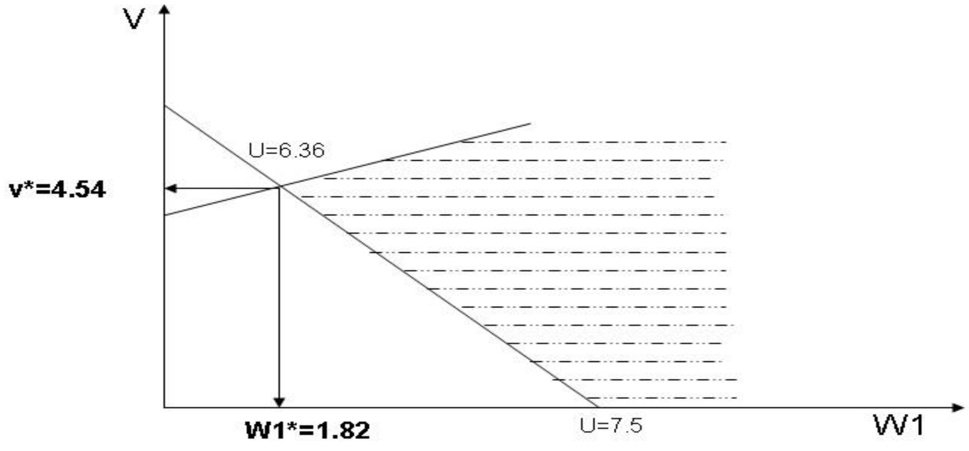

Figure 6.3: The Gittins index by LP

6.9.2 Results for $K = 5$

We will use the same details in the last example in Section 3, but assuming $K = 5$.

	Regret				
State M_j	$j = 1$	$j = 2$	$j = 3$	$j = 4$	$j = 5$
1	0.309	0.309	0.305	0.781	2
2	0.309	0.308	0.312	0.393	3
3	0.855	0.856	0.851	1.562	4
4	0.855	0.855	0.859	0.396	1

Table 6.4: Regrets for two-arm bandit problem when $K = 5$.

Dynamic regret-regression policies for the initial states (1,2,3,4) are (1, 1, 1, 1) at all time points from $j = 1$ until $j = K - 2$. Then it uses (1, 0, 1, 1) for $j = K - 1$ and (1, 0, 1, 0) for $j = K$. Choosing the wrong action will cost us the regrets, given in Table 6.4. Then the optimal expected rewards, using simulations from a dataset size 1000 are shown in Table 6.5.

Initial State M_1	1	2	3	4
Rewards	6 or 8	6 or 3	4 or 8	4 or 3
Optimal expected rewards	30.59	26.07	29.27	24.75
Overall mean	27.67			

Table 6.5: Regret-regression optimal mean expected rewards for two-arm bandit problem when $K = 5$.

Now we will compare regret-regression results with other methods such as a *random policy* (playing each arm with probability 0.5), *play-the-winner* (playing an arm which gives a maximum reward) and Gittins index policies. The last two methods have static policies, (1, 0, 1, 0) and (1, 0, 1, 1), respectively, for all time points. For reference we give the mean reward under four decision regimes:

Policy	Expected Reward
Random, $p = 0.5$	25.2
Play-the-winner	26.0
Gittins Index (or linear programming)	27.1
Optimal Dynamic	27.67

Table 6.6: Mean rewards under different strategies when $K = 5$.

6.9.3 Linear programming policy with an infinite horizon

In the previous section we showed how Gittins Index (linear programming) can be calculated using a finite horizon with K time points. In this section we will compare results of linear programming with regret-regression policies, when there are an infinite number of decision time points. First we describe how to find these policies, then we illustrate the comparison by a numerical example.

Suppose a bandit has four states and two arms. Arm one has two possible states, denoted 1 and 2. Arm 2 also has two possible states, denoted 3 and 4. Initially one of states is in arm 1 and the other in arm 2. Arms may change state only after completing service, according to Markovian transition probabilities. In arm 1, State 1 may thus either remain in the same state, with probability $P(1,1)$, or transfer to state 2, with probability $P(1,2) = 1 - P(1,1)$, and when arm 1 in state 2, state 1 may be entered with probability $P(2,1)$, or re-enter state 2, with probability $P(2,2) = 1 - P(2,1)$. In arm 2, State 3 may thus either remain in the same state, with probability $P(3,3)$, or transfer to state 4, with probability $P(3,4) = 1 - P(3,3)$, and when arm 2 in state 4, state 3 may be entered with probability

$P(4,3)$, or re-enter state 4, with probability $P(4,4) = 1 - P(4,3)$. Each time a state completes its service, a reward R_k for k =1, 2, 3 and 4 is earned, discounted in time by a discount factor $0 < \lambda < 1$. The objective is to find a policy $d \in D$, that maximise the total expected discount reward earned over an infinite horizon. By defining $M_j(t)$ as indicator variable, that takes value 1 if state $M(t)$ in service at time j and 0 otherwise, then we can write the stochastic optimization problem of interest as follows,

$$V = \max_{d \in D} \left\{ E_d \left[\sum_{j \geq 0} \sum_{t=1}^{2} \lambda^j R[M_j(t)] I[M_j(t)] \right] \right\}$$

Now, returning to the linear programming method. We define,

1- m_1, m_2, m_3, m_4 are the number of service completions of states 1, 2 in arm 1 and 3, 4 in arm 2 respectively. E. g., $m_i = \sum I[M_j(i)]$ is the total number of service completions of state i.

2- R_1, R_2, R_3, R_4 are the rewards of states 1, 2, 3 and 4 respectively.

3- The transition matrices will be,

$$P_1 = \begin{pmatrix} P(1,1) & P(1,2) \\ & \\ P(2,1) & P(2,2) \end{pmatrix} \quad \text{and} \quad P_2 = \begin{pmatrix} P(3,3) & P(3,4) \\ & \\ P(4,3) & P(4,4) \end{pmatrix}.$$

Performance measures :

As with all linear programming problems, the objective must be expressed in terms of suitable variables. For the multi-armed problem, a natural performance measure is the total discount number of service completions for each state, namely

$$m_i^d = E_d \left[\sum_{j \geq 0} \lambda^j I[M_j(i)] \right]$$

The variable m_i^d is the *performance* of state m_i under policy d. By varying the policy adopted, the corresponding performance vectors span the region of achievable performance

$$\Omega = \{m_1^{d_1}, m_2^{d_2}, m_3^{d_3}, , m_4^{d_4}\}$$

Hence our stochastic optimization problem of finding an optimal performance vector can now be reformulated as a mathematical program

$$V = \max_{(m_1,m_2,m_3,m_4)\in\Omega} \sum_{t=1}^{2} R[M(t)]]$$

Conservation laws:

To solve this equation a complete description of the performance region Ω needs to be identified. This is done in terms of linear constraints on the performance vectors, the physical laws that describe the effects on the system of the scheduling policy implemented. First, since at each time exactly one arm completes its service, it follows that the total expected discounted number of completed arm is the same under any policy. This conservation can be written as

$$\begin{aligned} m_1^d + m_2^d + m_3^d + m_4^d &= E\{\sum_{j\geq 0}\sum_{t\geq 1}^{2} \lambda^j I[M_j(t)]\} && (6.2)\\ &= \sum_{j\geq 0} \lambda^j = \frac{1}{1-\lambda} && (6.3)\end{aligned}$$

To construct other relationships, let d_3 be a scheduling policy that gives priority to state 3 over other states. Under such a policy, the conservation law can be seen that,

$$m_3^{d_3} = \frac{1}{1-\lambda p(3,3)} + \lambda p(4,3)\frac{1}{1-\lambda p(3,3)} m_4^{d_3}. \quad (6.4)$$

This equation expresses the intuitive fact that the total expected discounted number of state 3 completed, $m_3^{d_3}$, can be decomposed into two terms. The first term is a constant, $\frac{1}{1-\lambda p(3,3)}$, that accounts for state 3 completed until the arm that was initially in state 3 transfers for the first time to state 4. The second term, $\lambda p(4,3)\frac{1}{1-\lambda p(3,3)}m_4^{d_3}$, accounts for the number of service completions afterwards. Clearly, $m_3^{d_3}$ is the maximum number of service completions and so it can be concluded that,

$$m_3^d \leq \frac{1}{1-\lambda p(3,3)} + \lambda p(4,3)\frac{1}{1-\lambda p(3,3)}m_4^d \quad \text{for all} \;\; d \in D.$$

Using Equation 6.2 and we may rewrite the first inequality as

$$m_1^d + \left[1 + \lambda\frac{p(3,2)}{1-\lambda p(2,2)}\right]x_3^\pi \geq \frac{1}{1-\lambda} - \frac{1}{1-\lambda p(2,2)} \quad \text{for all} \;\; d \in D.$$

Then we derive the linear programming system as follows

$$Z = \;\; \max R_1 m_1 + R_2 m_2 + R_3 m_3 + R_4 m_4$$

subject to

$$(1+\lambda\frac{p(2,1)}{1-\lambda p(1,1)})m_2 + m_3 + m_4 \geq \frac{1}{1-\lambda}$$

$$m_1 + m_2 + (1+\lambda\frac{p(4,3)}{1-\lambda p(3,3)})m_4 \geq \frac{1}{1-\lambda} - \frac{1}{1-\lambda p(3,3)}$$

$$(1+\lambda\frac{p(1,2)}{1-\lambda p(2,2)})m_1 + m_3 + m_4 \geq \frac{1}{1-\lambda}$$

$$m_1 + m_2 + (1+\lambda\frac{p(3,4)}{1-\lambda p(4,4)})m_3 \geq \frac{1}{1-\lambda}$$

$$m_1 + m_2 + m_3 + m_4 = \frac{1}{1-\lambda}$$

$$m_1, m_2, m_3, m_4 \geq 0$$

6.9.4 Simplex method

The graphical method is limited to solve linear programming problems having one or two decision variables. However, it provides where the feasible and non-feasible regions are, as well as, *vertices* (corner points of its feasible region). If a linear program has a non-empty, bounded feasible region, then the optimal solution is always one the vertices. Therefore, what is needed to be done is to find all the intersection points (vertices) and then examine which one among all feasible vertices, provides the optimal solution. An algorithm for solving the classical linear programming problem is developed by George B. Dantzig in 1947. The *simplex method* (algebraic method) is an iterative procedure, solving a system of linear equations in each of its steps, and stopping when either the optimum is reached, or the solution proves infeasible. Standard maximization problem is a linear programming problem for which the objective function is to be maximized and all the constraints are inequalities. For more details and examples can be found in Bartels and Golub (1969). Our two-armed bandit problem example assuming infinite number of decision time points can be written as:

$$Z = \max\ 6m_1 + 4m_2 + 8m_3 + 3m_4$$

subject to

$$\begin{array}{rrrrl} & 1.374953m_2 + & m_3 + & m_4 & \geq 10000 \\ m_1 + & m_2 + & & 1.833194m_4 & \geq 9998.333 \\ 3.665778m_1 + & & m_3 + & m_4 & \geq 10000 \\ m_1 + & m_2 + & 2.19976m_3 & & \geq 10000 \\ m_1 + & m_2 + & m_3 + & m_4 & = 10000 \end{array}$$

$$m_1,\ m_2,\ m_3,\ m_4 \geq 0$$

When the simplex method is used in this problem, if the problem has a solution, then the solution occurs at one of the vertices of a solution region in a multi-dimensional space. We start at one of the vertices and check the neighboring vertices to see which ones provide a better solution. We then move to one of the vertices that give a better solution. The process is repeated until the target vertex is reached. The first step of the simplex method requires that each inequality be converted into an equation. Inequalities are converted to equations by including *slack variables* S_1, S_2, S_3 and S_4. The constraints become:

$$Z = \max\ 6m_1 + 4m_2 + 8m_3 + 3m_4$$

subject to

$$\begin{aligned}
1.374953m_2 + m_3 + m_4 - S_1 &= 10000\\
m_1 + m_2 + 1.833194m_4 - S_2 &= 9998.333\\
3.665778m_1 + m_3 + m_4 - S_3 &= 10000\\
m_1 + m_2 + 2.19976m_3 - S_4 &= 10000\\
m_1 + m_2 + m_3 + m_4 &= 10000
\end{aligned}$$

$$m_1,\ m_2,\ m_3,\ m_4,\ S_1,\ S_2,\ S_3,\ S_4 \geq 0,$$

Setting $m_1, m_2, m_3,$ and m_4 to 0, we can read off the values for the other variables: $S_1 = 10000, S_2 = 9998.333, S_3 = 10000$ and $S_3 = 10000$. This specific solution is called a dictionary solution. The slack variables can be included in the objective function with zero coefficients. The LP problem, with greater than or equal and equality constraints, are transformed to its standard form in the way, that one artificial variable A is added to each of the constraints to ensure an initial basic feasible solution. Therefore we add artificial variables (A_1, A_2, A_3, A_4 and A_5) to those equations and give them a large negative coefficient (M) in the objective function, to penalize them. The problem can now be considered as solving a system of 6 linear equations involving the 14 variables $Z, m_1, m_2, m_3, m_4, S_1, S_2, S_3, S_4, A_1, A_2, A_3, A_4$ and A_5.

$$Z - 6m_1 - 4m_2 - 8m_3 - 3m_4 + MA_1 + MA_2 + MA_3 + MA_4 + MA_5 = 0$$

$$1.374953\ m_2 + m_3 + m_4 - S_1 + A_1 = 10000$$

$$m_1 + m_2 + 1.833194\ m_4 - S_2 + A_2 = 9998.333$$

$$3.665778\ m_1 + m_3 + m_4 - S_3 + A_3 = 10000$$

$$m_1 + m_2 + 2.199760\ m_3 - S_4 + A_4 = 10000$$

$$m_1 + m_2 + m_3 + m_4 + A_5 = 10000,$$

in such a way that Z has the maximum value. The system of linear equations can be written as above, where all $m's$, $S's$ and $A's$ are non-negative. As shown, we select a large positive number M (thus the name, the *Big-M method*) and form a new objective function as

Maximize

$$\begin{aligned} Z &= 6m_1 + 4m_2 + 8m_3 + 3m_4 - M(A_1 + A_2 + A_3 + A_4 + A_5) \\ Z &= (6 + 6.665778M)m_1 + (4 + 4.374953M)m_2 + (8 + 5.199760M)m_3 \\ &+ (3 + 4.833194M)m_4 - MS_1 - MS_2 - MS_3 - MS_4 - 49998.333M \end{aligned}$$

In the simplex method, the augmented matrix is referred to as the tableau,

Var	m_1	m_2	m_3	m_4	S_1	S_2	S_3
Z	-72.657780	-47.749531	-59.99760	-51.331945	10	10	10
A_1	0.000000	1.374953	1.00000	1.000000	-1	0	0
A_2	1.000000	1.000000	0.00000	1.833194	0	-1	0
A_3	3.665778	0.000000	1.00000	1.000000	0	0	-1
A_4	1.000000	1.000000	2.19976	0.000000	0	0	0
A_5	1.000000	1.000000	1.00000	1.000000	0	0	0

Var	S_4	A_1	A_2	A_3	A_4	A_5	R.S
Z	10	0	0	0	0	0	-499983.334
A_1	0	0	1	0	0	0	10000.000
A_2	0	0	0	1	0	0	9998.333
A_3	0	0	0	0	0	0	10000.000
A_4	-1	0	0	0	0	1	10000.000
A_5	0	1	0	0	0	0	10000.000

Although the problem can be solved by using M, but to keep the calculations simple, we assumed that $M = 10$, i.e. the value of first element in the first row is $-(6 + 6.665778M = 6 + 6.665778 \times 10) = -72.657780$. The previous tableau represents an initial solution. The initial solution assumes that all available resources are unused i.e. the slack variables take the largest possible values therefore all m's and Z are equal to zero. The simplex method uses a four step process to go from one tableau to the next.

Step 1

Select the pivot column with the "most negative" element in the objective function row (determine which variable to enter into the next solution).

Step 2

Select the pivot row (determine which variable to replace in the new solution). Divide the last element in each row by the corresponding element in the pivot column. The pivot row is the row with the smallest non-negative result. A_3 should be replaced by m_1 in the new solution.

Step 3

Let r_i denotes the row i, where $i = 1, 2, \cdots, 6$. Calculate new values for the pivot row. Divide every number in the row by the pivot number, $r_4 =$ the previous $r_4/3.665778$. Thus r_4 will be, $m_1, 1, 0, 0.2727934, 0.2727934, 0, 0, -0.2727934, 0, 0, 0, 0, 0.2727934, 0, 2727.934$

Step 4

Use row operations to make all numbers in the pivot column equal to 0 except for the pivot number which remains as 1, $r_1 =$ previous $r_1 + 72.657780 \times r_4$, $r_2 =$ previous $r_2 - r_4$, $r_3 =$ previous $r_3 - r_4$, $r_5 =$ previous $r_5 - r_4$ and $r_6 =$ previous $r_6 - r_4$. Then the next solution table is

Var	m_1	m_2	m_3	m_4	S_1	S_2	S_3
Z	0	-47.749531	-40.1770382	-31.5113825	10	10	-9.8205620
A_1	0	1.374953	1.0000000	1.0000000	-1	0	0.0000000
A_2	0	1.000000	-0.2727934	1.5604011	0	-1	0.2727934
m_1	1	0.000000	0.2727934	0.2727934	0	0	-0.2727934
A_4	0	1.000000	1.9269666	-0.2727934	0	0	0.2727934
A_5	0	1.000000	0.7272066	0.7272066	0	0	0.2727934

Var	S_4	A_1	A_2	A_3	A_4	A_5	R.S
Z	10	0	0	0	19.8205620	0	-301777.714
A_1	0	0	1	0	0.0000000	0	7270.400
A_2	0	0	0	1	-0.2727934	0	7270.400
m_1	0	0	0	0	0.2727934	0	2727.934
A_4	-1	0	0	0	-0.2727934	1	7272.066
A_5	0	1	0	0	-0.2727934	0	7272.066

Now repeat the steps until there are no negative numbers in the first row. The solution gives the total numbers of many times we play each arm until reach to the maximum value of Z, as follows

Var	m_1	m_2	m_3	m_4	S_1	S_2	S_3
Z	0	0.9997500062	0	0	0.72747934	2.727479e+00	0
m_4	0	0.7500312492	0	1	-0.54549587	-5.454959e-01	0
S_3	0	0.0004582108	0	0	2.66577799	-6.017095e-16	1
m_1	1	-0.3749531262	0	0	1.00000000	-1.725757e-16	0
m_3	0	0.6249218770	1	0	-0.45450413	5.454959e-01	0
S_4	0	-0.0002749631	0	0	0.00019998	1.199960e+00	0

Var	S_4	A_1	A_2	A_3	A_4	A_5	R.S
Z	0	18.7274793	9.27252066	7.272521e+00	10	10	5.272975e+04
m_4	0	-0.5454959	0.54549587	5.454959e-01	0	0	5.454050e+03
S_3	0	3.6657780	-2.66577799	6.017095e-16	-1	0	8.830727e-13
m_1	0	1.0000000	-1.00000000	1.725757e-16	0	0	4.313967e-13
m_3	0	0.5454959	0.45450413	-5.454959e-01	0	0	4.545950e+03
S_4	1	2.1999600	-0.00019998	-1.199960e+00	0	-1	1.099449e-12

$$\text{Thus} \quad \begin{pmatrix} Z \\ m_4 \\ S_3 \\ m_1 \\ m_3 \\ S_4 \end{pmatrix} = \begin{pmatrix} 52730 \\ 5454 \\ 0 \\ 0 \\ 4546 \\ 0 \end{pmatrix}$$

The final solution shows the next important results

- The linear programming policies are to use arm 2 only at all time points (proportions of playing arm 1 states $\{m_1, m_2\}$ are equal to zero).

- The total optimal rewards using linear programming policies are equal $8 \times 4546 + 3 \times 5454 = 52730$.

- The proportions of arm 2 states $\{m_3, m_4\}$ are equal to $\{0.4546, 0.5454\}$.

- There are no unused resources, where both S_3 and S_4 are equal to zero.

6.10 Regret-regression policy for an infinite horizon

Equation 4.3 in Chapter 4 Section 2 denotes the regression formula when using a finite horizon. As we see the formula depends on the function of residuals and the regrets. In a case of infinite horizon and discounted rewards, when time point j approach to infinity then discount rewards $\lambda^j R\{M_j(t)\}$ approximately equal to zero. Hence we can write the regret-regression formula for discounted rewards and infinite horizon as follows

$$E(Y|\bar{M}_\tau, \bar{T}_\tau) = \beta_0(M_1) + \sum_{j\geq 2} \lambda^j \beta_j^T(\bar{M}_{j-1}, \bar{T}_{j-1}) Z_j - \sum_{j\geq 1} \lambda^j \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}), \tag{6.5}$$

where τ is the converges time. In this specific example, results in this chapter, Section 6.9.2 shows that regret-regression policies for time points $j = 1, \cdots, K-2$ are the same and equal to $\{1, 1, 1, 1\}$ which is playing only arm 2 after all of observed states. But their policies for time points $K-1$ and K are differs and equal to $\{1, 0, 1, 1\}$ and $\{1, 0, 1, 0\}$. Because it is a discounted reward example, then discounted rewards, when j approaches to converges time, should be approximately equal to zero. Hence we can conclude that regret-regression policy for this an infinite horizon example is to play arm 2 at all time points and against all states.

6.10.1 Simulation Results

We mentioned in Section 6.9.2, that the linear programming policy for a finite horizon is $d^{opt} = \{1, 0, 1, 1\}$. But it differs for an infinite horizon that we should always play only arm 2 for all states. The method shows that the total discounted rewards for an infinite horizon when $\lambda = 0.9999$ will be 52730, where λ is discounted rate for infinite horizon. It is affect on the total non-discounted horizon by $1/(1-\lambda)$ time points. We compare linear programming with regret-regression for an infinite horizon by a simulation of 1000 dataset. Table 6.6 shows simulation results of one hundred repetitions for an infinite horizon (using time points $j \geq 0$ until converges at time with discount rewards equal to zero). When $\lambda = 0.9999$ it converges in mean at $j = 126112$ and gives total rewards with mean 52729.86 and standard deviation 121.67.

λ	0.6	0.7	0.8	0.9	0.945
Total rewards	13.10	17.67	26.42	54.07	95.93
Converges time	27	38	59	122	225
Times playing $T_2(M_1)$	12	17	27	56	102
Proportion playing $T_2(M_1)$	0.4574	0.4580	0.4561	0.4584	0.4536
Times playing $T_2(M_2)$	14	20	32	66	123
Proportion playing $T_2(M_2)$	0.5426	0.5419	0.5438	0.5415	0.5464
λ	0.99	0.9945	0.999	0.99945	0.9999
Total rewards	528.92	958.37	5271.48	9576.68	52729.86
Converges time	1257	2289	12608	22926	126112
Times playing $T_2(M_1)$	570	1041	5728	10421	57325
Proportion playing $T_2(M_1)$	0.4539	0.4548	0.4542	0.4545	0.4545
Times playing $T_2(M_2)$	686	1248	6880	12505	68787
Proportion playing $T_2(M_2)$	0.5460	0.5451	0.5457	0.5454	0.5454

Table 6.7: Simulation results for an infinite state of two-armed bandit problem

So the mean in both methods are approximately the same. It is not surprising because both polices are exactly the same for all time points except those policies at last two time points. So if we assume that K increases indefinitely, then discounted rewards of last two time intervals (which approximately equal to zero) do not effect on the total discounted rewards.

Hence these compared results on the same simulation problem have been showed that both methods have identical policies and total discount rewards against different values of the discounted rate *lambda* for infinite horizon. On the other hand we can conclude that regret-regression polices in the two-armed bandit problem on the same simulation problem on the same simulation problem for infinite horizon gives the same results with those using IPTW method or Q-learning but it avoids problems which arise when using samples of modest size or when there is need to estimate high-dimensional parameters. On the other hand regret-regression gives the same results with linear programming policy for an infinite horizon. But it has an advantage comparing Gittins index (or linear programming) for a finite horizon.

Chapter 7

Diagnostics

Regression diagnostics are required, because assumptions that underlay an analysis may not hold in any particular case. The diagnostic procedures are intended to check whether the assumptions of the regression model are satisfied or not (Lin et al, 2002). While model misspecification can affect the validity and efficiency of regression models, model checking has not become routine practice in the optimal dynamic treatment field, in part due to lack of suitable tools. One problem with the blip or regret structural nested mean models is that they are not based on a model for an observable quantity. This is not a problem with the regret-regression method described in Chapter 4. Residuals can be used to assess model adequacy. This will be illustrated in the next section.

7.1 Graphical Informal Testing

Model violations can be detected by means of graphical procedures and formal statistical tests. Graphical procedures will often be sufficient for validating model assumptions, but they may be supplemented by statistical tests (Ritz and Streibig, 2008). If there is uncertainty about the interpretation of a plot, it would be helpful to get a second judgment by using a formal test such as the *Likelihood Ratio Test* or *Wild Bootstrap Tests* which will be considered here.

7.1.1 Residual plots for model adequacy

Analysis of residuals is an effective method for assessing the fit of the model to the data and determining whether the model is useful. The recommended approach is to study a variety of residual plots and look for patterns and trends. The basic idea is that if a model is correct, then the residuals should have zero mean at all states or regrets. The appearance of a systematic trend may indicate the absence of an important covariate or an incorrect functional form. However, determining whether a pattern observed in a residual plot is due to a model misspecification or due to natural variation can be difficult (Johnston and So 2003).

7.1.2 Simulation study

The following simulation study is based on the Murphy (2003) scenario which was described in Chapter 4. We used the regret-regression method to estimate the parameters of the mean final response $E[Y|\bar{M}_{10} = \bar{m}_{10}, \bar{T}_{(10)2} = \bar{t}_{(10)2}]$ in a fully parameterized model,

$$\begin{aligned} E[Y|\bar{M}_{10}, \bar{T}_{(10)2}] &= \beta_1 + \beta_2 \sum_{j=1}^{10} (M_j - mean_j) - \sum_{j=1}^{10} \psi_1 (t_{j1} - I\{M_j > \psi_2\})^2 \\ &- \sum_{j=1}^{10} \psi_4 T_{j1} \{t_{j2} - (\psi_3 + \psi_5 M_j)\}^2 + \psi_7 (1 - T_{j1}) \{t_{j2} (\psi_6 - \psi_8 M_j)\}^2. \end{aligned}$$

$$M_1 \sim N(0.5, 0.01).$$

For $j \geq 2$

$$M_j \sim N(mean_j, 0.01),$$

where

$$\text{mean}_j = \gamma_1 + \gamma_2 M_{j-1} + \gamma_3 T_{\{j-1\}1} T_{(j-1)2} + \gamma_4 (1 - T_{\{j-1\}1}) T_{\{j-1\}2}. \tag{7.1}$$

In this section, we consider four different models, Model 1, Model 2, Model 3 and Model 4. All these models have similar generating functions for states and actions with the same mean and variance as was described in Murphy scenario. Model 1 is the only correct one. It uses the following regret function either for generating or estimating parameters. For treatment T_{j1} at time j, the regret function is

$$\mu_{j1}(t_{j1}|\bar{M}_j, \bar{T}_{j-1}, \psi) = \psi_1 \{t_{j1} - I(M_j > \psi_2)\}^2,$$

and

$$\mu_{j2}(t_{j2}|\bar{M}_j, \bar{T}_{j1}, \psi) = \psi_4 T_{j1} \{t_{j2} - (\psi_3 + \psi_5 M_j)\}^2 + \psi_7 (1 - T_{j1}) \{t_{j2}(\psi_6 - \psi_8 M_j)\}^2,$$

for second treatment T_{j2}. The other models are miss-specified models. Model 2 and Model 3 provide incorrect description of the regret formulas. We can write any of the regrets above as quadratic functions of the form $\mu(u) = \psi u^2$. Model 2 and Model 3 were investigated by Murphy. They assume a different functional form. For Model 2 it is a quadratic link function

$$\mu(u) = \begin{cases} u^2 & \text{if } u^2 \geq 0.83 \\ 0 & \text{otherwise,} \end{cases}$$

for the regrets of T_{j1}, and the next for the regrets of T_{j2}

$$\mu(u) = \begin{cases} u^2 & \text{if } u^2 \geq 3.33 \\ 0 & \text{otherwise.} \end{cases}$$

In Model 3, Murphy used a different link function

$$\mu(u) = \begin{cases} \mid u \mid & \text{if } u^2 < 1.5 \\ u^2 - 1.5 + \sqrt{1.5} & \text{otherwise,} \end{cases}$$

for the regrets of T_{j1}, and the next for the regrets of T_{j2}

$$\mu(u) = \begin{cases} \mid u \mid & \text{if } \; u^2 < 1.5 \\ u^2 - 2.5 + \sqrt{2.5} & \text{otherwise.} \end{cases}$$

Finally, in Model 4, we fit each state in time j using only the state in time $j-1$ but we ignore the effect of treatments in that time point. Thus we falsely assume mean$_j = \gamma_1 + \gamma_2 M_{j-1}$ rather that the expression given at Equation 7.1. Hence we have one correct model and three miss-specified models.

Plots of residual against fitted values

Figure 7.1 shows residual scatter plots of sample size 500 for the different models.

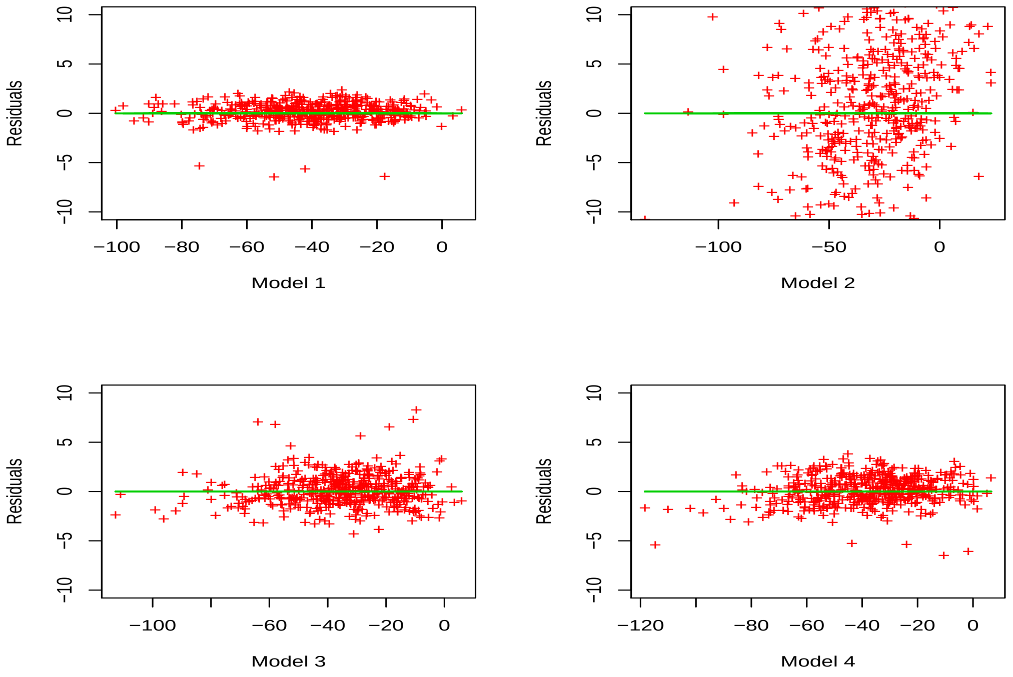

Figure 7.1: Plot of residuals against fitted values of sample size 500.

For Model 1 (the correct model), the scatter plot is roughly even distributed above and below zero with mean (-0.0000254) which is close to zero and a small standard deviation (0.973). On the other hand, the residuals of Model 2, Model 3 and Model 4 are distributed also with approximately zero mean values $\{0.0000017, -0.0000418, -0.0000002\}$ but larger standard deviations $\{7.846, 1.955, 1.891\}$ respectively. As we can see, the residual plots for Model 2, Model 3 and Model 4 show different variances against the residuals plot on the Model 1. Although the variances are larger for models 2, 3 and 4, there is no strong pattern.

The boxplots in Figure 7.2 were obtained from further datasets of the same sample size. For the correct model, the box covers the interval [-0.06, 0.06].

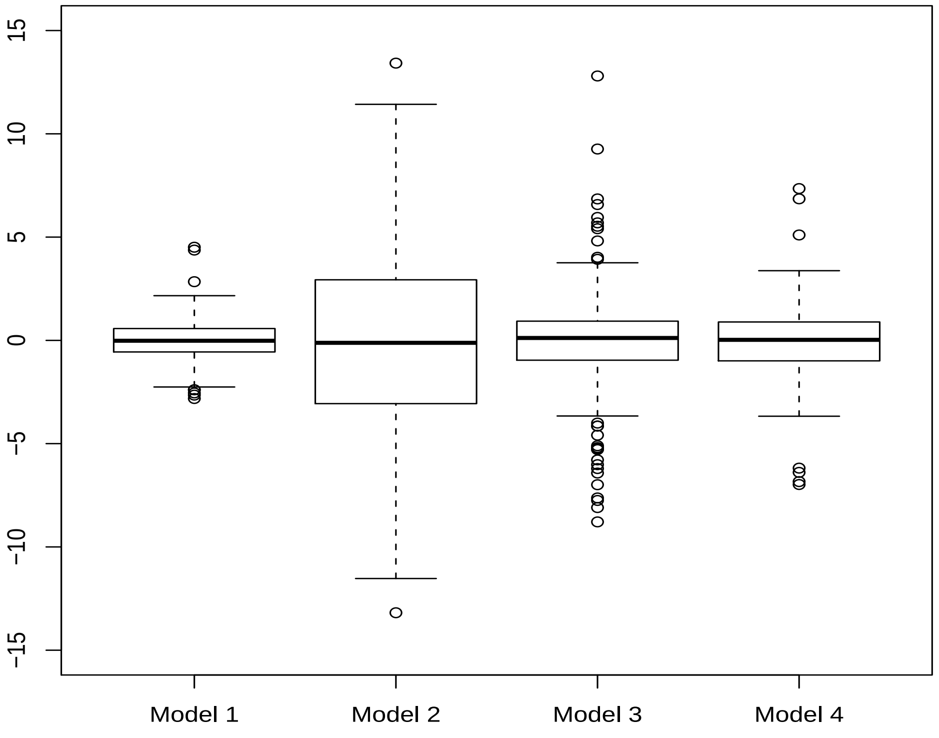

Figure 7.2: Boxplots of residuals from a sample size of 10000.

The median is very close to zero $\{-0.00176\}$. In other models, the medians of Model 2, Model 3 and Model 4 are

$\{-0.118, 0.115, 0.026\}$. Their inter-quartile ranges are $\{5.985, 1.883, 1.879\}$. As we see the differences from the correct model are large. Hence Model 2, Model 3 and Model 4 have a clear difference compared with the correct model. Following these results, we may say that Model 1 could be the correct model. But in fact we do not have a strong evidence to make this decision, because results differ from one dataset to another. Also without having Model 1 for reference we would not know that the variances were too high in the others. To compare these models, it would be better if we use other residuals plots against regrets or states, as used later in this chapter, Section 7.2.1 (see Figure 7.4).

Slope of the mean residuals line

The following suggestion is based on plots of residuals against regrets or states at different time points. If the model is correct, the residuals should not be dependent on each of them. Therefore the slope β between residuals and states or regrets should be equal to zero. To calculate the slope, we can use the formula

$$\hat{\beta} = \frac{\sum_{i=1}^{n} \hat{\mu}_i R_i}{\sum_{i=1}^{n} \hat{\mu}_i},$$

where $\hat{\beta}$ is the slope, R_i the residual of subject i and μ_i is the regret of i at any time point. By replacing M_i instead of $\hat{\mu}_i$, we can use the same formula to calculate the slope of the residuals against states. Table 7.1 gives results for $\hat{\beta}$, the average of slopes estimated from simulated samples of linear function of residuals on the regrets at the second decision part T_{j2} using the different models. It summarises a simulation of 1000 datasets each of size 500, at time points one, five and nine following the Murphy scenario. There are small values of the mean and standard error of the slope in Model 1. So regrets do not affect the mean residuals of Model 1 at these time points. On the other hand there is a clear

		Model 1	Model 2	Model 3	Model 4
First time point	Mean	0.0012	-0.1593	-0.2019	-0.0225
	Standard Deviation	0.050	0.223	0.121	0.076
	Median	0.0006	-0.1514	-0.1969	-0.0221
	Quartile Deviation	0.032	0.155	0.082	0.049
Fifth time point	Mean	-0.0033	-0.1698	-0.4723	-0.1329
	Standard Deviation	0.058	0.288	0.164	0.085
	Median	-0.0063	-0.1631	-0.4630	-0.1292
	Quartile Deviation	0.040	0.192	0.110	0.056
Ninth time point	Mean	-0.0001	-0.1545	-0.4666	-0.1267
	Standard Deviation	0.057	0.29	0.158	0.082
	Median	-0.0025	-0.1484	-0.4576	-0.1247
	Quartile Deviation	0.036	0.193	0.105	0.053

Table 7.1: Comparing different models through the average of slopes of residuals against regrets estimated from 1000 simulated samples of size 500, following the Murphy scenario.

negative relationship between the mean residuals and regrets for both Model 2 and Model 3. In Model 4 the mean slope values and their standard errors show effects of regrets on residuals but not so much compared with the other missspecifed models. Figure 7.3 shows the distribution of $\hat{\beta}$ in Model 1 is symmetric with zero mean. Histograms of the other models do not show these properties. Both Model 2 and Model 3 have non-zero mean and large variance compared with Model 1. However, the slopes of linear functions of residuals for the states for the different models indicate independence between states and residuals in Model 1 and might be similar in Model 2 but in Model 3 and Model 4 residuals are clearly far from zero at time points 5 and 9 (results are in the appendix).

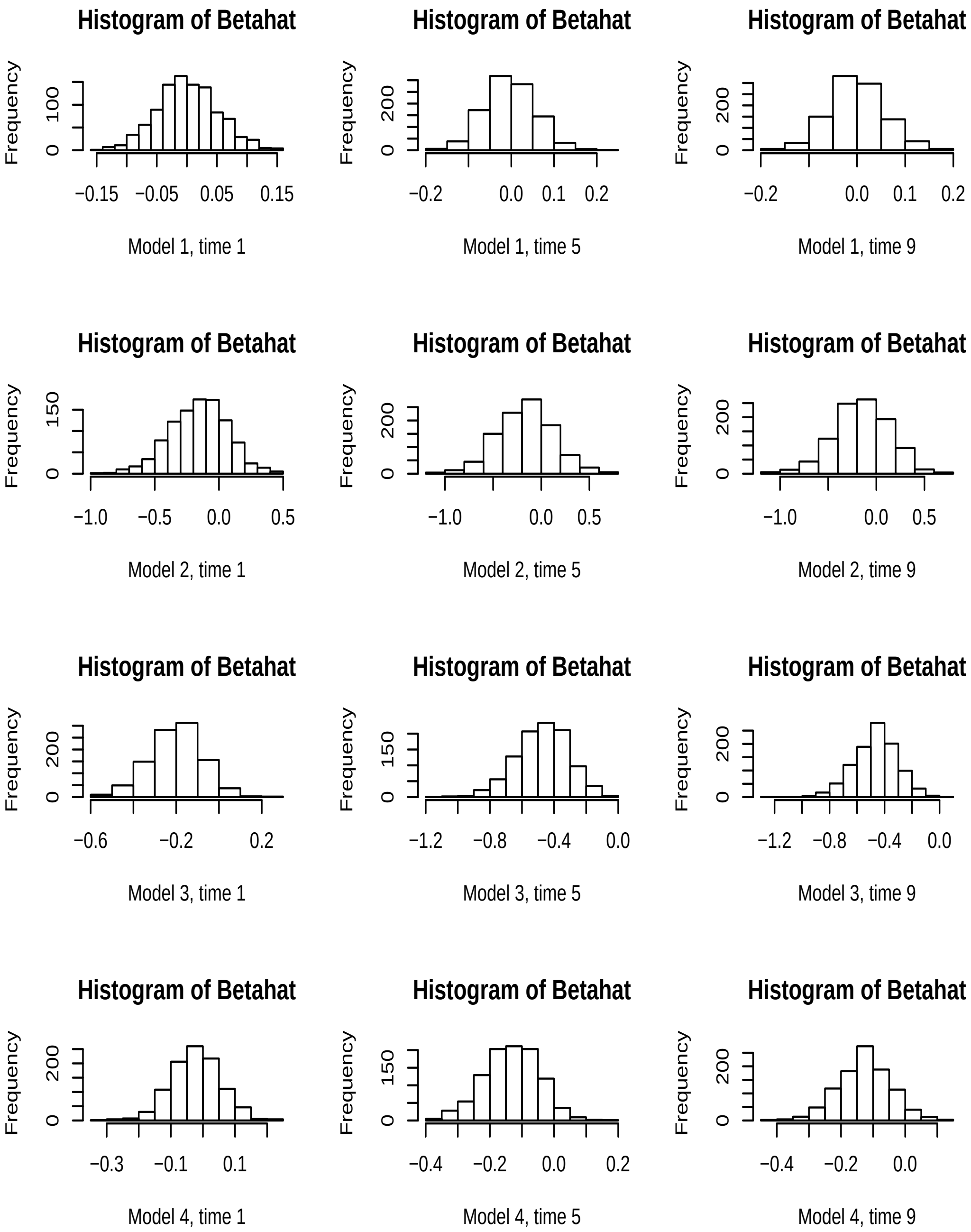

Figure 7.3: Histogram of $\hat{\beta}$ for different models at $j = 1, 5$ and 9 using simulations of 1000 datasets each of size 500, following the Murphy scenario.

According to the previous results, it would be helpful to get a second judgment by using a formal test. In the next section we will use a *Wild Bootstrap Test*, which is the first formal test. In the last section we will explore a *Likelihood Ratio Test* which was used to compute the p-value. If the model used was correct and the test suitable, then we may conclude that the p-value should be uniformly distributed on the interval [0, 1] if the usual error rate interpretation is to be valid. This will be approximately correct for bootstrap tests, but for other tests it can be far from correct.

7.2 Wild Bootstrap Test

The basic idea of any sort of hypothesis test is to compare the observed value of a test statistic, say $\hat{\beta}$, with the distribution that it would follow if the null hypothesis were true. The null is then rejected if $\hat{\beta}$ is sufficiently extreme relative to this distribution. Bootstrapping uses the sample data to estimate relevant characteristics of the population. The sampling distribution of a statistic is then constructed empirically by re-sampling from the sample. The key bootstrap analogy is the following: the population is to the sample as the sample is to the bootstrap samples. Often, an important point of bootstrapping is not just to evaluate estimates of the parameters, but also to obtain good estimates of standard errors from the distribution generated by the parameter estimates in bootstrapped iterations (Draper and Smith, 1998).

The wild bootstrap is a fairly simple modification on the standard bootstrap. The wild bootstrap method requires neither complete exchangeability nor a Gaussian distribution for the imaging data (Hongtu Zhu, 2007). The idea of the wild bootstrap is to use for the bootstrap disturbance associated with the i^{th} observation the actual residual for that observation, possibly transformed in some way, and multiplied by a random variable, independent of the data, with mean 0 and variance 1. Often, a binary random variable is used for this purpose. The details of this so-called wild bootstrap can be found in Liu (1988), Mammen (1993) and Wu (1986).

Suppose $D_1, D_2, \cdots, D_n$ are a sequence of independent zero mean random variables and suppose

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i,$$

converges in distribution to a random variable D. Let Z_i be iid random variables with zero mean and unit variance at this stage. There are, in principle, many ways of specifying the random variable Z_i. Liu (1988) and Mammen (1993) suggest alternative means of meeting the above requirements, the most widely used of which appeared to be the two point distribution

$$Z = \begin{cases} \frac{1+\sqrt{5}}{2}, \text{ with probability } p = \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \frac{1-\sqrt{5}}{2}, \text{ with probability } 1-p \end{cases}$$

The so-called Radamacher distribution is an alternative two point distribution:

$$Z = \begin{cases} 1 \text{ with probability } p = \frac{1}{2} \\ -1 \text{ with probability } 1-p \end{cases}$$

Both have the property $E[Z] = 0$ and $E[Z^2] = 1$. On the other hand we can use Z as standard normal, which does not have two points but still has $E[Z] = 0$ and $E[Z^2] = 1$. Then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i D_i,$$

also converges in distribution to D.

We would like to use their idea to form a test based on the residuals from a regret-regression model. An advantage over the standard bootstrap is that we do not need to assume as identical distribution for all subjects, though we do need to assume independence. This leads to a problem: If there is an intercept in the model, residuals R_i are not independent as $\sum_{i=1}^{n} R_i = 0$. A fix is to assume the contrast $n^{-1/2}\sum_{i=1}^{n} c_i R_i$, where $D_i = c_i R_i$ and $\sum_{i=1}^{n} c_i = 0$. Thus we will investigate whether a model is adequate or not using

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i c_i R_i.$$

For each dataset we repeated the wild bootstrap sampling 200 times. For example Figure 7.4 shows one dataset of size 500. It is divided to many parts dependent on different tests.

7.2.1 Test the different models based on Murphy scenario

We can plot residuals against states or regrets. As we know there are 10 states and 10 regrets. For the Murphy scenario, note we do not use regrets at the first decision part T_{1j} which is binary action $\{0, 1\}$, because a plot of residuals against only two regret values does not provide much information. So we will test residuals against states or even regrets for each j^{th} time point (in this particular case we test them only at first, fifth and ninth time points).

For the wild bootstrap test, we take a copy of the residuals and then we either multiplied each of them by a random standard normal $Z \sim N(0, 1)$, or a random uniform $Z \sim U\{-1, 1\}$. So each observed residual will be on the same vertical but it may go up or down. Then we compare the mean residuals of the original sample with the mean residuals of the wild bootstrap sample. After putting the data in the right order we design five different tests to investigate slope and trends. The following are these tests,

Test 1

This contrasts the mean of the first half of data (ordered by state or regret at time j) with the mean of the second half. Figure 7.4 shows these means for one simulated dataset under each of the four models. We can compare the observed difference between the two mean lines by the green lines. As stated, the problem in using the wild bootstrap is that the sum of the residuals is zero; hence they could not be independent. So we do a contrast, which means we multiply the first half by +1 and the second half by -1 (they will go up or down respectively), and the sum remains zero (that is why we need to order the data). We then find the difference between those two means. We do not impose any distributional assumptions and hence use the wild bootstrap tests. If that difference is significant, then we can conclude that the first half is different from the second one and there is a trend. We do these steps on the original sample and repeat that on 200 wild bootstrap samples. The proportion of times that the difference in the wild bootstrap samples is bigger than the difference in the real one is the estimated p-value. If the p-value is less than 0.05 then we can conclude there is something going on and we should reject the assumed model.

Test 2

This test contrasts the middle third with the other thirds. Because it could be that there is a pattern but not a trend, for example it could be up and then down. So we split the x-axis into three parts and calculate the difference between the means of them (see the blue lines in Figure 7.4). To test that, we multiply the middle part by 2 and the other two thirds by -1. That is also a contrast because the sum is zero. We do the same on the 200 wild bootstrap samples and calculate the difference on each. To decide whether the model is correct or not, we calculate the p-value by comparing the 200 bootstrap differences with the difference in the real sample.

Test 3

When the pattern does not have to be a major trend, but could be up or down in the left-hand tail, we should compare the first sixth with the second sixth (see the left hand

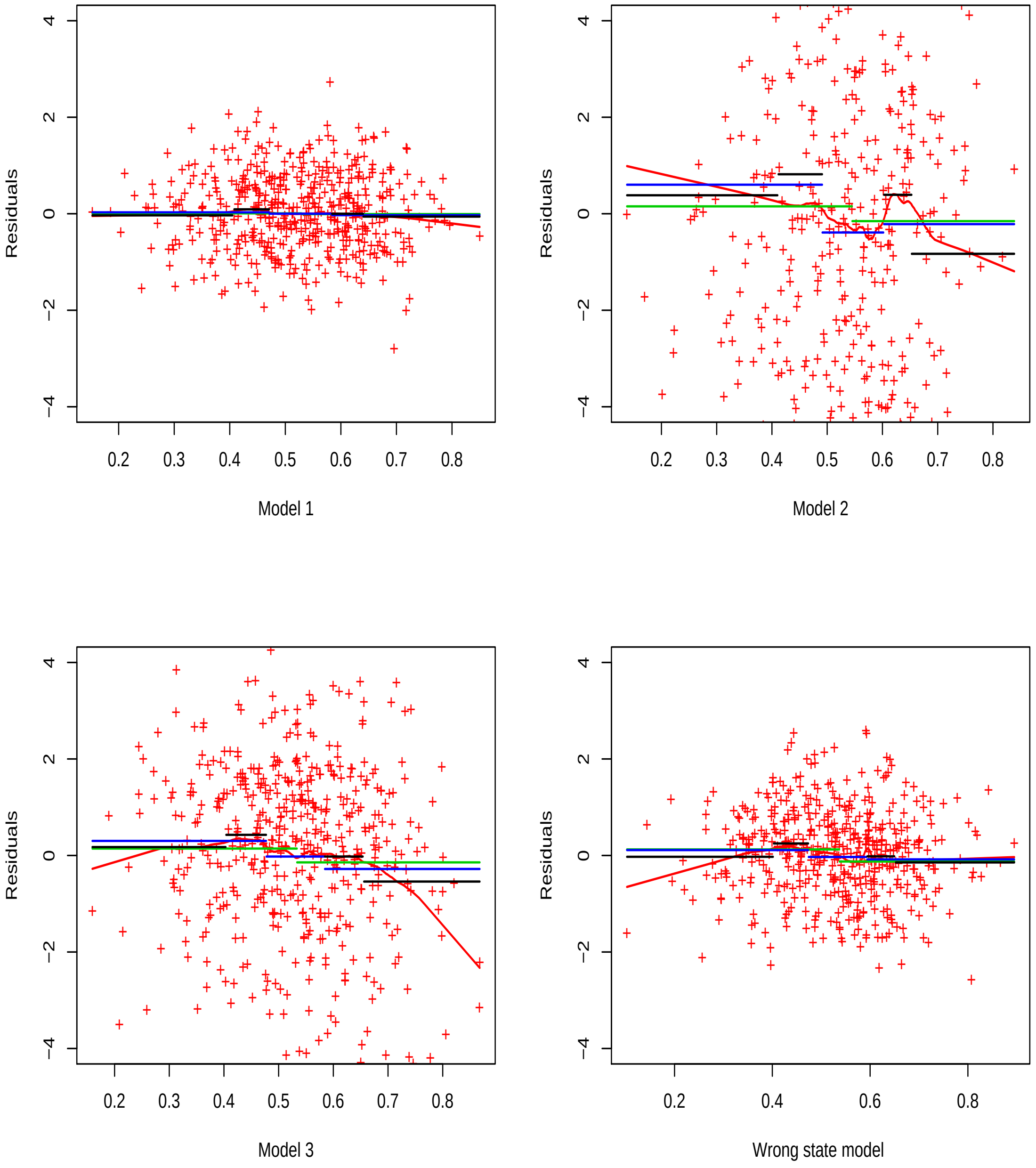

Figure 7.4: Different tests of the mean residuals against states of a dataset size of 500.

black lines in Figure 7.4). We split the first third of x-axis into two parts, then we test the difference as we did before.

Test 4

The trend could in the right-hand tail. The test compares the fifth sixth with the last sixth. So we split the last third of x-axis into two parts (see the right hand black lines in Figure 6.5) then we calculate the difference on each sixth to conclude the result of the test if significant or not.

Test 5

This is based on the maximum deviation from zero of the cumulative sum of residuals. This is no longer a contrast, but is investigated for completeness. We calculate that in the real dataset then in all bootstrap samples and compare the difference.

To see whether we have an evidence to reject the null hypothesis or not, the p-value can be calculated by how often the maximum deviation of the wild bootstrap samples is bigger than the original one. If the model is correct, then the proportion of rejections of the assumed model should be by chance equal about 0.05.

7.2.2 Simulation results

As shown in Figure 7.4, for Model 1, it is clear that all the means of each test are at the same horizontal level. It indicates that Model 1 could be correct, but other models are not. To test the wild bootstrap we used simulations of 500 datasets of size 500. Tables 7.2-7.4 show the proportions of rejection of the null hypothesis at time points 1, 5 and 9. Tests use residuals against states or regrets.

First time point

The first time point table shows that in Model 1, most of the proportions of rejections of the null hypothesis by chance are around 0.05. We found very similar results when using the wild bootstrap with either $Z \sim U\{-1, 1\}$ or $Z \sim N(0, 1)$. Against both states or regrets, the model is approximately correct, except a few particular cases, e.g., 0.068 in test three using standard normal against regrets, and 0.066 in test two using $U\{-1, 1\}$

Model 1		Test 1	Test 2	Test 3	Test 4	Test 5
$U\{-1, 1\}$	State	0.050	0.034	0.038	0.034	0.036
	Regrets	0.050	0.066	0.040	0.050	0.030
$Z(0, 1)$	State	0.040	0.042	0.050	0.058	0.052
	Regrets	0.040	0.058	0.068	0.060	0.050
Model 2						
$U\{-1, 1\}$	State	0.098	0.078	0.064	0.092	0.036
	Regrets	0.610	0.372	0.108	0.634	0.440
$Z(0, 1)$	State	0.120	0.072	0.038	0.098	0.044
	Regrets	0.672	0.372	0.092	0.578	0.422
Model 3						
$U\{-1, 1\}$	State	0.146	0.042	0.060	0.082	0.262
	Regrets	0.260	0.086	0.056	0.83	0.432
$Z(0, 1)$	State	0.152	0.064	0.054	0.100	0.266
	Regrets	0.252	0.076	0.056	0.806	0.398
Model 4						
$U\{-1, 1\}$	State	0.144	0.054	0.062	0.058	0.202
	Regrets	0.054	0.058	0.072	0.064	0.116
$Z(0, 1)$	State	0.120	0.038	0.060	0.076	0.230
	Regrets	0.058	0.036	0.072	0.082	0.108

Table 7.2: Estimated p-values on different models using first time point.

against regrets are a little higher. To make our decision, let us put a range of estimated p-value that we would get by chance if the model is correct, i.e.,

$$P \mp 2\sqrt{\frac{P \times (1-P)}{NS}} = 0.05 \mp 2\sqrt{\frac{0.05 \times 0.95}{500}} = [0.030, 0.069],$$

where NS is the number of the planned simulations. Hence values of 0.068 or 0.066 are not particularly unusual.

Fifth time point

Model 1		Test 1	Test 2	Test 3	Test 4	Test 5
$U\{-1,1\}$	State	0.082	0.056	0.040	0.060	0.068
	Regrets	0.058	0.046	0.054	0.060	0.050
$Z(0,1)$	State	0.076	0.066	0.054	0.044	0.056
	Regrets	0.048	0.059	0.045	0.068	0.055
Model 2						
$U\{-1,1\}$	State	0.458	0.09	0.158	0.08	0.17
	Regrets	0.810	0.358	0.160	0.474	0.402
$Z(0,1)$	State	0.394	0.098	0.172	0.068	0.160
	Regrets	0.824	0.370	0.178	0.460	0.400
Model 3						
$U\{-1,1\}$	State	0.766	0.054	0.098	0.198	0.890
	Regrets	0.112	0.216	0.054	0.982	0.902
$Z(0,1)$	State	0.728	0.050	0.100	0.206	0.874
	Regrets	0.098	0.246	0.074	0.982	0.922
Model 4						
$U\{-1,1\}$	State	0.090	0.082	0.040	0.072	0.148
	Regrets	0.058	0.056	0.038	0.076	0.106
$Z(0,1)$	State	0.106	0.064	0.046	0.092	0.148
	Regrets	0.064	0.036	0.036	0.084	0.104

Table 7.3: Estimated p-values on different models using fifth time point.

Table 7.3 repeats for time point five. Tests sizes under Model 1 are generally good and there are mixed powers for miss-specified models.

Ninth time point

In the ninth time point, results of Model 1 using regrets are similar as the previous time points. Now all the values of Model 1 are within the range.

Model 1		Test 1	Test 2	Test 3	Test 4	Test 5
$U\{-1,1\}$	State	0.074	0.056	0.06	0.044	0.068
	Regrets	0.048	0.064	0.062	0.078	0.056
$Z(0,1)$	State	0.068	0.064	0.052	0.060	0.072
	Regrets	0.052	0.058	0.060	0.082	0.060
Model 2						
$U\{-1,1\}$	State	0.42	0.074	0.174	0.064	0.164
	Regrets	0.832	0.388	0.138	0.494	0.416
$Z(0,1)$	State	0.456	0.088	0.170	0.062	0.152
	Regrets	0.694	0.058	0.086	0.186	0.848
Model 3						
$U\{-1,1\}$	State	0.72	0.058	0.072	0.172	0.872
	Regrets	0.086	0.222	0.062	0.972	0.914
$Z(0,1)$	State	0.668	0.064	0.086	0.186	0.848
	Regrets	0.112	0.258	0.052	0.964	0.890
Model 4						
$U\{-1,1\}$	State	0.082	0.084	0.046	0.062	0.140
	Regrets	0.072	0.046	0.064	0.086	0.114
$Z(0,1)$	State	0.058	0.070	0.044	0.078	0.122
	Regrets	0.052	0.064	0.042	0.076	0.100

Table 7.4: Estimated p-values on different models using ninth time point.

Hence we can conclude that model is fine. Models 2 and 3 are miss-specified fit models that we generate data using these models but we fit Model 1.

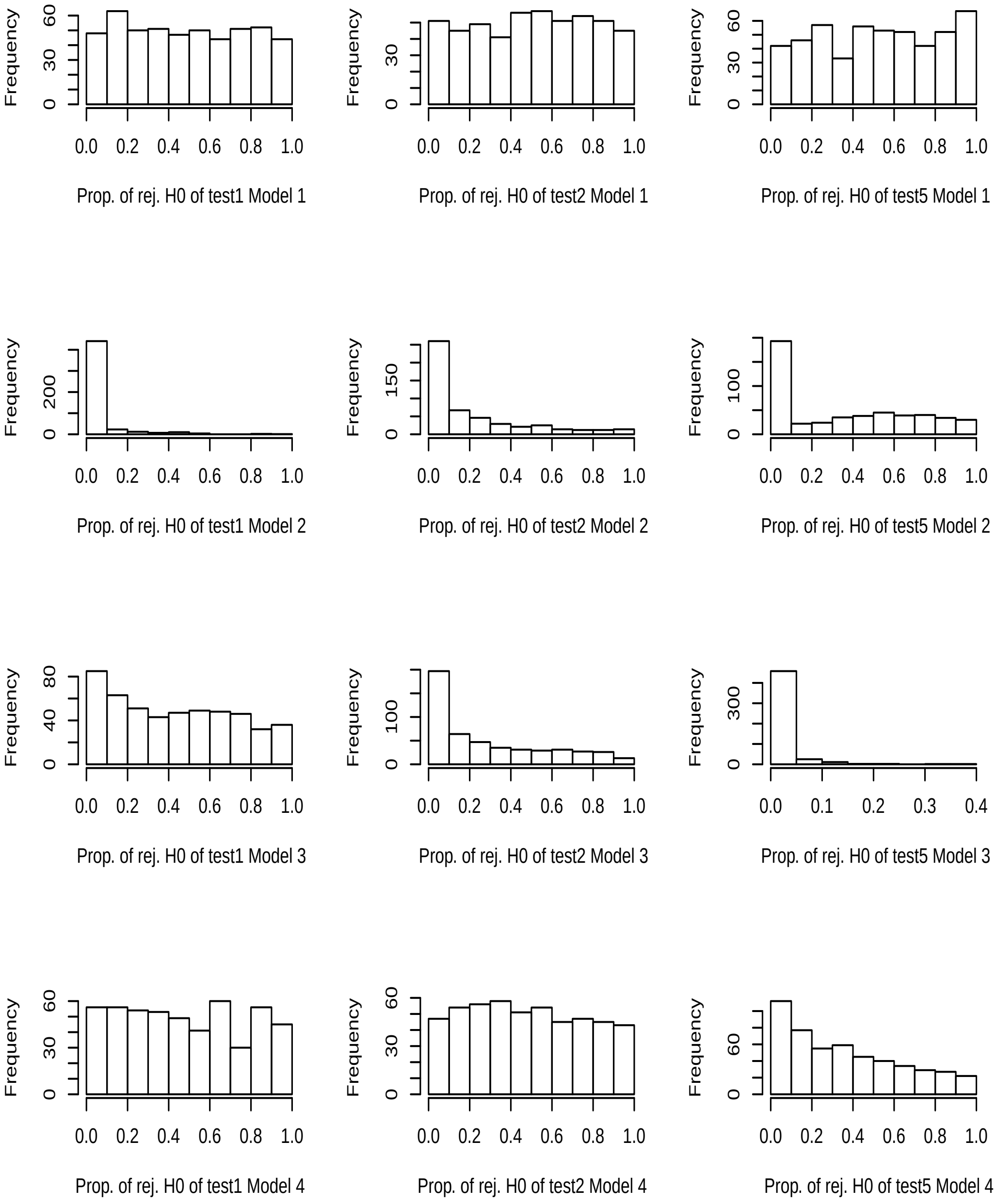

Figure 7.5: Histograms of p-value for model comparison

Therefore, we observe most other proportions are too far from 0.05 when using tests against states, with either $Z \sim U\{-1, 1\}$ or $Z \sim N(0, 1)$. Except, in Model 4, there are a few of proportions around 0.06 or 0.07. We also plot the p-value for the tests against regrets. The chosen simulations use the fifth time point. Columns 1-3 are histograms of p-value for tests 1, 2 and 5 respectively. The aim is to see whether the p-values are uniformly distributed or not. Figure 7.5 is an example to summarise the simulation results of 500 datasets of size 500. We test residuals against regrets using the wild bootstrap $U\{-1, 1\}$. It is clear that the histograms for Model 1 seem uniform, but not others.

According to the previous results, we can conclude that

- We get similar results with either $Z \sim U\{-1, 1\}$ or $Z \sim N(0, 1)$.
- Results using Test 1 and Test 5 are better than the other tests.
- Testing against regrets is better for Model 1, Model 2 and Model 3.
- For Model 4, it is better is to test against states.

7.3 Likelihood Ratio Test

We consider whether it is realistic to develop a diagnostic test for marginal structural models (MSMs) of Robins, see Hernán et al. (2002). First we provide some background. Suppose we have a classical regression model:

$$Y = \beta_0 + \beta_1 M_{i1} + \cdots + \beta_k M_{ik} + \epsilon_i,$$

with independent normal errors. The F test can be used to test the model. It involves estimating both (R) the restricted model (based on fewer parameters) and (U) the unrestricted model (the full model) and then calculating

$$\frac{(RSS_R - RSS_U)/M}{RSS_U/(N-K)} \sim F_{M,N-K}.$$

Here RSS is the sum of squared of residuals under the appropriate model, N is sample size, K the number of parameters in the U model and M parameters are deleted from the U model to form of the R model.

The Likelihood Ratio Test (LRT), the Wald Test (WT) and the Lagrange Multiplier Test (LMT), are frequently proposed as alternative means for testing parametric restrictions in a linear regression model. Since they all converge to the same limiting χ^2 distribution, they have asymptotically the same power characteristics. The three tests are considered equivalent alternatives for large samples (Kohler, 1982).

The LMT and WT involves the estimation of both the restricted and unrestricted models and a comparison of the values of their sum of squared residuals. They can be written as,

$$LMT = N\left[\frac{RSS_R - RSS_U}{RSS_R}\right]$$

and

$$WT = N\left[\frac{RSS_R - RSS_U}{RSS_U}\right].$$

The LRT involves the estimation of both the restricted and unrestricted models and a comparison of the values of the log likelihoods. If the difference is large, then we can reject the restrictions, otherwise we accept them.

The likelihood ratio statistic is

$$LRT = 2(l_U - l_R),$$

where

$$l_U = -\frac{N}{2}\ln(2\pi \times RSS_U/n) + 1$$

and

$$l_R = -\frac{N}{2}\ln(2\pi \times RSS_R/n) + 1,$$

leading to

$$LRT = N[\ln(RSS_U) - \ln(RSS_R)].$$

Our aim is to test an independence by comparing log likelihoods estimation of both the restricted and unrestricted models without put any assumptions. The comparison of log-likelihood values reflects the relative goodness-of-fit between two nested models. The observed difference indicates the effectiveness of the simpler model relative to the more complex model. When a parameter value or a set of parameter values is removed from a model and the log-likelihood value remains essentially unaffected (only a slight increase), the inference is made that the values eliminated are unimportant and likely have only random influences (Kohler, 1982).

The three tests have a χ^2 distribution with M degrees of freedom, because M parameters are deleted from the more complex model to form a simpler and, as required, a nested model. However, Savin (1976), has shown that the numerical value of LMT test statistic is always less than that of LRT, which is less than that of WT. This implies that if we use the same critical values, indicated by the fact that all three tests converge to the same limiting χ^2 distribution, the Wald test will reject the null hypothesis most often. This difference in rejection probabilities raises the possibility of conflicting conclusions from the three tests.

7.3.1 Independence Test

Returning to Chapter 3 Section 3.1, let us recall the formula,

$$H_j(\psi) = Y + \sum_{i=j}^{k} \mu_i(t_i|\bar{M}_i, \bar{T}_{i-1}),$$

where

$$\mu_j(t_j|\bar{M}_j, \bar{T}_{j-1}) = E(Y \mid \bar{M}_j, \bar{T}_{j-1}, \underline{d}_j^{opt}) - E(Y \mid \bar{M}_j, \bar{T}_{j-1}, t_j, \underline{d}_{j+1}^{opt}),$$

which is the expected difference between the average outcome if a patient received t_j instead of the optimal at time j, with given treatment and covariate history to time $j-1$ and who is subsequently treated optimally after time j.

As seen in Chapter 3, $H_j(\psi_{true})$ and T_j are conditionally independent given $\bar{M}_j, \bar{T}_{j-1}$. To test that, using LRT, let us assume the following hypothesis,

H_0: H_j and $T_j|\bar{M}_j, \bar{T}_{j-1}$ are independent,

H_a: H_j and $T_j|\bar{M}_j, \bar{T}_{j-1}$ are not independent.

We will fit H_j models, including and excluding $T_j|\bar{M}_j, \bar{T}_{j-1}$ and make a comparison of the LRT value with a χ^2 statistic for rejecting the null hypothesis or not. In this exploratory approach we will make a working assumption of normality for H_j, relying on the central limit theorem for at least partial support.

7.3.2 Results

In this section and the next two sections, we estimate the parameters of the model described in Section 7.1.2 using random samples of size 500 to examine whether $H_j(\hat{\psi})$ is independent of T_j or not. Then we assume the convergence of the likelihood ratio statistic to the assumed

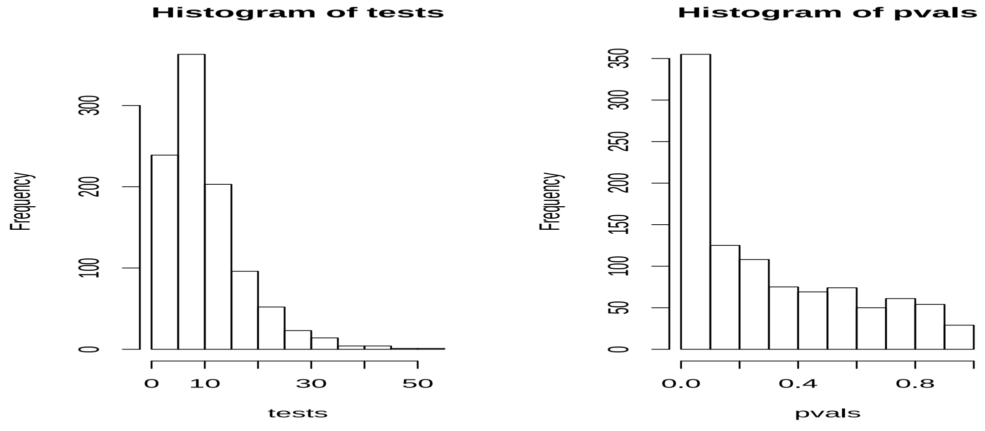

Figure 7.6: Histogram of test statistics and p-value using estimated ψ

distributions. Figure 7.6 shows histograms of the likelihood ratio statistics and associated p-values for 1000 simulated samples of size 500. If the test is performing correctly, the right hand histogram should be close to uniform. Clearly it is not, since there are too many small values. The proportion of p-values less than 0.05 is 0.27, indicating the χ^2 approximation to the null distribution is very poor.

We repeated the test using WT and LMT tests, and obtained similar results. The p-values are not uniformly distributed on the interval [0,1].

7.3.3 Using the true ψ values

Figure 7.7 repeats the procedure using the true ψ values instead the estimated ψ values. The p-values now form a more or less uniformly distributed pattern. The proportion of p-values less than 0.05 is 0.045. It seems performance of this test has very high sensitivity to the estimator's value even when they are very close to the true values.

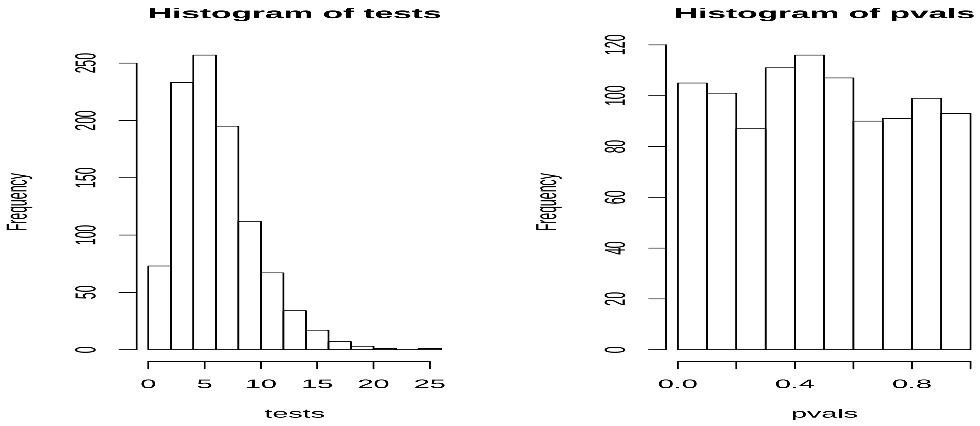

Figure 7.7: Histogram of test statistics and p-value using true parameters of ψ

7.3.4 Using subsets of mixed estimator and true ψ values

We now investigate which parameter elements of ψ seem to have most influence on the test performance. We have seen that the test performs poorly if all ψ are estimated, and well if all are fixed at their true values. We used different sub-sets of estimated and true ψ values. We found that any of estimators of ψ_1, ψ_2, ψ_3 and ψ_6 can be used but we must use ψ_4, ψ_5, ψ_7 and ψ_8 only as the true values. The results are shown in the next figure, where the histogram of p-values is uniformly distributed on the interval [0, 1].

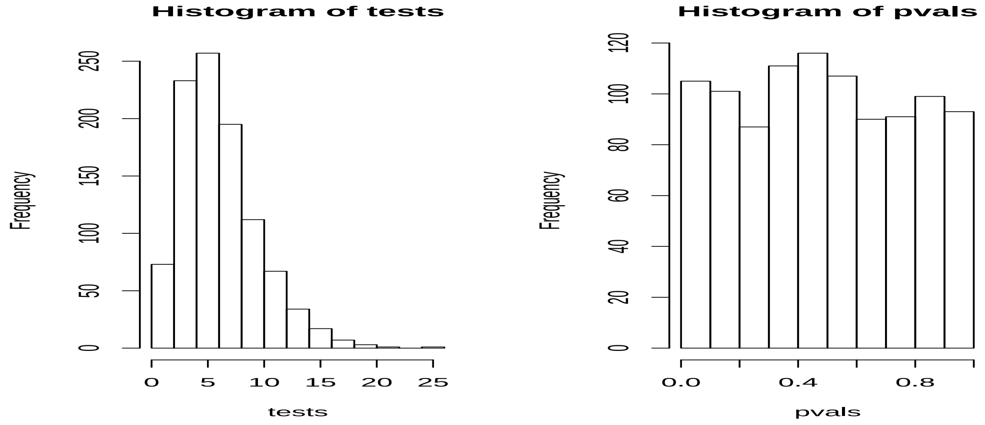

Figure 7.8: Histogram of test statistics and p-value using true values of $\{\psi_4, \psi_5, \psi_7, \psi_8\}$ and estimated values of $\{\psi_1, \psi_2, \psi_3, \psi_6\}$

Now suppose that we fix any three of these four true values and vary the fourth one. Figure 7.9 summarises a simulation of 100 datasets. We use the estimators of $\{\hat{\psi}_1, \hat{\psi}_2, \hat{\psi}_3, \hat{\psi}_6\}$ and the true values of $\{\psi_5, \psi_7, \psi_8\}$ and vary ψ_4 over the interval [1.3,1.7]. The figure shows how the probability of rejecting H_0 approaches to 0.05 only when the value of ψ_4 comes close to its true value (1.5). It goes far from 0.05 as the absolute difference between the true and the used value increases.

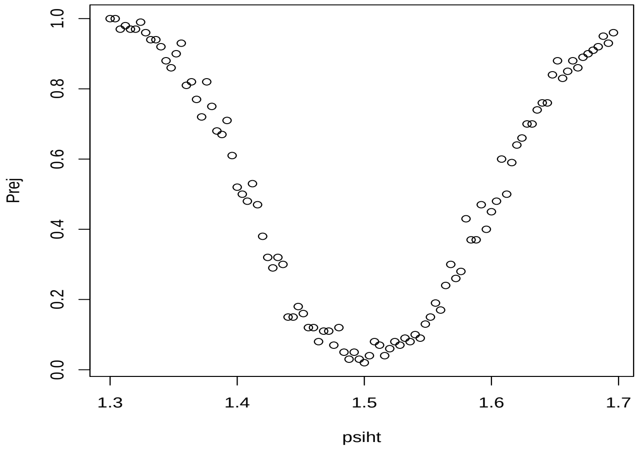

Figure 7.9: Probability of rejecting H_0 using varying values of ψ_4.

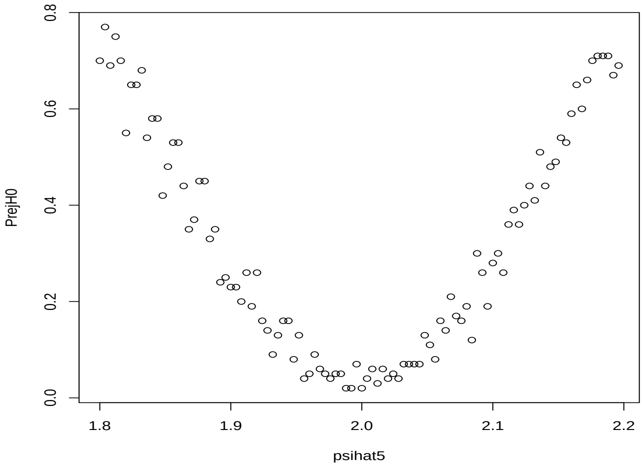

Figure 7.10: Probability of rejecting H_0 using varying values of ψ_5.

Figure 7.10 repeated this with ψ_5 selected as the variable parameter and with similar results. Recall the regret function for treatment j part two,

$$\mu_{j2}(t_{j2}|\bar{M}_j, \bar{T}_{j1}) = \psi_4 T_{j1}\{t_{j2} - (\psi_3 + \psi_5 M_j)\}^2 + \psi_7(1 - T_{j1})\{t_{j2} - (\psi_6 + \psi_8 M_j)\}^2.$$

Similar plots to Figure 7.9 can be obtained by varying ψ_7 and ψ_8 rather than ψ_4 and ψ_5, because of the symmetry of the regret function. Our conclusion therefore is that the proposal independence test is unreliable and should not be pursed farther.

7.4 Diagnostics for Warfarin Data

In this section we will test residuals of the M2 model used by Rosthøj et al (2006), but as shown we improve it by re-estimating its parameters using the regret-regression method. Then we will compare results with M7 which is the best model according to results as we described before. We can use here the same diagnostic plots of residuals against states which we used in Chapter 4, Section 4.3.3, but now we plot each residual for each time points (9 times) rather than pooling. Then we test these models by using the same five tests used in Section 7.2.1 in the current chapter and using a conditional multiplier (wild bootstrap) with random multiplier $N(0, 1)$ or $U\{-1, 1\}$.

7.4.1 Comparing warfarin models M2 vs. M7 using conditional multiplier tests.

Figure 7.11 shows results of testing the first half versus the second half using states at second and ninth time points to illustrate. The left plots use M2 and the right plots use M7. The upper plots use second time point and the lower plots use ninth time point. The plots test the difference between the two halves of the original sample (the blue horizontal lines). We need to see whether that difference is significant or not.

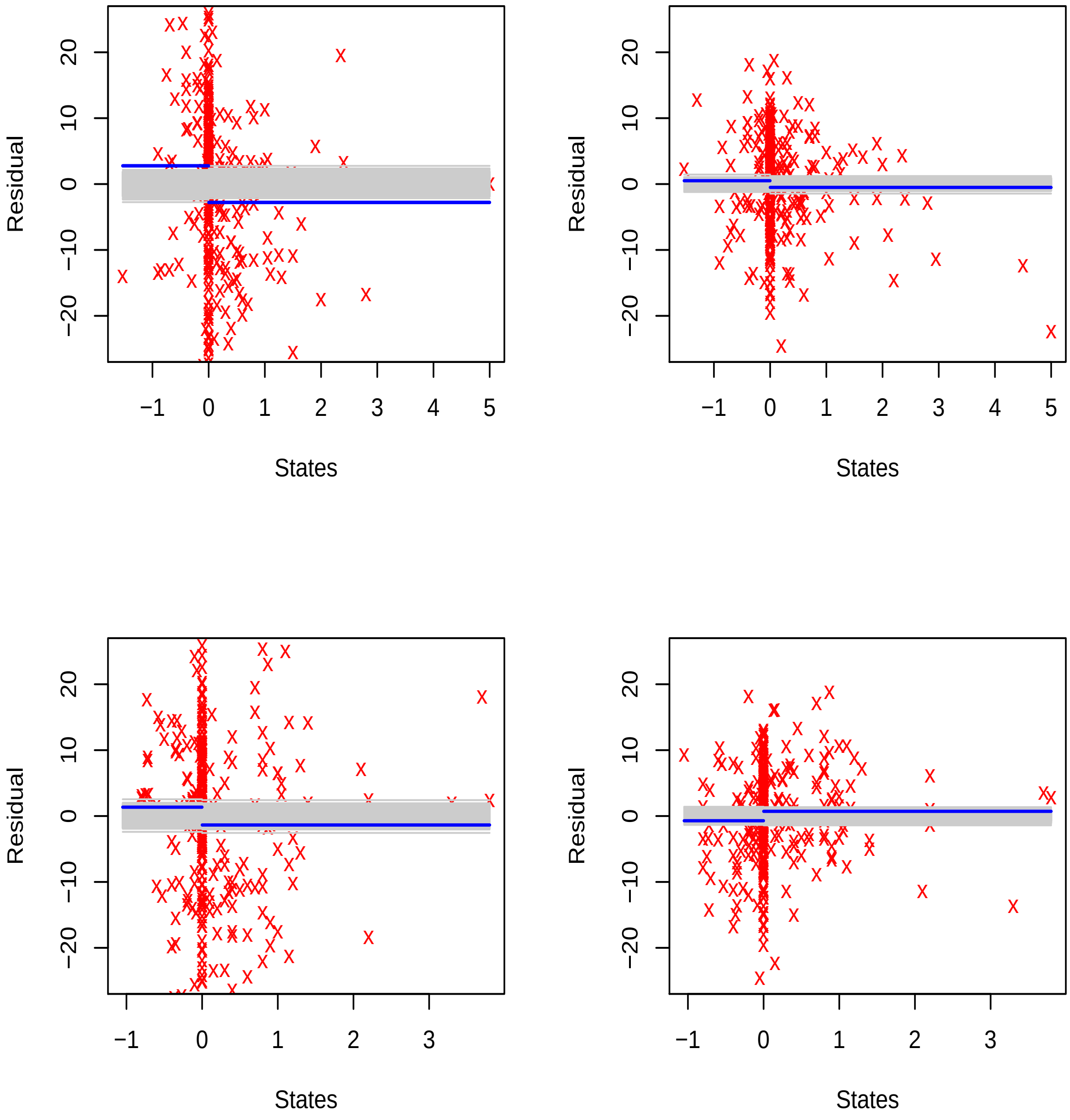

Figure 7.11: Comparing tests of models (M2 and M7) at second (first row) and ninth (second row) time points using 1000 wild bootstrap samples. Left colmn M2, right colmn M7.

	Model		Test 1	Test 2	Test 3	Test 4	Test 5
First time point							
	M2	$U\{-1,1\}$	0.035	0.028	0.000	0.452	0.013
		$Z(0,1)$	0.049	0.030	0.000	0.461	0.017
	M7	$U\{-1,1\}$	0.913	0.171	0.160	0.351	0.308
		$Z(0,1)$	0.900	0.150	0.157	0.377	0.298
Second time point							
	M2	$U\{-1,1\}$	0.000	0.177	0.099	0.980	0.000
		$Z(0,1)$	0.000	0.174	0.097	0.977	0.000
	M7	$U\{-1,1\}$	0.209	0.779	0.262	0.532	0.118
		$Z(0,1)$	0.247	0.743	0.305	0.533	0.137
Third time point							
	M2	$U\{-1,1\}$	0.011	0.391	0.607	0.029	0.005
		$Z(0,1)$	0.015	0.386	0.612	0.034	0.011
	M7	$U\{-1,1\}$	0.766	0.330	0.615	0.879	0.561
		$Z(0,1)$	0.736	0.360	0.587	0.890	0.504
Fourth time point							
	M2	$U\{-1,1\}$	0.000	0.964	0.046	0.775	0.000
		$Z(0,1)$	0.001	0.961	0.064	0.791	0.000
	M7	$U\{-1,1\}$	0.382	0.520	0.925	0.391	0.259
		$Z(0,1)$	0.390	0.516	0.918	0.399	0.225
Fifth time point							
	M2	$U\{-1,1\}$	0.002	0.059	0.702	0.116	0.000
		$Z(0,1)$	0.004	0.053	0.712	0.102	0.001
	M7	$U\{-1,1\}$	0.548	0.996	0.238	0.883	0.622
		$Z(0,1)$	0.545	0.993	0.221	0.882	0.604
Sixth time point							
	M2	$U\{-1,1\}$	0.060	0.280	0.184	0.141	0.011
		$Z(0,1)$	0.053	0.291	0.177	0.134	0.009
	M7	$U\{-1,1\}$	0.818	0.949	0.076	0.450	0.733
		$Z(0,1)$	0.805	0.950	0.075	0.435	0.736
Seventh time point							
	M2	$U\{-1,1\}$	0.000	0.195	0.929	0.696	0.000
		$Z(0,1)$	0.000	0.191	0.948	0.647	0.001
	M7	$U\{-1,1\}$	0.198	0.549	0.940	0.721	0.204
		$Z(0,1)$	0.200	0.548	0.937	0.720	0.195
Eighth time point							
	M2	$U\{-1,1\}$	0.000	0.625	0.415	0.818	0.001
		$Z(0,1)$	0.000	0.611	0.416	0.806	0.000
	M7	$U\{-1,1\}$	0.107	0.994	0.929	0.640	0.106
		$Z(0,1)$	0.120	0.998	0.936	0.599	0.136
Ninth time point							
	M2	$U\{-1,1\}$	0.063	0.289	0.821	0.313	0.049
		$Z(0,1)$	0.060	0.265	0.792	0.276	0.048
	M7	$U\{-1,1\}$	0.083	0.487	0.354	0.623	0.923
		$Z(0,1)$	0.100	0.466	0.371	0.635	0.901

Table 7.5: Estimated p-values for models M2 and M7 using warfarin data

Grey lines denote the means of the first and the second halves using the wild bootstrap method with $Z \sim N(0,1)$. If the blue lines are not included the grey lines range then we can say the model is not valid. The upper left plot shows that the blue lines are out of the grey lines range. To be sure we can use p-values (Table 7.5). The p-values for Test 1 with Model M2 are almost all very low, indicating the model is not appropriate as a residual trend remains. Some of the test 2 and test 3 results are also significant, indicating some quadratic effects or a problem in the left tail. Test 4 however does not detect any problem with Model M2 in the right tail. Test 5, the cumulative sum test, is consistent in suggesting Model M2 should be rejected at all time points. None of the tests indicate any problem with Model M7, our final selection.

7.4.2 Discussion

As shown model misspecification can affect the conclusions of a dynamic treatment analysis. We illustrated that through Murphy models by comparing her misspecifed models with her correct model. The chapter includes some diagnostic residuals plots. We do not impose any distributional assumptions and hence use the wild bootstrap tests. We investigated use of a likelihood ratio statistic for testing the independence of the $H_j(\psi)$ function from the actions $\underline{T}_j$, but do not recommend this. More investigation of our application on the anticoagulation data was performed. Residual plots and tests were used to compare Model M2 (Rosthøj et al 2006) after we improved it by re-estimating the parameters using the regret-regression method, and our developed Model M7. Hence M7 gives closed estimated values to the true model parameters, then optimal treatment strategies can be obtain. Finally we can conclude that the regret-regression method provides an important advantage of other methods is that we can compare a variety of candidate models then choose the best one to estimate the optimal dynamic treatment regimes.

Chapter 8

Conclusion and Further Developments

8.1 Conclusion

This dissertation explores problems of optimal dynamic treatment regimes. The most important problem is choosing an appropriate model when using any semi-parametric method. Blips with Robins G-estimation method or regrets with Murphy iterative minimization method produce apparently unbiased estimators, with less efficiency when using non appropriate models. Each of these methods and the relation between them are investigated in Chapter 3 after providing background on causal effects and dynamic programming as basics of our research area. In Chapter 4 we proposed a method which we hope will assist in the practical implementation of methodology for finding optimal dynamic treatment regimes. Almirall et al (2009), have a similar aim. Like us, they include residuals Z_j in a model for the observed mean of Y, but unlike us they assume the remaining terms (our regrets, their blips) are also linear and there is no discussion of diagnostics. Both methods require modelling of $E(M_j|\bar{M}_{j-1}, \bar{T}_{j-1})$, which is not necessary if either the Murphy (2003) iterative estimation procedure or the Robins (2004) G-estimation method is used. On the other hand we have not also required $Pr(t_j|\bar{M}_j, \bar{T}_{j-1})$, which is needed for the other methods. Inclusion of the linear combination of residuals in Equation 4.3 is reminiscent of a path analysis method for dealing with time-dependent confounders, as exemplified by Borgan et al (2006) for instance. This brings additional modelling assumptions, not needed by Murphy or Robins, and hence the possibility of misspecification. If the model is correct however, or close to correct, then we expect gains in efficiency, as seen in the simulation study of Section 3.3 in Chapter 4.

Murphy (2003) had primary interest in the parameters of the regrets, and considered other unknown functions involved in data generation as nuisance parameters. We agree in part only: unless sample size is enormous we see no alternative to assuming parametric models for all components. In that case some form of diagnostic is essential for good statistical practice, and development of diagnostics based on models for observables is an obvious way forward. The use of residual plots to detect misspecified regret functions was illustrated in Chapter 4, Section 3.3. Chapter 5 investigated a comparison between the regret-regression approach and inverse probability of treatment weighting. Although they are different methods both have had the same results. We showed that regret-regression and inverse probability of treatment weighting are equal in some cases at least (see proof in Appendix 9.1). Then we prepared a chapter for applying regret-regression to decision making on the multi-armed bandit problem. The chapter compared the regret-regression policy for solving that problem with other policies using some traditional approaches, such as Gittins index, Q-learning etc. Some diagnostic methods were used in Chapter 7, to propose and investigate goodness of fit tests for instance, to understand more how residual means react to different types of misspecification, and to explore the bias/variance tradeoffs and the possibility of overfitting.

8.2 Further Developments

The area of dynamic treatment regimes is very young and exciting. Many challenges remain in the area of dynamic regime estimation. We have assumed a discrete time scale for state measurement and action decisions and that there is no censoring. These conditions apply to the warfarin data example which motivated our work, but in general clinic visits will take place at different times for different patients and might be thought of as a point process in continuous time, with differential follow-up. It will be of interest to try to develop our methods for this situation, perhaps with the counting process approach of Lok (2008) as a starting point. Further, we have not incorporated covariates. In principle we can include covariate terms in our parameterized regret functions and state models, but clearly it would be useful to test this approach in practice, especially as there is much current interest in genetic variation in anticoagulant response (Schwarz et al, 2008). On the other hand if M_j is high dimensional our modelling may become problematic and needs to be tested. We have not seen application on real data with high dimensional M_j of any of the proposed methods in this literature.

Chapter 9

Appendix

9.1 Proof that the regret-regression and the inverse probability of treatment weighting are identical.

Our aim in this section is to prove that the regret-regression and the inverse probability of treatment weighting are identical, through the same calculated regret at time 1 in the particular example described in Chapter 4 of two time point situation and binary states and actions.

Let us recall the stabilized weights using the IPTW

$$SW = \frac{Pr[T_2 = t_2|T_1 = t_1]}{Pr[T_2 = t_2|M_2 = m_2, T_1 = t_1]}.$$

Let $SW_{t_1 m_1 t_2} = SW(T_1 = t_1, M_2 = m_1, T_2 = t_2)$ be the stabilized inverse probability of treatment weighted for $YI(t_1, m_1, t_2) = YI(T_1 = t_1, M_2 = m_1, T_2 = t_2)$ and let $\hat{\mu_1}$ be the regret when choosing the wrong action at the first time point, then

$$SW_{001} = \frac{\frac{\sum I(001)+\sum I(011)}{\sum I(0)}}{\frac{\sum I(001)}{\sum I(00)}}, \quad SW_{011} = \frac{\frac{\sum I(001)+\sum I(011)}{\sum I(0)}}{\frac{\sum I(011)}{\sum I(01)}},$$

$$SW_{100} \;=\; \frac{\frac{\sum I(100)+\sum I(110)}{\sum I(1)}}{\frac{\sum I(100)}{\sum I(10)}}, \quad SW_{110} = \frac{\frac{\sum I(100)+\sum I(110)}{\sum I(1)}}{\frac{\sum I(110)}{\sum I(11)}}$$

$$\hat{\mu_1} = \frac{\sum YI(001).SW_{001}+\sum YI(011).SW_{011}}{\sum I(001).SW_{001}+\sum I(011).SW_{011}} - \frac{\sum YI(100).SW_{100}+\sum YI(110).SW_{110}}{\sum I(100).SW_{100}+\sum I(110).SW_{110}}$$

$$= \frac{\frac{\sum I(001)+\sum I(011)}{\sum I(0)} \frac{\sum I(00)}{\sum I(001)} \sum YI(001) + \frac{\sum I(001)+\sum I(011)}{\sum I(0)} \frac{\sum I(01)}{\sum I(011)} \sum YI(011)}{\frac{\sum I(001)+\sum I(011)}{\sum I(0)} \frac{\sum I(00)}{\sum I(001)} \sum I(001) + \frac{\sum I(001)+\sum I(011)}{\sum I(0)} \frac{\sum I(01)}{\sum I(011)} \sum I(011)}$$
$$- \frac{\frac{\sum I(100)+\sum I(110)}{\sum I(1)} \frac{\sum I(10)}{\sum I(100)} \sum YI(100) + \frac{\sum I(100)+\sum I(110)}{\sum I(1)} \frac{\sum I(11)}{\sum I(110)} \sum YI(110)}{\frac{\sum I(100)+\sum I(110)}{\sum I(1)} \frac{\sum I(10)}{\sum I(100)} \sum I(100) + \frac{\sum I(100)+\sum I(110)}{\sum I(1)} \frac{\sum I(11)}{\sum I(110)} \sum I(110)}$$

Thus

$$\hat{\mu_1} = \frac{\sum I(00)\overline{YI(001)}+\sum I(01)\overline{YI(011)}}{\sum I(00)+\sum I(01)} - \frac{\sum I(10)\overline{YI(100)}+\sum I(11)\overline{YI(110)}}{\sum I(10)+\sum I(11)}$$

Now, regarding to regret-regression method, let us recall equation 4.2 in Chapter 4 Section 3

$$E[Y|\bar{M}_K, \bar{T}_K] = \beta_0(M_1) + \sum_{j=2}^{K} \beta_j^{(}\bar{M}_{j-1}, \bar{T}_{j-1})Z_j - \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}). \tag{9.1}$$

$$E[Y|\bar{M}_K, \bar{T}_K] = \beta_0(M_1) + \sum_{j=2}^{K} \beta_j^T(\bar{M}_{j-1}, \bar{T}_{j-1})Z_j - \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}).$$

The parameters are estimated by minimising

$$SS = \sum \left(E[Y|\bar{M}_K, \bar{T}_K] - \beta_0(M_1) - \sum_{j=2}^{K} \beta_j^T(\bar{M}_{j-1}, \bar{T}_{j-1})Z_j + \sum_{j=1}^{K} \mu_j(T_j|\bar{M}_j, \bar{T}_{j-1}) \right)^2.$$

this equation will be as follows

$$SS = \sum (Y - \beta_0 - \beta_1 I(0)(M_2 - \hat{p}_0) \;-\; \beta_2 I(1)(M_2 - \hat{p}_1) + \mu_1 I(1) + \mu_{00} I(000)$$

$$+ \quad \mu_{01}I(010) + \mu_{10}I(101) + \mu_{11}I(111))^2,$$

where $M_2 = I(01) + I(11)$, $\hat{p}_0 = \frac{\sum I(01)}{\sum I(0)}$, $\hat{p}_1 = \frac{\sum I(11)}{\sum I(1)}$ and $\hat{\beta}_0$ is the overall mean. The values of $\mu_{00}, \mu_{01}, \mu_{10}$ and μ_{11} are as follows

$$\begin{aligned}
\hat{\mu_{00}} &= \frac{\sum YI(001)}{\sum I(001)} - \frac{\sum YI(000)}{\sum I(000)} \\
&= \frac{\sum I(000)\sum YI(001) - \sum I(001)\sum YI(000)}{\sum I(000)\sum I(001)} \\
&= \frac{\sum I(000)[\sum YI(00) - \sum YI(000)] - [\sum I(00) - \sum I(000)]\sum YI(000)}{\sum I(000)[\sum I(00) - \sum I(000)]} \\
&= \frac{\sum I(000)\sum YI(00) - \sum I(00)\sum YI(000)}{\sum I(000)[\sum I(00) - \sum I(000)]}
\end{aligned}$$

By the same way we find the other μ's

$$\begin{aligned}
\hat{\mu_{01}} &= \frac{\sum YI(011)}{\sum I(011)} - \frac{\sum YI(010)}{\sum I(010)} \\
&= \frac{\sum I(010)\sum YI(01) - \sum I(01)\sum YI(010)}{\sum I(010)[\sum I(01) - \sum I(010)]}
\end{aligned}$$

$$\begin{aligned}
\hat{\mu_{10}} &= \frac{\sum YI(100)}{\sum I(100)} - \frac{\sum YI(101)}{\sum I(101)} \\
&= \frac{\sum I(10)\sum YI(100) - \sum I(100)\sum YI(10)}{\sum I(100)[\sum I(10) - \sum I(100)]}
\end{aligned}$$

$$\begin{aligned}
\hat{\mu_{11}} &= \frac{\sum YI(110)}{\sum I(110)} - \frac{\sum YI(111)}{\sum I(111)} \\
&= \frac{\sum I(11)\sum YI(110) - \sum I(110)\sum YI(11)}{\sum I(110)[\sum I(11) - \sum I(110)]}
\end{aligned}$$

$$\begin{aligned}
\text{Let } y &= Y + \mu_{00}I(000) + \mu_{01}I(010) + \mu_{10}I(101) + \mu_{11}I(111). \quad \text{then} \\
SS &= \sum (y - \beta_0 - \beta_1 I(0)(M_2 - \hat{p}_0) - \beta_2 I(1)(M_2 - \hat{p}_1) + \mu_1 I(1))^2, \\
\text{and } \frac{\partial SS}{\partial \beta_0} &= -2\sum[y - \beta_0 - \beta_1 I(0)(M_2 - \hat{p}_0) - \beta_2 I(1)(M_2 - \hat{p}_1) + \mu_1 I(1)]. \\
\text{From this } \hat{\beta}_0 &= \overline{y} - \beta_1 \overline{I(0)(M_2 - \hat{p}_0)} - \beta_2 \overline{I(1)(M_2 - \hat{p}_1)} + \mu_1 \overline{I(1)},
\end{aligned}$$

$SS = \sum[\{y - \overline{y}\} - \beta_1\{I(0)(M_2 - \hat{p}_0) - \overline{I(0)(M_2 - \hat{p}_0)}\}$

$- \beta_2\{I(1)(M_2 - \hat{p}_1) - \overline{I(1)(M_2 - \hat{p}_1)}\} + \mu_1\{I(1) - \overline{I(1)}\}]^2.$

Note

$\sum\{I(0)[M_2 - \hat{p}_0]\} = \sum\{I(0)[I(01) + I(11) - \frac{\sum I(01)}{\sum I(0)}]\} = \sum\{I(01) - I(0)\frac{\sum I(01)}{\sum I(0)}\} = \sum I(01) - \sum I(01) = 0$

Similarly

$\sum\{I(1)[M_2 - \hat{p_1}]\} = 0$

This means $\overline{I(0)(M_2 - \hat{p}_0)} = \overline{I(1)(M_2 - \hat{p}_1)} = 0$, then

$SS = \sum\left(\{y - \overline{y}\} - \beta_1\{I(0)(M_2 - \hat{p}_0)\} - \beta_2\{I(1)(M_2 - \hat{p}_1)\} + \mu_1\{I(1) - \overline{I(1)}\}\right)^2.$

Now, let

$Y_1 = y - \overline{y}$, $X_1 = I(0)(M_2 - \hat{p}_0)$, $X_2 = I(1)(M_2 - \hat{p}_1)$ and $X_3 = I(1) - \overline{I(1)}$.

To estimate β_1, β_2 and μ_1, we have to minimises

$$SS = \sum(Y_1 - \beta_1 X_1 - \beta_2 X_2 + \mu_1 X_3)^2$$

The first partial derivative of SS with respect to each of β_1, β_2 and μ_1 are

$$\begin{aligned}
\frac{\partial SS}{\partial \beta_1} &= \sum[Y_1 - \beta_1 X_1 - \beta_2 X_2 + \mu_1 X_3]X_1 \\
\frac{\partial SS}{\partial \beta_2} &= \sum[Y_1 - \beta_1 X_1 - \beta_2 X_2 + \mu_1 X_3]X_2 \\
\frac{\partial SS}{\partial \mu_1} &= \sum[Y_1 - \beta_1 X_1 - \beta_2 X_2 + \mu_1 X_3]X_3
\end{aligned}$$

Now we will turn to matrix notation $(X^T X) = (\sum)\, X_1^2 \sum X_1 X_2 \sum X_1 X_3$

$\sum X_2X_1 \sum X_2^2 \sum X_2X_3$

$\sum X_3X_1 \sum X_3X_2 \sum X_3^2$

Note

$\sum X_1X_3 = \sum\{[I(0)(M_2 - \hat{p}_0)][I(1) - \overline{I(1)}] = -\overline{I(1)}\sum[I(0)(M_2 - \hat{p}_0)] = 0,$

and by similar argument $\sum X_2X_3 \;=\; 0.$

Thus $\qquad \hat{\mu_1} = -\frac{\sum Y_1X_3}{\sum X_3^2}$

Note $\sum I(1) = \sum I(10) + \sum I(11)$, and $n = \sum I(00) + \sum I(01) + \sum I(10) + \sum I(11)$. Hence

$$\begin{aligned}\sum X_3^2 &= \sum\left[I(1) - \overline{I(1)}\right]^2 = \sum I(1) - \overline{I(1)}^2 \\ &= \frac{1}{n}\left[n(\sum I(10) + \sum I(11)) - (\sum I(10) + \sum I(11))^2\right] \\ &= \frac{1}{n}\left[\sum I(10) + \sum I(11)) \times (\sum I(10) + \sum I(11))\right]]\end{aligned}$$

Note $\sum Y_1\overline{I(1)} = \overline{I(1)}\sum Y_1 = 0$,

so

$$\begin{aligned}\sum Y_1X_3 &= \sum Y_1(I(1) - \overline{I(1)}) = \sum Y_1I(1) \\ &= \sum(y - \overline{y})I(1),\end{aligned}$$

$$\begin{aligned}\sum Y_1X_3 = \sum[(Y - \overline{Y}) &+ \hat{\mu_{00}}(I(000) - \overline{I(000)}) + \hat{\mu_{01}}(I(010) - \overline{I(010)}) \\ &+ \hat{\mu_{10}}(I(101) - \overline{I(101)}) + \hat{\mu_{11}}(I(111) - \overline{I(111)})]I(1).\end{aligned}$$

We split this into 5 terms as follows

1-

$$\begin{aligned}\sum(Y-\overline{Y})I(1) &= \sum Y_1 I(1) - \sum \overline{Y} I(1)\\ &= \frac{1}{n}[n\sum Y_1(I(1) - \sum Y_1 \sum I(1)]\end{aligned}$$

$= \frac{1}{n}[\{\sum I(00) + \sum I(01) + \sum I(10) + \sum I(11)\}(\sum YI(10) + \sum YI(11)) - \{\sum YI(00) + \sum YI(01) + \sum YI(10) + \sum YI(11)\}(\sum I(10) + \sum I(11))]$

$= \frac{1}{n}[\{\sum I(00) + \sum I(01)\}(\sum YI(10) + \sum YI(11)) - \{\sum I(10) + \sum I(11)\}(\sum YI(00) + \sum YI(01))].$

2-

$\hat{\mu_{00}} \sum(I(000) - \overline{I(000)})I(1)$

$$\begin{aligned}&= -\mu_{00}\overline{I(000)}\sum I(1)\\ &= -(\tfrac{\sum I(000)\sum YI(00) - \sum I(00)\sum YI(000)}{\sum I(000)[\sum I(00) - \sum I(000)]})\tfrac{\sum I(000)}{n}(\sum I(10) + \sum I(11))\\ &= -(\tfrac{\sum I(00)\sum YI(000)\sum I(00) - \sum I(000)\sum YI(00)}{n\sum I(001})(\sum I(10) + \sum I(11))\\ &= (\tfrac{(\sum I(10) + \sum I(11))}{n\sum I(001)})[\sum I(00)\sum YI(000) - \sum I(000)\sum YI(00)].\end{aligned}$$

3-

$\hat{\mu_{01}} \sum(I(010) - \overline{I(010)})I(1)$

$$\begin{aligned}&= -\hat{\mu_{01}}\overline{I(010)}\sum I(1)\\ &= -(\tfrac{\sum I(010)\sum YI(01) - \sum I(01)\sum YI(010)}{\sum I(010)[\sum I(01) - \sum I(010)]})\tfrac{\sum I(010)}{n}(\sum I(10) + \sum I(11))\\ &= -(\tfrac{\sum I(01)\sum YI(010) - \sum I(010)\sum YI(01)}{n\sum I(011})(\sum I(10) + \sum I(11))\\ &= (\tfrac{(\sum I(10) + \sum I(11))}{n\sum I(011)})[\sum I(01)\sum YI(010) - \sum I(010)\sum YI(01)].\end{aligned}$$

4-

$$\hat{\mu_{10}} \sum (I(101) - \overline{I(101)}) I(1)$$

$$= -\hat{\mu_{10}} \overline{I(101)} \sum I(1)$$
$$= -(\tfrac{\sum I(10) \sum YI(100) - \sum I(100) \sum YI(10)}{\sum I(100)[\sum I(10) - \sum I(100)]}) \tfrac{\sum I(101)}{n} (\sum I(00) + \sum I(01))$$
$$= -(\tfrac{\sum I(10) \sum YI(100) - \sum I(100) \sum YI(10)}{n \sum I(101}) (\sum I(00) + \sum I(01))$$
$$= (\tfrac{(\sum I(00) + \sum I(01))}{n \sum I(100)}) [\sum I(10) \sum YI(100) - \sum I(100) \sum YI(10)].$$

5-

$$\hat{\mu_{11}} \sum (I(111) - \overline{I(111)}) I(1)$$

$$= -\hat{\mu_{11}} \overline{I(111)} \sum I(1)$$
$$= -(\tfrac{\sum I(11) \sum YI(110) - \sum I(110) \sum YI(11)}{\sum I(110)[\sum I(11) - \sum I(110)]}) \tfrac{\sum I(111)}{n} (\sum I(00) + \sum I(01))$$
$$= -(\tfrac{\sum I(00) \sum YI(000) - \sum YI(00) \sum I(000)}{n \sum I(001}) (\sum I(00) + \sum I(01))$$
$$= (\tfrac{(\sum I(00) + \sum I(01))}{n \sum I(110)}) [\sum I(11) \sum YI(110) - \sum I(110) \sum YI(11)].$$

So,

$\sum Y_1 X_3 = \frac{1}{n}[\{\sum I(00) + \sum I(01)\}(\sum YI(10) + \sum YI(11)) - \{\sum I(10) + \sum I(11)\}(\sum YI(00) +$
$\sum YI(01))$
$+ (\frac{(\sum I(10) + \sum I(11))}{n \sum I(001)}) [\sum I(00) \sum YI(000) - \sum I(000) \sum YI(00)]$
$+ (\frac{(\sum I(10) + \sum I(11))}{n \sum I(011)}) [\sum I(01) \sum YI(010) - \sum I(010) \sum YI(01)]$
$+ (\frac{(\sum I(00) + \sum I(01))}{n \sum I(100)}) [\sum I(10) \sum YI(100) - \sum I(100) \sum YI(10)]$
$+ (\frac{(\sum I(00) + \sum I(01))}{n \sum I(110)}) [\sum I(11) \sum YI(110) - \sum I(110) \sum YI(11)]].$

$= \frac{1}{n}[\{\sum I(00) + \sum I(01)\}[\sum YI(10) + \sum YI(11)$
$+ \frac{\sum I(10) \sum YI(100)}{\sum I(100)} - \frac{\sum I(100) \sum YI(10)}{\sum I(100)} + \frac{\sum I(11) \sum YI(110)}{\sum I(110)} - \frac{\sum I(110) \sum YI(11)}{\sum I(110)}]$

$$- \{\sum I(10) + \sum I(11)\}[\sum YI(00) + \sum YI(01)$$
$$- \frac{\sum I(00)\sum YI(000)}{\sum I(001)} + \frac{\sum I(000)\sum YI(00)}{\sum I(001)} - \frac{\sum I(01)\sum YI(010)}{\sum I(011)} + \frac{\sum I(010)\sum YI(01)}{\sum I(011)}]].$$

$$= \tfrac{1}{n}[\{\sum I(00) + \sum I(01)\}[\frac{\sum I(10)\sum YI(100)}{\sum I(100)} + \frac{\sum I(11)\sum YI(110)}{\sum I(110)}]$$
$$- \{\sum I(10) + \sum I(11)\}[\sum YI(000) + \sum YI(001) + \sum YI(010) + \sum YI(011)$$
$$- \frac{\sum I(000)\sum YI(000)}{\sum I(001)} - \frac{\sum I(001)\sum YI(000)}{\sum I(001)} + \frac{\sum I(000)\sum YI(000)}{\sum I(001)} + \frac{\sum I(000)\sum YI(001)}{\sum I(001)}$$
$$- \frac{\sum I(010)\sum YI(010)}{\sum I(011)} - \frac{\sum I(011)\sum YI(010)}{\sum I(011)} + \frac{\sum I(010)\sum YI(010)}{\sum I(011)} + \frac{\sum I(010)\sum YI(011)}{\sum I(011)}]].$$

$$= \tfrac{1}{n}[\{\sum I(00) + \sum I(01)\}[\frac{\sum I(10)\sum YI(100)}{\sum I(100)} + \frac{\sum I(11)\sum YI(110)}{\sum I(110)}]$$
$$- \{\sum I(10) + \sum I(11)\}[\sum YI(001) + \frac{\sum I(000)\sum YI(001)}{\sum I(001)} + \sum YI(011) + \frac{\sum I(010)\sum YI(011)}{\sum I(011)}]].$$

$$= \tfrac{1}{n}[\{\sum I(00) + \sum I(01)\}[\frac{\sum I(10)\sum YI(100)}{\sum I(100)} + \frac{\sum I(11)\sum YI(110)}{\sum I(110)}]$$
$$- \{\sum I(10) + \sum I(11)\}[\frac{\sum I(00)\sum YI(001)}{\sum I(001)} + \frac{\sum I(01)\sum YI(011)}{\sum I(011)}]].$$

$$\hat{\mu_1} = -\tfrac{\sum Y_1 X_3}{\sum X_3^2}$$

$$= \tfrac{(\sum I(10)+\sum I(11))[\sum I(00)\overline{YI(001)}+\sum I(01)\overline{YI(011)}]-(\sum I(00)+\sum I(01))[\sum I(10)\overline{YI(100)}+\sum I(11)\overline{YI(110)}]}{(\sum I(00)+\sum I(01))(\sum I(10)+\sum I(11))}$$

Thus

$$\hat{\mu_1} = \frac{\sum I(00)\overline{YI(001)}+\sum I(01)\overline{YI(011)}}{\sum I(00)+\sum I(01)} - \frac{\sum I(10)\overline{YI(100)}+\sum I(11)\overline{YI(110)}}{\sum I(10)+\sum I(11)}$$

This is the same of $\hat{\mu_1}$ using $IPTW$ formula, as required.

9.2 Additional plot and table of residuals against states

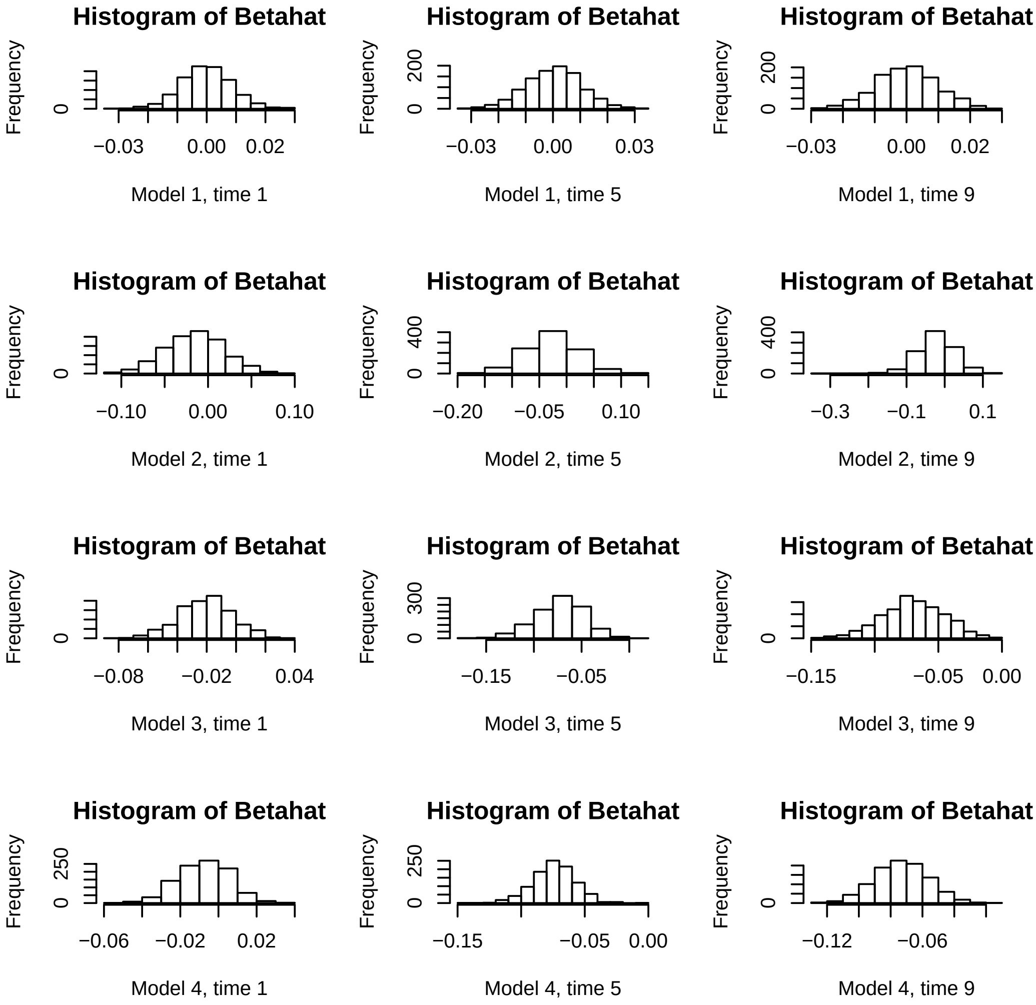

Figure 9.1: Histograms of $\hat{\beta}$ on different models using simulations of 1000 datasets each of size 500.

		Model 1	Model 2	Model 3	Model 4
Against states					
First time point	Mean	-0.0001	-0.0152	-0.0208	-0.0078
	Standard Deviation	0.008	0.033	0.018	0.013
	Median	-0.0002	-0.0149	-0.0201	-0.0077
	Quartile Deviation	0.005	0.022	0.012	0.009
Fifth time point	Mean	-0.0003	-0.0271	-0.0723	-0.0741
	Standard Deviation	0.010	0.046	0.025	0.016
	Median	-0.0007	-0.0271	-0.0708	-0.074
	Quartile Deviation	0.006	0.030	0.017	0.010
Ninth time point	Mean	0.0001	-0.0226	-0.0698	-0.0732
	Standard Deviation	0.009	0.048	0.024	0.016
	Median	0.0006	-0.0229	-0.0699	-0.0729
	Quartile Deviation	0.006	0.031	0.016	0.011

Table 9.1: Comparing different models through the slope residuals against states using simulations of 1000 datasets each of size 500, following the Murphy scenario.

Table 9.2: $E[H_3|\bar{M}_3, \bar{T}_2]$ for optimal outcome using G-estimation equation 2.2.

T_1	T_2	M_1	M_2	M_3	$E[H_2 \mid \bar{M}_2, \bar{T}_1]$
0	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1$
1	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1$
0	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1$
1	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1$
0	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1)$
1	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1)$
0	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1)$
1	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1)$
0	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{20} + \psi_{21}M_2)$
1	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{20} + \psi_{21}M_2)$
0	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{20} + \psi_{21}M_2)$
1	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{20} + \psi_{21}M_2)$
0	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1) - (\psi_{20} + \psi_{21}M_2)$
1	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1) + (\psi_{20} + \psi_{21}M_2)$
0	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1) + (\psi_{20} + \psi_{21}M_2)$
1	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$\geq -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1) - (\psi_{20} + \psi_{21}M_2)$
0	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{30} + \psi_{31}M_3)$
1	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{30} + \psi_{31}M_3)$
0	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{30} + \psi_{31}M_3)$
1	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{30} + \psi_{31}M_3)$
0	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1) - (\psi_{30} + \psi_{31}M_3)$
1	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1) - (\psi_{30} + \psi_{31}M_3)$
0	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1) - (\psi_{30} + \psi_{31}M_3)$
1	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1) - (\psi_{30} + \psi_{31}M_3)$
0	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$
1	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$
0	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$
1	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$
0	0	$< -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1) - (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$
1	1	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1) + (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$
0	1	$< -\frac{\psi_{10}}{\psi_{11}}$	$\geq -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 - (\psi_{10} + \psi_{11}M_1) + (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$
1	0	$\geq -\frac{\psi_{10}}{\psi_{11}}$	$< -\frac{\psi_{20}}{\psi_{21}}$	$< -\frac{\psi_{30}}{\psi_{31}}$	$400 + 2M_1 + (\psi_{10} + \psi_{11}M_1) - (\psi_{20} + \psi_{21}M_2) - (\psi_{30} + \psi_{31}M_3)$

Bibliography

[1] Almirall, D., Ten Have, T., and Murphy, S.A. (2009). *Structural nested mean models for assessing time-varying effect moderation.* Biometrics, to appear.

[2] Baglin, T.P., Keeling D.M., Watson H.G., (2006). *Guidelines on oral anticoagulation (warfarin): third edition-2005 update.* Br. J. Haematol. 132 (3): 277-285.

[3] Bartels, R. H. and Golub, G. H., (1969) *The Simplex Method of Linear Programming.* Comm. ACM, **12**.

[4] Bellman, R. (1957). *Dynamic Programming.* Princeton University Press.

[5] Bellman, R. and Dreyfus, E. (1962). *Applied Dynamic Programming.* Princeton University Press,

[6] Berry, D., and B. Fristedt (1985). *Bandit Problems.* Chapman and Hall, London.

[7] Borgan Ø., Fiaccone R.L., Henderson R. and Barreto M.L. (2006). *Dynamic analysis of recurrent event data with missing observations, with application to infant diarrhoea in Brazil.* Scandinavian Journal of Statistics, **34**, 53-69.

[8] Canner, P. L. (1970). Selecting one of two treatments when the responses are dichotomous. Journal of the American Statistical Association, **65**, 293-306.

[9] Clayton, M.K. and Witmer, J.A. (1988). *Two-stage Bandits.* Ann. Statist, **16**, 887-894.

[10] Cormen, Thomas H., Leiserson, Charles E., Rivest, Ronald L.(1990). *Introduction to Algorithms.* The MIT Electrical Engineering and Computer Science Series. MIT Press, Cambridge, MA; McGraw-Hill Book Co., New York, **8**, 260-262.

[11] Davidian, M. (2006). *Applied Longitudinal Data Analysis.* New York: Springer-Verlag, **55**, 257-263.

[12] Dawid, A. P. (1979). *Conditional independence in statistical theory.* Journal of the Royal Statistical Society Series B , **41**, 1-31.

[13] Dawid, A.P. and Didelez, V. (2008). *Identifying optimal sequential decisions.* In Proceedings of the 24th Annual Conference on Uncertainty in Artifical intelligence, AUAI Press, 113-120.

[14] Diggle, P.J., Liang, K.Y., and Zeger, S.L. (1994). *Analysis of Longitudinal Data.* New York: Oxford University Press.

[15] Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis.* John Wiley and Sons.

[16] Laber E.,Qian M., Murphy S.A. (2010). *Statistical Inference in Dynamic Treatment Regimes.* Technical Report, Department of Statistics, University of Michigan, Number **506**.

[17] Eric V. D. (1982). *Dynamic Programming: Models and Applications.* Prentice-Hall, New Jersey.

[18] Ernst, D., Pierre G. and Louis, W. (2005). *Iteratively extending time horizon reinforcement learning* Proceedings of ECML, 96-107.

[19] Gittins, J.C. (1989). *Multi-Armed Bandit Indices.* Chichester: Wiley.

[20] Gittins J.C. and Jones, D.M. (1974). *A dynamic allocation index for the sequential design of experiments.* Progress in Statistics, European Meeting of Statisticians, **1**, 241-266.

[21] Gittins, J.C. and Wang, Y. (1992). *The learning component of dynamic allocation indices.* Ann. Statist, **20**, 1625-1636.

[22] Henderson, R., Ansel, P. and Alshibani, D. (2009). *Regret-regression for optimal dynamic treatment regimes.* Biometrics (to appear).

[23] Hernán, M.A. (2005). *Invited commentary: Hypothetical interventions to define causal effects: afterthought or prerequisite?* American Journal of Epidemiology, **162**, 618-620.

[24] Hernán, M.A., Brumback B, Robins J.M. (2002). *Estimating the causal effect of zidovudine on CD_4 count with a marginal structural model for repeated measures.* Statistics in Medicine, **21**, 1689-1709.

[25] Hernán, M.A., Robins J.M. (2006). *Estimating causal effects from epidemiological data.* Journal of Epidemiology and Community Health, **60**, 578-586.

[26] Hogan J.W. and Lee J.Y. (2004). *Marginal structural quantile models for longitudinal observational studies with time-varying treatment.* Statistica Sinica, **14**, 927-944.

[27] Holbrook A.M., Pereira J. A., Labiris R. (2005). *Systematic overview of warfarin and its drug and food interactions.* Arch. Intern. Med, **165**(10), 1095-1096.

[28] Hongtu Zhu, Joseph G. I., Niansheng, T., Daniel B. R., Xuejun, H., Ravi, B., and Bradley S. P. (2007). *A Statistical Analysis of Brain Morphology Using Wild Bootstrapping.* Published in IEEE Trans Med Imaging, **26(7)**, 954966.

[29] Horton J.D., Bushwick BM (1999). *Warfarin therapy: evolving strategies in anticoagulation.* Am Fam Physician. **59** (3), 635-646.

[30] Feldman D. (1962). *Contributions to the two-armed bandit problem.* Ann. Math. Statist, **33**, 847-856.

[31] Johnson B.A. (2008). *Treatment-competing events in dynamic regimes.* Lifetime Data Analysis, **14**, 196-215.

[32] Johnston, G. and So, Y. (2003). *Let the Data Speak: New Regression Diagnostics Based on Cumulative Residuals.* SAS Institute Inc. Cary, North Carolina, USA

[33] Kohler, D.F. (1982). *The Relation Among the Likelihood Ratio-, Wald-, and Lagrange Multiplier Tests and Their Applicability to Small Samples.* Santa Monica, Carolina, USA, **93**,893-898

[34] Lagoudakis, M. G. and Parr, R. (2003). *Reinforcement Learning as Classification.* Leveraging Modern Classifiers. Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-03), Acapulco, Mexico, 1432-1434.

[35] Landefeld C.S., Beyth R. J. (2001). *Anticoagulant-related bleeding: clinical epidemiology, prediction, and prevention.* Am J Med, **95** 315-28.

[36] Lavori, P.W. and Dawson, R. (2001). *Dynamic treatment regimes: Practical design considerations.* Statistics in Medicine, **20**, 1487-1498.

[37] Lawlor D. A., Davey S. G. and Ebrahim S. (2004). *Commentary: the hormone replacement-coronary heart disease conundrum: is this the death of observational epidemiology?.* Int J Epidemiol, **33** (3), 464-7.

[38] Lin, D.Y., Wei, L.J., and Ying, Z. (2002). *Model-checking techniques based on cumulative residuals.* Biometrics, **58**, 1-12.

[39] Liu, R.Y. (1988) *Bootstrap procedure under some non-I.I.D. models.* Annals of Statistics, **16**, 1696-1708.

[40] Lok J., Gill R., van de Vart A., Robins J.M. (2004). *Estimating the causal effect of a time-varying treatment on time-to-event using structural nested failure time models.* Statistica Neerlandica, **58**, 1-25.

[41] Mammen, E. (1993). *Bootstrap and wild bootstrap for high dimensional linear models.* Annals of Statistics **21**, 255-285.

[42] Moodie E., Richardson T.S. and Stephens, D. (2007). *Demystifying optimal dynamic treatment regimes.* Biometrics, **63**, 447-455.

[43] Murphy S. (2003). *Optimal dynamic treatment regimes* (with discussion). Journal of the Royal Statistical Society Series B, **65**, 331-366.

[44] Murphy, S.A. (2005). *A generalization error for Q-learning.* Journal of Machine Learning Research, **6**, 1073-1097.

[45] Nemhauser, G.L. (1967). *Introduction to Dynamic Programming.* New York, John Wiley and Sons.

[46] Norman, J.M. (1975). *Elementary Dynamic Programming.* New York, Crane, Russak & Company, Inc.

[47] Ormoneit, D. and Sen, S. (2002) *Kernel-Based Reinforcement Learning.* Kluwer Academic Publishers Hingham, USA, **49**, 161-178.

[48] Pearl J. (1995). *Causal diagrams for empirical research.* Biometrika, **82**, 669-710.

[49] Petersen M.L., Deek, S.G. Martin, J.N. and van der Laan M.J. (2007). *History-adjusted marginal structural models for estimating time-varying effect modification.* American Journal of Epidemiology, **166**, 985-993.

[50] Ritz, C. and Streibig, J. C. (2008). *Nonlinear Regression with R.* New. York: Springer

[51] Robins J.M. (1986). *A new approach to causal inference in mortality studies with sustained exposure preiods: application to control the healthy worker survivor effect.* Math. Mod, **7**, 1393-1512.

[52] Robins, J.M. (1994). *Correcting for non-compliance in randomized trials using structural nested mean models.* Communications in Statistics, **23**, 2379-2412.

[53] Robins J.M. (2004). *Optimal structured nested models for optimal sequential decisions.* Proceedings of the Second Seattle Symposium on Biostatistics, ed D.Y. Lin and P.J. Heagerty P.J. New York: Springer, 189-326.

[54] Robins J. M., Greenland S. (2000). *Comment on (Causal inference without counterfactuals) by Dawid A. P.* Journal of the American Statistical Association -Theory and Methods **95**, 477-482.

[55] Robins J.M., Hernán, M.A. and Brumback, B. (2000). *Marginal structural models and causal inference in epidemiology.* Epidemiology, **11**, 550-560.

[56] Robins J.M., Orellana, L., Rotnitzky,. (2008). *Estimation and extrpolation of optimal treatment and testing strategies.* Statist. Med. **27**, 4678-4721.

[57] Robins, J.M., Rotnitzky, A. and Scharfstein, D. O. (1999). *Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In Statistical Models in Epidemiology* (E. Halloran, ed.) Springer, New York.

[58] Robinson D.R. (1982). *Algorithms for evaluating the dynamic allocation index.* Operations Research Letters, **1**, 72-74.

[59] Rosenbaum, Paul R., (2002). *Observational studies. Second edition.* Springer Series in Statistics. Springer-Verlag, New York, 62-07.

[60] Rosenbaum P. R., Rubin D.B. (1983). *The central role of the propensity score in observational studies for causal effects.* Biometrika **70**, 41-55.

[61] Roderick J.L. and Rubin, D. B. (2000). *Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches.* Annual Review of Public Health, **21**, 121-145.

[62] Ross, S. (1970). *Applied Probability Models with Optimization Applications.* Holden-Day, San Francisco.

[63] Ross, S.M. (1983. *Introduction to Stochastic Dynamic Programming.* Academic Press, Inc. Orlando, USA.

[64] Rosthøj S., Fullwood C., Henderson R. and Stewart S. (2006). *Estimation of optimal dynamic anticoagulation regimes from observational data: a regret-based approach.* Statistics in Medicine, **25**, 4197-4215.

[65] Schwarz, U.I., Ritchie, M.D., Bradford, Y., Li, C., Dudek, S.D., Frye-Anderson, A., Kim, R.B., Roden, D.M., and Stein, M. (2008). *Genetic determinants of response to warfarin during initial anticoagulation.* New England Journal of Medicine, **358**, 999-1008.

[66] Scott, L. and Kung, Y. (1986). *Longitudinal data analysis for discrete and continuous outcomes.* Biometrics, **42** (1), 121.

[67] Spirtes P., Glymour C., Scheines R. (1993). *Causation, Prediction, and Search.* Lecture Notes in Statistics. New York: Springer-Verlag.

[68] Sundaresh, S.,Leong, T.Y. and Haddawy, P., (1999). *Supporting multi-level multi-perspective dynamic decision making in medicine.* Proceedings of the 1999 AMIA Annual Symposium, 161-165.

[69] Sutton, R. and Barto, A.G.(1998). *Reinforcement Learning: An Introduction.* MIT Press, Cambridge.

[70] Taha, H.A. (1992). *Operations Research: An Introduction.* USA, Winston (Collier-Macmillan).

[71] Thompson, W. R., (1933). *On the likelihood that one unknown probability exceeds another in view of the evidence of two samples.* Biometrika, **25** (3-4), 285-294.

[72] Upton, G., and Cook, I. (2002). *A Dictionary of Statistics*, Oxford, U.K.: Oxford University Press.

[73] Watkins, C.J.(1989). *Learning from Delayed Rewards.* King's College, Cambridge, UK.

[74] Whittle P. (1980). *Multi-armed bandits and the Gittins index.* Journal of the Royal Statistical Society Series B, **42**, 143-149.

[75] Witmer, J.A. (1986). *Bayesian multistage decision problems.* Ann. Statist, **14** (1), 283-297.

[76] Wu, C.F.G. (1986). *Jackknife bootstrap and other resampling methods in regression analysis.* Annals of Statistics, **14**, 1261-1295.

[77] Yong Z., Xiaobo Z., Alexei D., Marta L., Donald A., Junying Y. and Stephen T.W. (2007). *Automated neurite extraction using dynamic programming for high-throughput screening of neuron-based assays.* National Institutes of Health, **35** (4), 1502-15.