GAINS, CLAIMS AND PAINS

Mathematical and Statistical Problems in Occupational Health and Safety

Samuel N. Cohen

Supervisor: Elder Professor Charles E. M. Pearce

November 1, 2007

Thesis submitted for the degree of Honours in Statistics

SCHOOL OF MATHEMATICAL SCIENCES
DISCIPLINE OF STATISTICS

THE UNIVERSITY
OF ADELAIDE
AUSTRALIA

Declaration

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, except where due acknowledgement is made in the thesis. Furthermore, I believe that it contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution.

Signed: .....

Date: .....

Acknowledgements

As with any piece of work, there are a large number of people who need to be thanked:

- The members of the Mathematics Department at Adelaide University, in particular Finnur Larusson and Gary Glonek, for their patience with me and my persistent rambling ideas, always presented to them in a decidedly half-baked fashion. To Eric Parsonage, for keeping me on my toes, and to Jono Tuke, for always being there to help.
- My good friend Adrian Ahrens, for his extraordinary work in managing to download and format large segments of the OSHA database when it looked as though I might have no data.
- Cathy Parsonage, for giving me her time and insights into OH&S regulation in South Australia.
- All those who proof-read this work (of course, any remaining errors are completely my own fault).
- The late Shengjun Guo, for introducing me to this topic and giving me guidance on what to work on.
- Charles Pearce, for his brilliant supervision of me, and somehow managing to guide me to do something, even when it seemed that all we ever did was digress.
- My new baby boy, John, for being such a wonderful sleeper and allowing me to finish this without too much stress, for keeping me entertained for hours on end simply by looking around the room, and for constantly reminding me of how unimportant all this is.
- To Juli, my love, for her extraordinary patience, and making me coffee when I haven't been sleeping enough.
- And finally, to Christ Jesus, by whose grace comes all that I am and all that I have.

Soli Deo Gloria

S.Cohen, November 1, 2007

Errata

- p. 14, The final equation should read

$$\gamma(i) := E[C(i)|\kappa^{n+1}] + \beta \sum_j p_{ij}^{\kappa^{n+1}} v_j - E[C(i)|\kappa^n] - \beta \sum_j p_{ij}^{\kappa^n} v_j.$$

- p. 15 The second equation should read

$$\gamma = (I - \beta P_{\kappa^{n+1}})(E[C^*|\kappa^{n+1}] - E[C^*|\kappa^n])$$

- p. 44, The argument for the expectation and variance of $\hat{\boldsymbol{\phi}}$ is flawed. A better argument is as follows:

$$S(\boldsymbol{\phi}) \approx S(\hat{\boldsymbol{\phi}}) + S'(\hat{\boldsymbol{\phi}}) \cdot (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})$$

  from a Taylor expansion around $\hat{\boldsymbol{\phi}}$. As $S(\hat{\boldsymbol{\phi}}) \equiv 0$, evaluating this at $\boldsymbol{\phi}_0$ implies

$$\hat{\boldsymbol{\phi}} \approx \boldsymbol{\phi}_0 - \mathcal{I}(\hat{\boldsymbol{\phi}})^{-1} S(\boldsymbol{\phi}_0).$$

  Hence as $E[S(\boldsymbol{\phi}_0)] = \mathbf{0}$, we know $\hat{\boldsymbol{\phi}}$ has expectation $\boldsymbol{\phi}_0$ and variance $\approx \mathcal{I}(\hat{\boldsymbol{\phi}})^{-1} V \mathcal{I}(\hat{\boldsymbol{\phi}})^{-1}$, where $V := V[S(\boldsymbol{\phi}_0)]$. We can then show that $\mathcal{I}(\hat{\boldsymbol{\phi}}) \to V$ as $n \to \infty$, in some sense, and the stated result follows.

- p. 53, The equation half way down the page, the $\boldsymbol{\beta}$ should be omitted. In other words, it should read

$$\log f(\xi, \eta) = \phi(\beta_1 + \beta_2 \xi) + \phi(\beta_3 + \beta_4 \eta).$$

"ΟΙ ΚΥΡΙΟΙ ΤΟ ΔΙΚΑΙΟΝ ΚΑΙ ΤΗΝ ΙΣΟΤΗΤΑ ΤΟΙΣ ΔΟΥΛΟΙΣ
ΠΑΡΕΧΕΣΘΕ ΕΙΔΟΤΕΣ ΟΤΙ ΚΑΙ ΥΜΕΙΣ ΕΧΕΤΕ ΚΥΡΙΟΝ ΕΝ ΟΥΡΑΝΩ"

ΠΡΟΣ ΚΟΛΟΣΣΑΕΙΣ 4:1

" Masters, give unto your servants that which is just and equal;
knowing that ye also have a Master in heaven."

Colossians 4:1 (AV)

Contents

1 Introduction **1**

1.1 Motivation 2

1.2 Previous Work 2

2 The Economics of Injury **5**

2.1 Firms, Workers and Regulators 6

2.2 Crime & Punishment 7

2.2.1 Regulation Issues 8

3 The Regulatory Game **11**

3.1 Dynamic Programming and Markov Decision Processes 11

3.1.1 The Policy Improvement Algorithm 13

3.2 Harrington's Markov Model 16

3.2.1 Finding Optimal Policies 18

3.2.2 Final Analysis 20

3.3 An Insurance-Based System 23

3.3.1 A Mathematical Model 24

3.3.2 Finding Optimal Policies 26

3.3.3 Final Analysis 30
3.4 Capturing Diminishing Returns 34
3.5 Further extensions 37

4 Estimation with Missing Information 41

4.1 Maximum Likelihood Estimation 42
4.1.1 The EM Algorithm 43
4.1.2 Estimation of Errors 44
4.2 Detection Controlled Estimation 45
4.3 Binary Choice Models 46
4.3.1 Single Index Models 48
4.4 Poisson Process Models 48
4.5 Overdispersion 50

5 Parameter Identifiability 51

5.1 A Formal Definition 52
5.2 A Small Example 53
5.3 Ein Gedankenexperiment 55
5.4 A General Response 55
5.4.1 Extending to Non-Binary Models 57
5.4.2 A Geometric Observation 57
5.5 Examples of failure 58
5.5.1 Uniform Distributions 59
5.5.2 Exponential Functions 60

5.5.3 Insufficient Points 61
5.5.4 Insufficient Differences in data 61
5.5.5 A Peculiar Counterexample 62
5.6 And An Example of Success 62
5.7 Some Heuristics 64
5.8 Semi-parametric estimation 65
5.9 Data Requirements 66

6 Data and Analysis **67**
6.1 A Model of Violations 69
6.2 Modelling Inspector Suspicion 73

7 Conclusions **77**
7.1 Economic Models 77
7.2 Statistical Methods 78
7.3 Practical Questions 79

Bibliography **81**

A Further Derivations **85**
A.1 Feinstein's Theorems 85
A.2 Identifiability with Logistic regression 86
A.3 Left-invertibility and Covariates 87

B Further Results **89**
B.1 Grouping Industries 89

B.2 Using Transformed Durations 90

B.3 Incorporating Interactions 90

C Computer Code **93**

C.1 Code to implement DCE methods 93

C.2 Code used to generate graphs 95

C.2.1 Code used to generate Figure 5.2 95

C.2.2 Code used to generate Figure 5.3 96

C.2.3 Code used to generate Figure 3.2 97

C.2.4 Code used to generate Figure 3.3 98

C.2.5 Code used to generate Figure 3.4 98

C.2.6 Code used to generate Figure 3.6 98

C.2.7 Code used to generate Figure 3.7 98

C.2.8 Code used to generate Figure 3.8 99

C.2.9 Code used to generate Figure 3.9 99

C.2.10 Code used to generate Figures 3.10 and 3.11 99

Chapter 1

Introduction

" When I was One, I had just begun..."

A.A. Milne (1927)
The End* from *Now We are Six

In this thesis, we shall explore various mathematical and statistical problems associated with the regulation of Occupational Health and Safety (OH&S). This is a significant area of economic thought, and is of importance in modern economic systems.

We shall in particular explore two key questions, namely:

- In what ways can a regulator optimally enforce the law with minimal effort? and
- When looking at records of injuries, how can we distinguish between differences in detection and compliance?

This thesis can be loosely separated into the responses to each of these questions. In Chapter 2 we shall outline some of the economic arguments surrounding this area of research, and the theory behind the approach that shall be taken. Following on from this, in Chapter 3, we shall discuss a model for targeted enforcement, and develop various extensions, which better mimic observed systems.

After this, we shall move on to the second question, first pausing (in Chapter 4) to discuss the general statistical theory and methodology needed to model and estimate the detection of violations of the law. This leads directly into Chapter 5, where we will discuss conditions under which estimation is possible, and improve on the conditions available in the literature. In Chapter 6 we will then apply this work to a data set obtained from the United States Department of Labor, and see what policy implications could be drawn from it.

We shall conclude in Chapter 7 with a brief summary of the results obtained and by outlining possible avenues for future research.

1.1 Motivation

As a motivation for looking at these two questions, we raise the same question as discussed by Bartel and Thomas (1985). We observe that despite regulation, occupational injuries are still common. Is the problem with regulation that it is ineffective at forcing compliance (the "noncompliance hypothesis"), or is it that compliance with regulations is not enough to prevent injury (the "inefficacy hypothesis")? Answering this question is very important to policy makers, as it determines whether the appropriate response is to alter the regulations, the regulator, or both.

Understanding how regulation works from an economic perspective and developing statistical methods with which to analyse the data available for such systems is crucial in answering this question, and it is to this end that we consider the questions raised above.

1.2 Previous Work

There has been a considerable variety of work done on this issue, streching through the disciplines of Law, Economics, Statistics and Industrial Relations. For this reason, we here outline only work particularly pertinent to the issues we shall address.

First, the economic discussion of the regulation began with Becker's (1968) work on the economics of crime and punishment. From this general framework (see Section 2.2 for details), the various aspects of regulation have been addressed from a game-theoretic perspective, for example in Laffont and Tirole (1986) and Laffont and Tirole (1991). The key model which we shall address in Chapter 3 was first developed by Harrington (1988) to particularly address regulation of this type, where a regulator seeks to prevent a firm from violating some law, and does so through the use of fines and inspections.

With regards to the particular situation of Occupational Health and Safety, Bartel and Thomas (1985) discuss a model for regulation, given that workers are also able to negotiate some of the conditions of their employment. Bartel and Thomas (1985) also raise the question of the significance of indirect effects of regulation (if it is asymmetrically enforced between firms of different types) and discuss to what extent workplace accidents are the fault of poor enforcement or poor regulations. These issues, and also those of the inefficiency caused by regulation, are very significant economically – for example Gray (1987) claims that approximately 30% of the slowdown in productivity

growth in maufacturing industries in the US during the 1970's can be attributed to increased regulation.

From a statistical perspective, the main approach used to estimate parameters in these types of models was discussed by Dempster, Laird and Rubin (1977). In particular, a variant on this was developed by Feinstein (1989) and then further discussed in Feinstein (1990) which applies more precisely to this situation.

Chapter 2

The Economics of Injury

> *" ... a class of labourers, who live only so long as they find work, and who find work only so long as their labour increases capital. These labourers, who must sell themselves piecemeal, are a commodity, like every other article of commerce, and are consequently exposed to all the vicissitudes of competition, to all the fluctuations of the market."*
>
> **Karl Marx and Friedrich Engels**
> ***The Communist Manifesto (1848)***

Before we can develop any models of how firms will behave with respect to regulation and occupational safety, we need to understand the basic principles of their behaviour. The fundamental paradigm that we shall be using is that of classical economics. Following Adam Smith's 'The Wealth of Nations' (1776), this sees production mainly as a function of two key inputs: capital and labour. Since the industrial revolution, these two factors have generally been separated, leading to the situation where the owner of a factory will hire workers to provide labour. In a modern economy, ownership is generally further separated from the day to day running of the factory, as it is the stockholders who act as the primary providers of capital, but do not necessarily take an active management role.

In general, it is agreed that it is better for society when work takes place in a safe environment. All parties involved in industry have a duty to ensure this safety, however as it is primarily the worker who suffers when accidents occur, and workers may have little ability to dictate their environment, the focus of the law is on the duties of an employer. In the U.K. and its colonies, since the institution of the "Health and Morals of Apprentices Act" (1802), and various developments throughout the early to mid. 19th Century, these duties have been explicitly stated. In particular, the "Factory Act" (1833) provided for Salaried Inspectors to enforce regulations, thereby beginning the situation discussed here. (For further historical discussion, see Innes (2002).)

The exact details of the law vary between nations and states, however in Australia in 2007 they can be summarised under the following three aims (Sappey, Burgess, Lyons and Buultjens, 2006):

1. The prevention of the occurrence of workplace injury, disease and death.
2. Providing compensation to workers who have suffered work-related illness and/or injury, or to their relatives in the case of a fatal accident.
3. The rehabilitation of workers suffering from work-related injury or illness in order to assist their return to work.

From an economic standpoint, this summary is incomplete. The prevention of injury is a costly exercise, and so it is usually not optimal to attempt to prevent all injuries. Instead, it is preferable to force all decision makers to internalise the costs of injury to the community, thereby forcing them to take a position which is better for society. For this reason regulations typically go beyond what a firm will do naturally, and so society, acting through the government, creates regulatory bodies to enforce the law. These regulators, through the use of fines, levies and other punishments or rewards, attempt to alter the relative cost of (non-)compliance to the firm. It is the nature and practice of this enforcement, and of the penalties and violations associated with it, that we intend to investigate.

2.1 Firms, Workers and Regulators

Loosely following Guo (1999) we make the following comments on the interaction between workers, management and a regulator[1]. We shall first consider how management and workers interact without the regulator.

In the absence of a regulator, it would first appear that management has few incentives to provide a safe workplace. This is somewhat naïve, as workers will demand a premium for working in an unsafe environment, and 'downtime', etc. can be costly. In most economies, there is also the possibility of civil action being brought against a firm in the case of negligence. Therefore, the firm will adopt a certain (typically low) level of safety.

At the same time, workers may undertake actions to decrease the risk of injury to themselves[2]. When a firm's management sees this, it may attempt to lower their own efforts, as it can pass the costs of prevention on to the worker. Workers resist this, and

[1]For the purposes of this analysis, we disregard any possible agency costs between the stockholders and management of a firm, or between government and a regulator.

[2]It is also possible that workers take actions which increase the risk to themselves, if doing so requires less effort or allows them to negotiate higher wages.

therefore we have a repeated/continuous time multi-person game. The solution to this game gives an equilibrium level of wages, worker's effort and workplace safety. Other factors will affect what solution is chosen – for example the availability of appropriately skilled labour[3].

If a regulator now enters the arena, the presence of punishments and rewards will result in a change in relative cost of safety – a firm's management now has to negotiate the risk of being fined, as well as the response of their workers. Simultaneously, workers also gain the power to report violations to the regulator, giving them more bargaining power when negotiating their position. It is therefore clear that the presence of an active regulator will work to decrease the relative marginal costs of safety to the firm, and so the firm will act to lessen the risk of injury.

Regulators may also fulfill other roles, by providing expertise in the area of workplace safety and assisting employers with preventative measures and worker rehabilitation.

2.2 Crime & Punishment

In a seminal paper, Becker (1968) expanded the ideas of economic choice theory to apply to criminal activity. The basic assumptions underlying this approach are:

- Obedience to law is not taken for granted.
- Conviction is not considered sufficient punishment in itself – for example the classic criteria of vengeance, deterrence, compensation and rehabilitation may be considered important.
- Crime incurs a social cost which is to be minimised, however actions taken to minimise it will also incur a social cost.
- The decision of an individual whether to commit an offence depends on their opinion of the costs and benefits of doing so. In particular, this includes their opinion of the probability that they will be caught, and the likely penalties that will follow from this. This decision is *rational*.

At a simple level, a Government then has a few tools with which to attempt to control criminal activity, for example:

- It can control the laws and regulations in place, determining what activities are worthy of punishment.

[3]This fact is made abundantly clear in Horwitz's (1994) Pulitzer prize winning article about the U.S. chicken industry, and in many of the discussions about 'third world sweatshops'.

- It can control the level of funding given to policing and regulation.
- It can set the types of punishment (a fine, imprisonment, etc...) appropriate for different types of offence.
- It can set the scale of punishment (fine quantity, length of prison term, etc...) appropriate to each type of offence.

Each of these tools is double-edged – high levels of policing/regulation and certain types of punishment (e.g. imprisonment) are costly to maintain; some effective punishments (e.g. torture) may be considered inappropriate in themselves; overly restrictive laws and totalitarian regimes incur further social and economic costs. In many ways, these measures can best be seen as the lesser of two evils, and therefore governments will not act indiscriminately to eliminate all crime, as the costs of doing so are too high.

In the context of OH&S, Mendeloff and Gray (2005), among others, have addressed the question of whether Becker's (1968) assumptions of utility maximisation are appropriate. Mendeloff & Gray propose that firms react to the shock induced by fines by paying more attention to all safety issues, leading to a higher responsiveness to enforcement. Weil (1996) examines empirical evidence that enforcement is effective, despite low levels of inspection (a phenomenon also noticed and modelled by Harrington (1988), to which we shall return in Chapter 3). Becker's (1968) conclusions have also been questioned more generally under different modelling assumptions, for example by Boyer, Lewis and Liu (2000), who model the regulations themselves as flexible.

2.2.1 Regulation Issues

In particular in this thesis we are looking at the regulation of Occupational Health and Safety within industry. In this context, there are issues associated with possible 'societal agency costs', arising from the situation with an independent regulatory and legislative system (i.e. costs to society caused by the fact that government and regulators may act in their own interests).

This can generally be summarised under the title of 'regulatory capture'. This notion was first proposed in some sense by Marx (1867), and was then developed by Stigler (1971), in work for which he received the 1982 Nobel Prize in Economics. Stigler's key claim can be summarised as "regulation is acquired by the industry and is designed and operated primarily for its benefit." (p. 3) In other words, industry may interfere with the regulator, to ensure that the regulator acts in the interests of industry. Therefore, public protection and socially optimal outcomes may not be pursued by the regulator as a primary goal.

Laffont and Tirole (1993) outline five methods through which various interest groups

can influence a regulatory official: (1) monetary bribes, (2) making the offer of future employment in regulated firms[4], (3) personal relationships between regulators and members of the interest group, (4) political collusion between the industry and the regulator, where industry refrains from publicly criticising the regulator, and (5) through political action including monetary contributions to political campaigns, with the aim of persuading sympathetic officials with influence over the agency (see Gordon and Hafer (2005) for a particular discussion of this).

While this issue is worth bearing in mind, and may help with the interpretation of some of our conclusions, the extent to which regulatory capture is present in the OH&S setting is not a question which we further address here. A good summary of the area can be found in Laffont and Tirole (1991) or Laffont and Tirole (1993, Ch. 11).

[4]This is a part of the so called 'revolving door hypothesis' (see for example, Gormley (1979)). Here a regulator (particularly an individual) currently or previously employed by a regulated industry acts (consciously or unconsciously) in a manner more or less favourable to the industry or company in question. In the context of OH&S, this is probably not a major issue, as the possible scope for regulation is large (there are many companies that can be regulated – unlike in, for example, the nuclear power industry in the US); however if regulation were highly specialised this could reappear as a significant concern.

Chapter 3

The Regulatory Game

> *" 'You were present on the occasion of the destruction of these trinkets, and indeed are the more guilty of the two, in the eye of the law; for the law supposes that your wife acts under your direction.'*
>
> *'If the law supposes that,' said Mr Bumble, squeezing his hat emphatically in both hands, 'the law is a ass – a idiot...' "*
>
> **Charles Dickens**
> ***Oliver Twist (1837)***

In this chapter, we shall outline models through which a regulator can enforce the law with less effort. While these models have clear flaws, they can help us to understand targeted enforcement and its consequences for compliance. To begin with, we need to establish the following basic theory.

3.1 Dynamic Programming and Markov Decision Processes

To solve these types of problems, we need to understand 'Markov Decision Processes'. These were developed in Howard (1960) and Bellman (1957).

We wish to model a system where there are a number of states that a firm can find itself in. In each state it has a decision to make, which results in an immediate cost and in altering the probabilities of moving to other states for the following period. We shall assume that the transition probabilities and costs depend only on the firm's decision at this time (they are *memoryless*), and so we describe this as a 'Markov Decision Process'.

Formally, a Markov Decision Process (MDP) can be described by:

- A set of states S.
- An 'action space' K_i of possible actions $\kappa(i)$ that can be taken when in state i.
- Each action $\kappa(i)$ leads to transition probabilities $p_{ij}^{\kappa(i)}$.
- Each action $\kappa(i)$ leads to an expected payoff/cost $E[C(i)|\kappa(i)]$ in each state.

We assume that when in state i, firms wish to minimise the expected long run discounted cost $E[C^*(i)]$. Let $C^{(k)}$ refer to the expected cost in k periods' time and $p_{ij}^{\kappa(i)}$ to the transition probabilities from state i to state j under action $\kappa(i)$. We can now write (omitting $\kappa(i)$ throughout for clarity)

$$\begin{aligned}
E[C^*(i)] &= E[C(i)] + \sum_{k=1}^{\infty} \beta^k E[C^{(k)}] \\
&= E[C(i)] + \beta E\left[C^{(1)} + \sum_{k=2}^{\infty} \beta^{k-1} C^{(k)}\right] \\
&= E[C(i)] + \beta \sum_j p_{ij}\left[E[C(j)] + \sum_{k=2}^{\infty} \beta^{k-1} E[C^{(k)}|j = \text{state after one period}]\right] \\
&= E[C(i)] + \beta \sum_j p_{ij} E[C^*(j) \text{ after one period}],
\end{aligned}$$

where $\beta \in [0, 1)$ is a fixed discount rate. In other words, we can consider the expected long run payoff from time zero to be the immediate payoff at time zero plus the (one-period discounted) long run payoff from time one.

Hence, as a matrix-vector equation, (with $P_\kappa = [p_{ij}^\kappa(i)]$ being the transition matrix under actions $\kappa = (\kappa(1), \kappa(2), ...)$, and obvious notation elsewhere)

$$E[C^*|\kappa] = E[C|\kappa] + \beta P_\kappa \cdot E[C^* \text{ after one period}|\kappa]. \tag{3.1}$$

We wish to find an optimal policy, that is, a decision rule to be followed in each state which will minimise this expected cost. We shall follow Bellman (1957) in

Definition 3.1.1 (The Principle of Optimality)**.** An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

Next, we shall cite (without proof) a theorem from Puterman (1994, p. 154) which states

Theorem 3.1.1. *Consider a Markov Decision Process with discrete state space. If, for all i, the action space K_i is compact, and $E[C(i)]$ and p_{ij} are continuous in $\kappa(i)$ (for all j), then there exists an optimal deterministic stationary policy.*

We have also, from p. 153,

Proposition 3.1.1. *When the state-space is discrete, an optimal stationary policy must be optimal for every starting state.*

Therefore, without loss of generality, we shall consider only stationary (time-invariant) policies. Because of this, $E[C^*] = E[C^*$ after one period], and so we can immediately rewrite (3.1) as

$$E[C^*|\kappa] = (I - \beta P_\kappa)^{-1} E[C|\kappa]. \tag{3.2}$$

We can characterise mathematically an optimal policy as one which satisfies the *Bellman Equations*

$$E[C^*(i)] = \inf_{\kappa(i)} \left\{ E[C(i)|\kappa(i)] + \beta \sum_j p_{ij}^{\kappa(i)} E[C^*(j)] \right\}$$

for all i, or in matrix-vector form,

$$E[C^*(i)] = \inf_{\kappa} \left\{ (I - \beta P_\kappa)^{-1} E[C|\kappa] \right\}.$$

As pointed out by Howard (1960), for any policy κ, P_κ is a stochastic matrix and so it has eigenvalues numerically less than one and therefore βP_κ has eigenvalues numerically less than one for any $\beta \in [0, 1)$. Hence $(I - \beta P_\kappa)^{-1} = \sum_{j=0}^{\infty} (\beta P_\kappa)^j$ is well defined. Moreover, as P is nonnegative, $(I - \beta P_\kappa)^{-1}$ is a matrix with nonnegative elements, and with elements at least as great as one on the main diagonal.

3.1.1 The Policy Improvement Algorithm

Given the above theory we now outline some useful methods and results for determining optimal policies. The first of these is *The Policy Improvement Algorithm*[1]. This is an iterative algorithm for determining the optimal policy for an arbitrary Markov Decision Process.

To implement this Algorithm, first assume some initial policy κ^0.

[1]This is taken directly from Howard (1960, p. 83ff.), with minor modifications.

At stage $n+1$, we have a policy κ^n from the previous iteration. We can determine the expected present values $\mathbf{v} := E[C^*|\kappa^n]$ under this policy using (3.1).

Then, for each i, a better policy for each state can be found by choosing $\kappa^{n+1}(i)$ such that we minimise (cf. (3.1))

$$E[C(i)|\kappa^{n+1}(i)] + \beta \sum_j p_{ij}^{\kappa^{n+1}(i)} v_j.$$

Provided that our action space K_i is a compact set (like [0,1]), this is an attainable objective. In general, if more than one minimum exists, then we shall choose one arbitrarily. Note that the immediate costs in state i depend only on $\kappa(i)$, that is, $E[C(i)]$ and p_{ij} do not depend on $\kappa(k)$ for any $i \neq k$, so conditioning only on $\kappa(i)$ does not cause any difficulty.

Choosing $\kappa^{n+1}(i)$ in this way for each i defines a new policy, which we can use to iterate this process. This process is known as the Policy Improvement Algorithm.

We now need to show that

Theorem 3.1.2. *For any sub-optimal policy, the Policy Improvement Algorithm decreases the expected costs in all states.*

Proof. We shall only show that the Policy Improvement Algorithm does not increase the expected cost in any state, as this is sufficient for our purposes. Proving that it converges to an optimal policy is difficult in full generality, as it requires the use of the contraction mapping theorem. The interested reader is referred to Puterman (1994).

We know that

$$\begin{aligned} E[C^*(i)|\kappa^{n+1}] &= E[C(i)|\kappa^{n+1}] + \beta \sum_j p_{ij}^{\kappa^{n+1}} E[C^*(j)|\kappa^{n+1}], \\ E[C^*(i)|\kappa^n] &= E[C(i)|\kappa^n] + \beta \sum_j p_{ij}^{\kappa^n} v_j \\ &= E[C(i)|\kappa^n] + \beta \sum_j p_{ij}^{\kappa^n} E[C^*(j)|\kappa^n]. \end{aligned}$$

Now let $\gamma(i)$ be the improvement the algorithm was able to achieve when in state i, keeping v fixed, that is

$$\gamma(i) := E[C(i)|\kappa^{n+1}] + \beta \sum_j p_{ij}^{\kappa^{n+1}} v_j - E[C(i)|\kappa^n] + \beta \sum_j p_{ij}^{\kappa^n} v_j.$$

By construction, it is clear that $\gamma(i) \leq 0$ for all i.

We can now write

$$E[C^*(i)|\kappa^{n+1}] - E[C^*(i)|\kappa^n] = \gamma(i) + \beta \sum_j p_{ij}^{\kappa^{n+1}} (E[C^*(j)|\kappa^{n+1}] - E[C^*(j)|\kappa^n])$$

and hence in matrix-vector form

$$\gamma = (I - \beta P_{\kappa^{n+1}})^{-1}(E[C^*|\kappa^{n+1}] - E[C^*|\kappa^n])$$

We noted above that $(I - \beta P_{\kappa^{n+1}})^{-1}$ has all nonnegative elements. Therefore $\gamma(i) \leq 0$ for all i implies

$$E[C^*(i)|\kappa^{n+1}] \leq E[C^*(i)|\kappa^n]$$

as desired. *Q.E.D*

This allows us to state the following results

Theorem 3.1.3. *For an optimal policy κ, when the assumptions of Theorem 3.1.1 are satisfied,*

$$E[C^*(i)|\kappa] = \inf_\lambda \left\{ E[C(i)|\lambda] + \beta \sum_j p_{ij}^\lambda E[C^*(i)|\lambda] \right\}$$

and

$$E[C^*(i)|\kappa] = \inf_\lambda \left\{ E[C(i)|\lambda] + \beta \sum_j p_{ij}^\lambda E[C^*(i)|\kappa] \right\}.$$

Proof. The first of these statements is simply the definition of an optimal policy, the second flows from it being a stationary point of the policy improvement algorithm. *Q.E.D*

From this it is clear that

Lemma 3.1.4. *Suppose the assumptions of Theorem 3.1.1 are satisifed. If $E[C(i)|\kappa(i)]$ and $p_{ij}^{\kappa(i)}$ are both linear in $\kappa(i)$ for all i and j, then an optimal policy must occur at a vertex of the action space.*

Proof. This follows immediately from the second statement of Theorem 3.1.3, recalling that the minimum of any linear function on a compact space must occur at a vertex. *Q.E.D*

3.2 Harrington's Markov Model

In a significant paper[2], Harrington (1988) developed a model of 'targeted' enforcement for violations of environmental regulations in discrete time. This work can equally be applied to the OH&S situation. The basic model is as follows:

At each 'play of the game,' the enforcement agency faces a choice: to inspect or not to inspect. Simultaneously, the firm faces a choice: to violate or not to violate. The aim of the regulator is to have the firm in compliance as frequently as possible, subject to some budgetary constraint. The aim of the firm is to minimise costs. We assume that the firm faces a significant cost to compliance c, and that the agency can impose a significant (but finite) penalty F, if violation is detected. It is clear that if $F > c$ it may not always be optimal for firms to be in non-compliance, and therefore we have a nontrivial process.

Harrington then makes the following development. Instead of treating all firms the same, the regulator can 'target' some firms based on their past performance[3]. In particular, firms are partitioned into one of two groups: G_0 and G_1. Each of these groups has an inspection probability ϕ_i and a fine F_i, with $\phi_1 \leq \phi_0$ and $F_1 \leq F_0$. Violations in G_1 are punished by exile to G_0 with probability p, while compliance in G_0 is rewarded by possible return to G_1 with probability g. (Note: this is a slight extension to Harrington's model, as he assumes that $p = 1$.) These dynamics are outlined in Figure 3.1.

The payoff to a firm is then dependent on its decision of when to violate. Let $\kappa(i)$ be the probability that a firm will violate when in state i. Then the one-period expected cost to the firm is

$$E[C(i)|\kappa(i)] = \kappa(i)\phi_i F_i + (1 - \kappa(i))c. \tag{3.3}$$

There are two alternative interpretations of this equation: (1) The firm is choosing a mixed strategy between violating and not violating, and the expected cost is therefore a weighted average of the expected costs under each strategy, or (2) the firm is choosing a strategy which gives a probability $\kappa(i)$ of violation, and the cost of such a strategy is linear in $\kappa(i)$. We do not address which interpretation is better, but progress with (2) in Section 3.4.

It is clear that given a policy κ, this entire process forms a Markov Chain, with state space $\{0, 1\}$ and transition matrix

$$P_\kappa = \begin{bmatrix} 1 - (1 - \kappa(0))\phi_0 g & (1 - \kappa(0))\phi_0 g \\ \kappa(1)\phi_1 p & 1 - \kappa(1)\phi_1 p \end{bmatrix}. \tag{3.4}$$

[2]Much of the analysis in the following section follows that of Harrington's original paper, however with slight modifications to allow increased flexibility in the model, and to allow the notation to extend more readily to some other cases.

[3]Friesen (2003) develops an optimal targeting system, based on steady-state dynamics rather than past performance, however the results are quite similar.

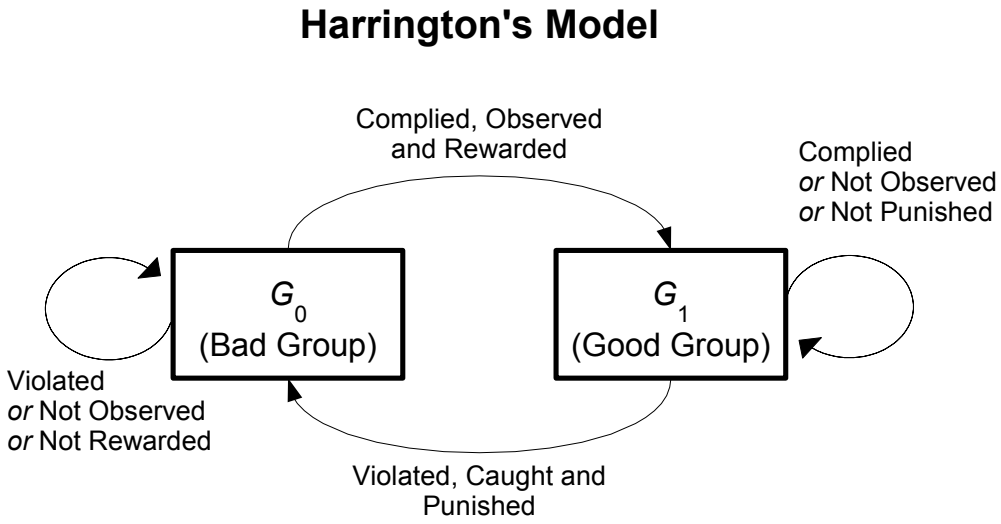

Figure 3.1: Harrington's Model of Regulatory Dynamics

We assume that the firm has a constant one-period discount factor $\beta \in [0, 1)$, and that it is an expected cost minimiser. Hence, let $C_\kappa^{(j)}(i)$ denote the cost at time j for a given policy κ (which may be omitted for notational simplicity), when a firm starts (at time 0) in group i. The firm will choose a policy to minimise

$$E[C^*(i)|\kappa] = \sum_{j=0}^{\infty} \beta^j E[C^{(j)}(i)|\kappa] = E[C(i)|\kappa(i)] + \sum_{j=1}^{\infty} \beta^j E[C^{(j)}(i)|\kappa]. \tag{3.5}$$

To summarise, a firm selects a policy function

$$\kappa : \{0, 1\} \to [0, 1]$$

and uses this to minimise $E[C^*|\kappa]$ as defined in (3.5). The regulator has strategy space

$$0 \leq \phi_1 \leq \phi_0 \leq 1,$$

$$0 \leq F_1 \leq F_0 \leq F^*$$

for some maximal fine F^*, as it is not possible for an unbounded fine to be imposed on a firm.

In a game-theoretic sense, these strategies are 'pre-mixed', so hopefully it should be possible to find an optimal strategy within these spaces[4]. This is implicitly claiming that firms are unable to perfectly collude, and therefore act as though the regulation parameters are fixed. Under the various theories of regulatory capture, we would clearly need to be more careful about this.

[4]It is not certain that we will find a truly optimal strategy this way, as the Von Neumann Minimax Theorem applies only to convex/finite zero-sum games, which this clearly is not.

To find a solution, we first determine the response of the firm to a given regulator-strategy $\{\phi_0, \phi_1, F_0, F_1\}$, and then incorporate this response in considering the values to be selected by the regulator.

3.2.1 Finding Optimal Policies

We now set about finding solutions to Harrington's Model. To begin, we know from Theorem 3.1.3 that

$$E[C^*(i)|\kappa] = \inf_{\lambda} \left\{ E[C(i)|\lambda] + \beta \sum_j p_{ij}^{\lambda} E[C^*(i)|\kappa] \right\}.$$

In this case, we know that $E[C(i)|\kappa]$ and $p_{ij}^{\kappa(i)}$ are linear in $\kappa(i)$. Therefore by Lemma 3.1.4, we can see that an optimal policy must occur for

$$\kappa : \{0,1\} \to \{0,1\},$$

that is, when a firm will simply choose whether to violate or not in each group, and will never choose a policy of violating with some probability $\kappa(i) \in (0,1)$.

Hence, there are four possible strategies that a firm could have:

$$(\kappa(0), \kappa(1)) = \begin{cases} (0,0) \\ (0,1) \\ (1,0) \\ (1,1) \end{cases}$$

For each, it is easy to find the value of $E[C^*|\kappa]$.

For $(\kappa(0), \kappa(1)) = (0,0)$, i.e. the policy of "Never violate," we have

$$P_{(0,0)} = \begin{bmatrix} 1-\phi_0 g & \phi_0 g \\ 0 & 1 \end{bmatrix}$$

and therefore

$$\begin{aligned} E[C^*|\kappa] &= (I - \beta P)^{-1} E[C|\kappa] \\ &= \begin{bmatrix} \frac{c}{1-\beta} \\ \frac{c}{1-\beta} \end{bmatrix}. \end{aligned}$$

For $(\kappa(0), \kappa(1)) = (0,1)$, i.e. the policy of "Violate when in the Good Group only," we have

$$P_{(0,1)} = \begin{bmatrix} 1-\phi_0 g & \phi_0 g \\ \phi_1 p & 1-\phi_1 p \end{bmatrix}$$

and therefore

$$E[C^*|\kappa] = \begin{bmatrix} \frac{A_0}{A_0+B_0} \cdot \frac{c}{1-\beta} + \frac{B_0}{A_0+B_0} \cdot \frac{\phi_1 F_1}{1-\beta} \\ \frac{A_1}{A_1+B_1} \cdot \frac{c}{1-\beta} + \frac{B_1}{A_1+B_1} \cdot \frac{\phi_1 F_1}{1-\beta} \end{bmatrix},$$

where $A_0 = 1 - \beta(1 - \phi_1 p), B_0 = \beta\phi_0 g, A_1 = \beta\phi_1 p, B_1 = 1 - \beta(1 - \phi_0 g)$.

For $(\kappa(0), \kappa(1)) = (1, 0)$, i.e. the policy of "Violate when in the Bad Group only," we have

$$P_{(1,0)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

and therefore

$$E[C^*|\kappa] = \begin{bmatrix} \frac{\phi_0 F_0}{1-\beta} \\ \frac{c}{1-\beta} \end{bmatrix}.$$

Finally, for $(\kappa(0), \kappa(1)) = (1, 1)$, i.e. the policy of "Always violate," we have

$$P_{(1,1)} = \begin{bmatrix} 1 & 0 \\ \phi_1 p & 1 - \phi_1 p \end{bmatrix}$$

and therefore

$$E[C^*|\kappa] = \begin{bmatrix} \frac{\phi_0 F_0}{1-\beta} \\ \frac{A}{A+B} \cdot \frac{\phi_0 F_0}{1-\beta} + \frac{B}{A+B} \frac{\phi_1 F_1}{1-\beta} \end{bmatrix},$$

where $A = \beta\phi_1 p, B = 1 - \beta$.

To summarise:

Policy κ	$E[C^*(0)\|\kappa]$	$E[C^*(1)\|\kappa]$
$(0,0)$	$\frac{c}{1-\beta}$	$\frac{c}{1-\beta}$
$(0,1)$	$\frac{A_0}{A_0+B_0} \cdot \frac{c}{1-\beta} + \frac{B_0}{A_0+B_0} \cdot \frac{\phi_1 F_1}{1-\beta}$	$\frac{A_1}{A_1+B_1} \cdot \frac{c}{1-\beta} + \frac{B_1}{A_1+B_1} \cdot \frac{\phi_1 F_1}{1-\beta}$
$(1,0)$	$\frac{\phi_0 F_0}{1-\beta}$	$\frac{c}{1-\beta}$
$(1,1)$	$\frac{\phi_0 F_0}{1-\beta}$	$\frac{A_2}{A_2+B_2} \cdot \frac{\phi_0 F_0}{1-\beta} + \frac{B_2}{A_2+B_2} \frac{\phi_1 F_1}{1-\beta}$

Here $A_0 = 1 - \beta(1 - \phi_1 p), B_0 = \beta\phi_0 g$, $A_1 = \beta\phi_1 p, B_1 = 1 - \beta(1 - \phi_0 g)$, $A_2 = \beta\phi_1 p$ and $B_2 = 1 - \beta$ are non-negative weights.

We can now see that $(0, 1)$ is a suboptimal policy, as

- If $c \leq \phi_0 F_0$, then we can see that $E[C^*]$ is no larger under policy $(0, 0)$.
- If $\phi_0 F_0 > c$, then as $\phi_1 F_1 < \phi_0 F_0$ we can see that $E[C^*]$ is no larger under policy $(1, 1)$ or policy $(0, 1)$.

Therefore this policy is (weakly) dominated, and so will never be chosen.

As a result of this analysis, we have that there are 3 strategies that a firm might pursue:

$$(\kappa(0), \kappa(1)) = \begin{cases} (0,0) \\ (0,1) \\ (1,1) \end{cases}$$

3.2.2 Final Analysis

Consider the variety of firms. In the setting of this model, they differ on two main counts: c and β. We investigate the effects of these on a firm's optimal policy choice and costs. Harrington's original work considers only the effects of c (except for the case $\beta = 0$); we shall see that β also determines significant differences.

To compare costs, note that a firm is indifferent between strategies $(0,0)$ and $(0,1)$ when $c = \phi_1 F_1 =: L_0$. Next, by comparing the costs for strategies $(0,1)$ and $(1,1)$, we find that a firm is indifferent between them when

$$c = \phi_0 F_0 + \frac{\beta\phi_0 g[\phi_0 F_0 - \phi_1 F_1]}{1 - \beta[1 - \phi_1 p]} =: L_1$$

and it is clear that higher costs will cause a firm to be more likely, on average, to violate. This is shown in Figure 3.2. See C.2.3 for details of this graph[5].

When $c \neq \phi_1 F_1$, a firm tends to indifference between strategies $(0,0)$ and $(0,1)$ if

$$\frac{\beta\phi_1 p}{1 - \beta[1 - \phi_0 g]} \to \infty.$$

For $\phi_1, g \neq 0$, this will not occur for any $\beta \in [0,1)$. A firm will be indifferent between strategies $(1,0)$ and $(1,1)$ if

$$\beta = \frac{\phi_0 F_0 - c}{[1 - \phi_1 p](\phi_0 F_0 - c) - \phi_0 g[\phi_0 F_0 - \phi_1 F_1]} =: M_1.$$

This is demonstrated in Figure 3.3. See C.2.4 for details.

It is informative to show the changes with respect to c and β simultaneously, as in Figure 3.4 (details in C.2.5). In this image, a darker colour indicates a policy which entails more frequent violation. Also, as $\beta \to 1$, our expected costs increase without bound, regardless of the policy chosen. We therefore also plot the cost multiplied by $(1 - \beta)$ to give an alternative visualisation without this effect.

[5]In this figure and others like it, $E[C^*(0)]$ is plotted for each policy, the optimal value is shown in bold and the policy thus selected is given below.

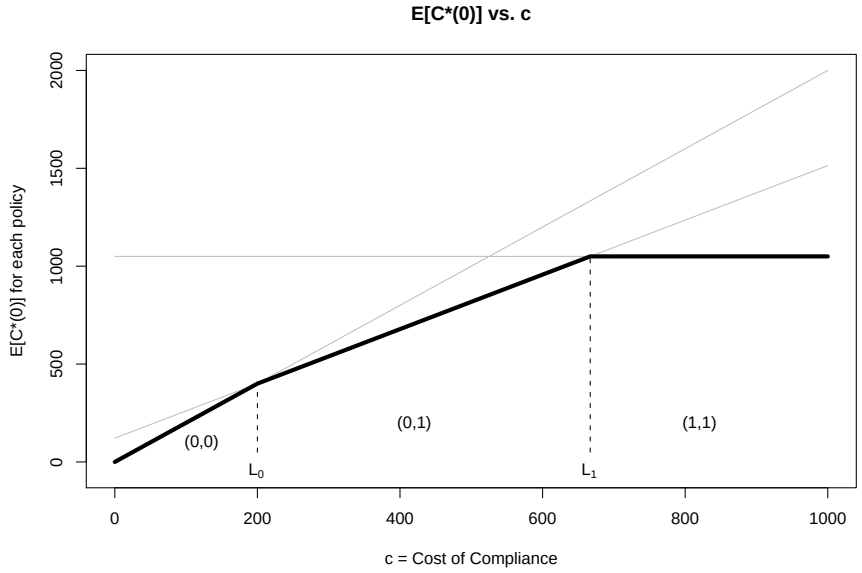

Figure 3.2: Indicative graph of $E[C^*(0)]$ vs c

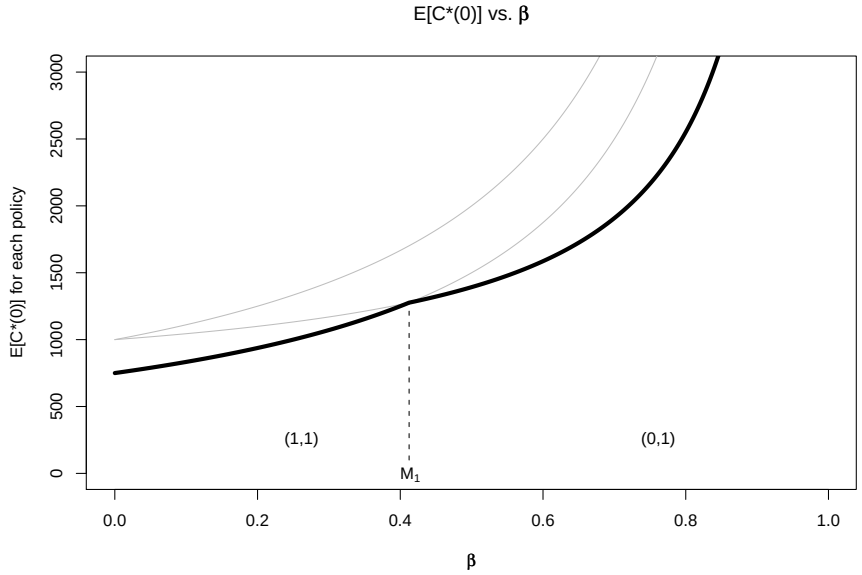

Figure 3.3: Indicative graph of $E[C^*(0)]$ vs β

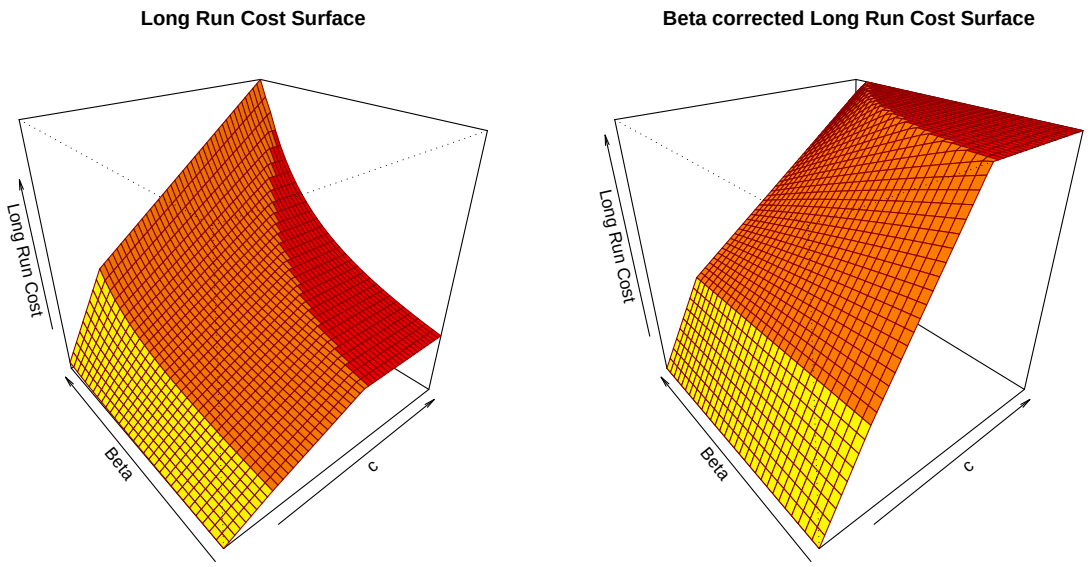

Figure 3.4: $E[C^*(0)]$ and $E[(1-\beta)C^*(0)]$ vs c and β

The regulator has the aim of minimising the overall violation rate. This can be done using two key steps: minimising the number of firms which adopt strategy $(1,1)$ and maximising the time that $(0,1)$ firms spend in G_0. Any $(0,0)$ firms are not a major concern, as they can be treated like $(0,1)$ firms and will continue to not violate.

To find the amount of time that a firm spends, on average, in G_0, we need to determine the steady-state probabilities for this Markov Chain. To do this, note that only $(0,1)$ firms are interesting in this regard – clearly $(0,0)$ firms have probabilities $\boldsymbol{\pi} = (0,1)$ and $(1,1)$ firms have probabilities $\boldsymbol{\pi} = (1,0)$. Therefore, considering only $(0,1)$ firms, the partial balance equation gives

$$\phi_1 p \pi_1 = \phi_0 g \pi_0,$$

so

$$\pi_1 = \frac{\phi_0 g}{\phi_1 p + \phi_0 g},$$
$$\pi_0 = \frac{\phi_1 p}{\phi_1 p + \phi_0 g},$$

or

$$\boldsymbol{\pi} \propto (\phi_1 p, \phi_0 g).$$

From this, we can see the following difficulty: to maximise the time that $(0,1)$ firms spend in G_0, then the regulator will set g low and p high. However, this will have the effect of driving these firms to become $(1,1)$ firms, as this will imply a low value of L_1

and a high value of M_1. Thus more firms will prefer to violate in both states. The optimal solution will therefore depend considerably on the distribution of c and β within the population of firms under regulation. Nevertheless, the following conclusions can be made:

- F_0 should be set as high as possible, as this has the effect of decreasing M_1 and increasing L_1.
- Typically, p should be set high and g low, to maximise the time $(0,1)$ firms will comply.
- As the regulator can force compliance by keeping $(0,1)$ firms in G_0, they will probably prefer to set F_1 low, as this will increase L_1 and decrease M_1, rendering more firms in this category. It will also drive most $(0,0)$ firms to become $(0,1)$ firms, but as noted earlier, these can then be forced to usually remain in G_0, which will result in compliance. This can be done using a very small number of inspections.

This analysis also confirms the following intuitive phenomena:

- Firms with higher costs are more likely to violate (as their costs are more likely to exceed L_0 and L_1).
- Firms in distress are more likely to violate (as they become more myopic – that is, their discount rate approaches zero – and therefore their discount rate is more likely to be below M_1).

3.3 An Insurance-Based System

When considering the OH&S situation in South Australia, we note that Harrington's model is not without flaws. In particular:

- Fines are very rare, as Harrington's model predicts, however there is no evidence of a 'large' fine for repeat offenders.
- The primary enforcement mechanism used is based on different insurance categories – firms which violate have to pay a higher levy.

A summary of the system in force in South Australia can be obtained on the Workcover SA website, in particular in the documents Workcover SA (2006a), Workcover SA (2007a) and Workcover SA (2007b). A few comments about this system are needed before we proceed to developing a mathematical model.

- The levies are proportional to the total payroll of the organisation, with the "Base Levy" rate being determined by the industry grouping which a firm is in.
- In general, the size of fines is considerably lower than the cost of the Levy.
- Firms are required by law to report all incidents, and various schemes are in place to encourage workers to report violations directly to the regulator (Safework SA).
- Schemes are in place to further penalise unusually bad workplaces (see Workcover SA (2006b)) and to reward unusually good workplaces (see Workcover SA (2007c)).

In light of the above, we will build a new model of this enforcement system.

3.3.1 A Mathematical Model

We shall make the following simplifying assumptions:

- All incidents are reported (or at least, the probability of an incident being reported is incorporated into the company's choice of violation probability, rather than being set by the regulator)[6].
- All fines are incorporated into the levy imposed on a company.
- All violations are treated equally.
- We ignore the mechanisms in place to reward/penalise unusually good/bad workplaces, and model only the basic levy system.

With the above, we construct the following basic reward/penalty system:

- There are three groups: G_{-1}, G_0, G_1. A firm violates when in group i with probability $\kappa(i)$.
- A firm in G_1 has a levy F_1, a firm in G_{-1} or G_0 has levy F_0. For notational simplicity, we may also write $F_{-1} \equiv F_0$.
- Violations in G_1 are punished with probability p by exile to G_{-1} in the following period[7].

[6] It is quite possible to build a model without this assumption, but the equations quickly become intractable. Further, such a model simply obscures the primary qualitative results that we wish to obtain.

[7] We ignore the one year of lag time present in the SA system, as this simply serves to complicate the model, without providing significant qualitative results.

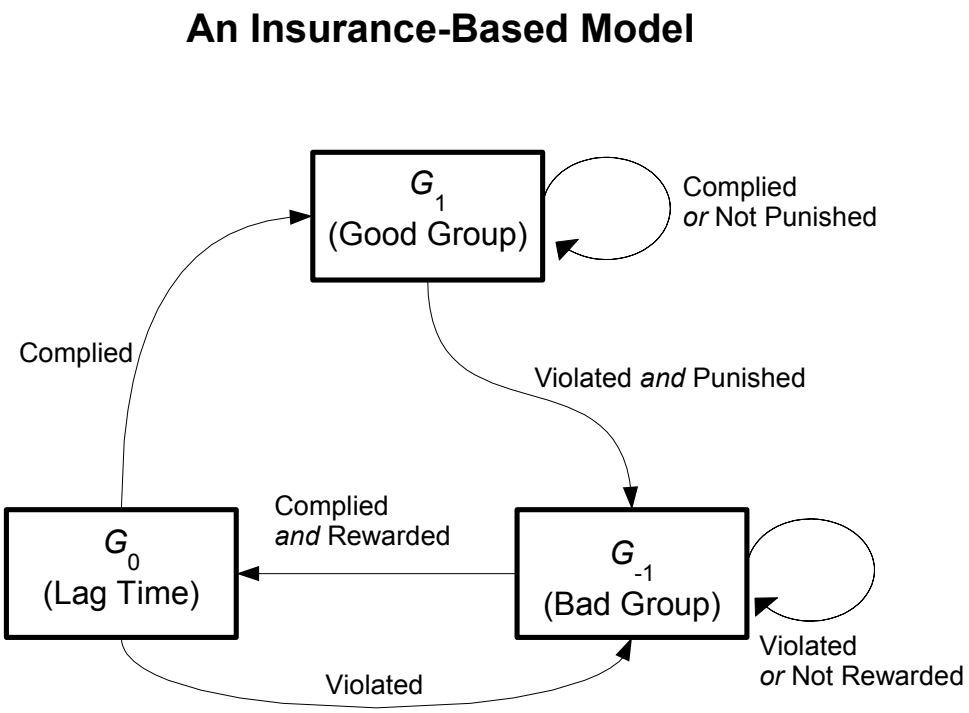

Figure 3.5: A Regulation Model based on Insurance

- Compliance in G_{-1} is rewarded by return to G_1 with probability g, however this requires passing through G_0 for one period, in which violation will be automatically punished by a return to G_{-1}.

Therefore, this gives a Markov chain with transition matrix[8]

$$P = \begin{bmatrix} 1-(1-\kappa(-1))g & (1-\kappa(-1))g & 0 \\ \kappa(0) & 0 & 1-\kappa(0) \\ \kappa(1)p & 0 & 1-\kappa(1)p \end{bmatrix}.$$

We again assume that one-step costs are linear in the probability of violation, that is,

$$E[C(i)|\kappa(i)] = F_i + (1-\kappa(i))c$$

and that firms wish to minimise the long run discounted cost

$$E[C^*|\kappa] = E[C|\kappa] + \beta P_\kappa \cdot E[C^*|\kappa] = (I - \beta P_\kappa)^{-1} E[C|\kappa]$$

for a constant discount rate $\beta \in [0, 1)$.

3.3.2 Finding Optimal Policies

Under this model, $p_{ij}^{\kappa(i)}$ and $E[C(i)|\kappa(i)]$ are again linear in $\kappa(i)$, therefore by Lemma 3.1.4, an optimal policy will occur at a vertex – that is, where $\kappa(i) \in \{0, 1\}$ for all i.

It is intuitively clear that $\kappa(0)$ will never equal 1 unless $\kappa(-1)$ also equals 1. This is because a firm would never go to the effort of compliance in G_{-1} only to deprive itself of the reward of ending up in G_1. Mathematically, this is because when $\kappa(0) = 1$ then either $\{G_{-1}, G_0\}$ is a recurrent class (when $\kappa(-1) = 0$), or G_{-1} is an absorbing state (when $\kappa(-1) = 1$) which will be immediately reached from G_0. In the former case, $E[C^*(0)]$ is a weighted average of $\frac{F_0}{1-\beta}$ and $\frac{F_0+c}{1-\beta}$, in the latter case $E[C^*(0)] = \frac{F_0}{1-\beta}$, which is clearly lower.

Therefore, we have 6 possible policies to consider:

$$(\kappa(-1), \kappa(0), \kappa(1)) = \begin{cases} (0,0,0) \\ (0,0,1) \\ (1,0,0) \\ (1,0,1) \\ (1,1,0) \\ (1,1,1) \end{cases}$$

[8] Where states are ordered as (G_{-1}, G_0, G_1).

We consider each in turn.

For $(\kappa(-1), \kappa(0), \kappa(1)) = (0, 0, 0)$, we have

$$P_{(0,0,0)} = \begin{bmatrix} 1-g & g & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

and therefore

$$\begin{aligned} E[C^*] &= (I - \beta P)^{-1} E[C] \\ &= \begin{bmatrix} \frac{1+\beta g}{1-\beta(1-g)} F_0 + \frac{\beta^2 g}{1-\beta(1-g)} \frac{F_1}{1-\beta} + \frac{c}{1-\beta} \\ F_0 + c + \beta \frac{F_1+c}{1-\beta} \\ \frac{F_1+c}{1-\beta} \end{bmatrix}. \end{aligned}$$

For $(\kappa(-1), \kappa(0), \kappa(1)) = (0, 0, 1)$, we have

$$P_{(0,0,1)} = \begin{bmatrix} 1-g & g & 0 \\ 0 & 0 & 1 \\ p & 0 & 1-p \end{bmatrix}$$

and therefore

$$E[C^*] = \begin{bmatrix} \frac{(1+\beta g)(1-\beta(1-p))}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_0+c}{1-\beta} + \frac{\beta^2 g}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_1}{1-\beta} \\ \frac{\beta^2 p + (1-\beta(1-g))(1-\beta(1-p))}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_0+c}{1-\beta} + \beta \frac{1-\beta(1-g)}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_1}{1-\beta} \\ \beta \frac{p(1+\beta g)}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_0+c}{1-\beta} + \frac{1-\beta(1-g)}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_1}{1-\beta} \end{bmatrix}.$$

For $(\kappa(-1), \kappa(0), \kappa(1)) = (1, 0, 0)$, we have

$$P_{(1,0,0)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

and therefore

$$E[C^*] = \begin{bmatrix} \frac{F_0}{1-\beta} \\ F_0 + c + \beta \frac{F_1+c}{1-\beta} \\ \frac{F_1+c}{1-\beta} \end{bmatrix}.$$

For $(\kappa(-1), \kappa(0), \kappa(1)) = (1, 0, 1)$, we have

$$P_{(1,0,1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ p & 0 & 1-p \end{bmatrix}$$

and therefore

$$E[C^*] = \begin{bmatrix} \frac{F_0}{1-\beta} \\ \frac{\beta^2 p}{1-\beta(1-p)} \frac{F_0}{1-\beta} + F_0 + c + \beta \frac{F_1}{1-\beta(1-p)} \\ \frac{\beta p}{1-\beta(1-p)} \frac{F_0}{1-\beta} + \frac{F_1}{1-\beta(1-p)} \end{bmatrix}.$$

For $(\kappa(-1), \kappa(0), \kappa(1)) = (1, 1, 0)$, we have

$$P_{(1,1,0)} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and therefore

$$E[C^*] = \begin{bmatrix} \frac{F_0}{1-\beta} \\ \frac{F_0}{1-\beta} \\ \frac{F_1+c}{1-\beta} \end{bmatrix}.$$

For $(\kappa(-1), \kappa(0), \kappa(1)) = (1, 1, 1)$, we have

$$P_{(1,1,1)} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ p & 0 & 1-p \end{bmatrix}$$

and therefore

$$E[C^*] = \begin{bmatrix} \frac{F_0}{1-\beta} \\ \frac{F_0}{1-\beta} \\ \frac{\beta p}{1-\beta(1-p)} \frac{F_0}{1-\beta} + \frac{F_1}{1-\beta(1-p)} \end{bmatrix}.$$

To summarise:

Policy κ	$E[C^*(-1)\|\kappa]$	$E[C^*(0)\|\kappa]$
$(0,0,0)$	$\frac{1+\beta g}{1-\beta(1-g)}F_0 + \frac{\beta^2 g}{1-\beta(1-g)}\frac{F_1}{1-\beta} + \frac{c}{1-\beta}$	$F_0 + c + \beta\frac{F_1+c}{1-\beta}$
$(0,0,1)$	$\frac{A_{-1}}{A_{-1}+B_{-1}} \cdot \frac{F_0+c}{1-\beta} + \frac{B_{-1}}{A_{-1}+B_{-1}} \cdot \frac{F_1}{1-\beta}$	$\frac{A_0}{A_0+B_0} \cdot \frac{F_0+c}{1-\beta} + \frac{B_0}{A_0+B_0} \cdot \frac{F_1}{1-\beta}$
$(1,0,0)$	$\frac{F_0}{1-\beta}$	$F_0 + c + \beta\frac{F_1+c}{1-\beta}$
$(1,0,1)$	$\frac{F_0}{1-\beta}$	$\frac{\beta^2 p}{1-\beta(1-p)}\frac{F_0}{1-\beta} + F_0 + c + \beta\frac{F_1}{1-\beta(1-p)}$
$(1,1,0)$	$\frac{F_0}{1-\beta}$	$\frac{F_0}{1-\beta}$
$(1,1,1)$	$\frac{F_0}{1-\beta}$	$\frac{F_0}{1-\beta}$

Policy κ	$E[C^*(1)\|\kappa]$
$(0,0,0)$	$\frac{F_1+c}{1-\beta}$
$(0,0,1)$	$\frac{A_1}{A_1+B_1} \cdot \frac{F_0+c}{1-\beta} + \frac{B_1}{A_1+B_1} \cdot \frac{F_1}{1-\beta}$
$(1,0,0)$	$\frac{F_1+c}{1-\beta}$
$(1,0,1)$	$\frac{\beta p}{1-\beta(1-p)}\frac{F_0}{1-\beta} + \frac{F_1}{1-\beta(1-p)}$
$(1,1,0)$	$\frac{F_1+c}{1-\beta}$
$(1,1,1)$	$\frac{\beta p}{1-\beta(1-p)}\frac{F_0}{1-\beta} + \frac{F_1}{1-\beta(1-p)}$

where $A_{-1} = \beta(1+\beta g)(1-\beta(1-p))$, $B_{-1} = \beta^2 g$, $A_0 = \beta^2 p + (1-\beta(1-g))(1-\beta(1-p))$, $B_0 = \beta(1-\beta(1-g))$, $A_1 = p(1+\beta g)$ and $B_1 = 1-\beta(1-g)$ are all nonnegative weights.

If $(1,0,0)$ is optimal (in state G_1), then by Proposition 3.1.1

$$\begin{aligned}
\frac{F_1+c}{1-\beta} &\leq \frac{\beta p}{1-\beta(1-p)}\frac{F_0}{1-\beta} + \frac{F_1}{1-\beta(1-p)} \\
&\Leftrightarrow c \leq \frac{\beta p(F_0-F_1)}{1-\beta(1-p)} \\
&\Rightarrow c \leq \beta(F_0-F_1) \\
&\Leftrightarrow \frac{F_0}{1-\beta} \leq F_0 + c + \beta\frac{F_1+c}{1-\beta}
\end{aligned}$$

which implies $(1,1,0)$ is at least as good a policy in state G_0, and therefore we can ignore $(1,0,0)$.

Similarly, if $(1,1,0)$ is optimal (in state G_0), then

$$\begin{aligned}
\frac{F_0}{1-\beta} &\leq F_0 + c + \beta\frac{F_1+c}{1-\beta} \\
&\Leftrightarrow \beta(F_0-F_1) \leq c \\
&\Rightarrow \frac{\beta p(F_0-F_1)}{1-\beta(1-p)} \leq c \\
&\Leftrightarrow \frac{\beta p}{1-\beta(1-p)}\frac{F_0}{1-\beta} + \frac{F_1}{1-\beta(1-p)} \leq \frac{F_1+c}{1-\beta}
\end{aligned}$$

which implies $(1, 1, 1)$ is at least as good a policy in state G_1, and therefore we can ignore $(1, 1, 0)$.

3.3.3 Final Analysis

We now look only at the state G_0, as it is in this state that the expected costs from all the remaining policies are different.

Policy κ	$E[C^*(0)\|\kappa]$
$(0,0,0)$	$F_0 + c + \beta\frac{F_1+c}{1-\beta}$
$(0,0,1)$	$\frac{\beta^2 p+(1-\beta(1-g))(1-\beta(1-p))}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_0+c}{1-\beta} + \frac{\beta(1-\beta(1-g))}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{F_1}{1-\beta}$
$(1,0,1)$	$\frac{\beta^2 p}{1-\beta(1-p)}\frac{F_0}{1-\beta} + F_0 + c + \beta\frac{F_1}{1-\beta(1-p)}$
$(1,1,1)$	$\frac{F_0}{1-\beta}$

Differentiation with respect to c gives

Policy κ	$\partial E[C^*(0)\|\kappa]/\partial c$
$(0,0,0)$	$\frac{1}{1-\beta}$
$(0,0,1)$	$\frac{\beta^2 p+(1-\beta(1-g))(1-\beta(1-p))}{(1+\beta p)(1+\beta g)-\beta} \cdot \frac{1}{1-\beta}$
$(1,0,1)$	1
$(1,1,1)$	0

It is possible, (albeit tedious,) to show that this second term is non-increasing in g, non-decreasing in p, and therefore must take values in the range $\left[1, \frac{1}{1-\beta} - \beta\right]$. Therefore these policies are listed in order of decreasing slope with respect to c. Thus higher costs will lead a firm to increase the number of groups in which they violate – first to violating when in the Good group, then when in the Bad group, and finally when in the Intermediate Group. An example of this is shown in Figure 3.6.

Similarly, we could differentiate with respect to β, however the outcome of doing so is quite unmanageable. However, it is straightforward to plot the four policies for a range of β's, as is done in Figures 3.7 and 3.8.

Again, we shall plot the changes with respect to c and β simultaneously (code to do this is in C.2.9), as in Figure 3.9. (Here a darker colour again indicates a policy which entails more frequent violation.)

It is interesting to note that these figures display one significant difference between the insurance-based model and Harrington's Model – under the insurance-based model,

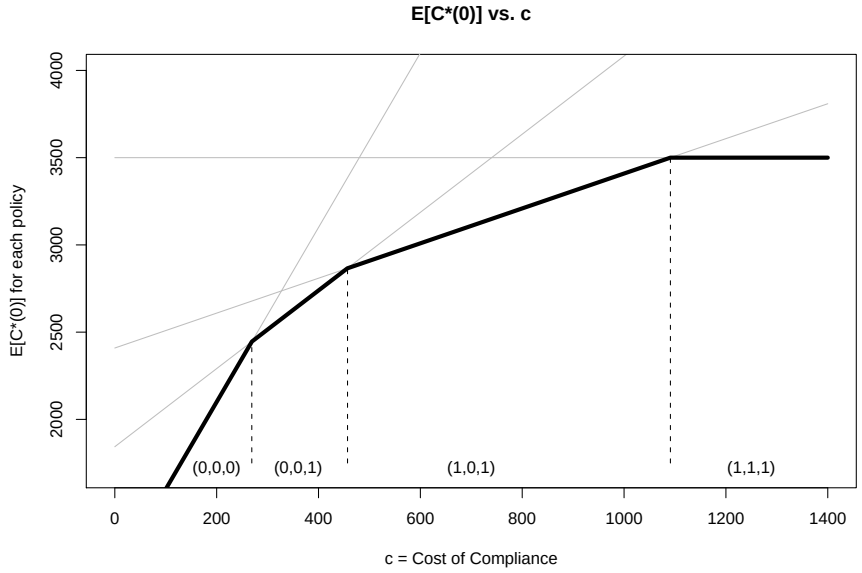

Figure 3.6: Indicative graph of $E[C^*(0)]$ vs c (see C.2.6 for details)

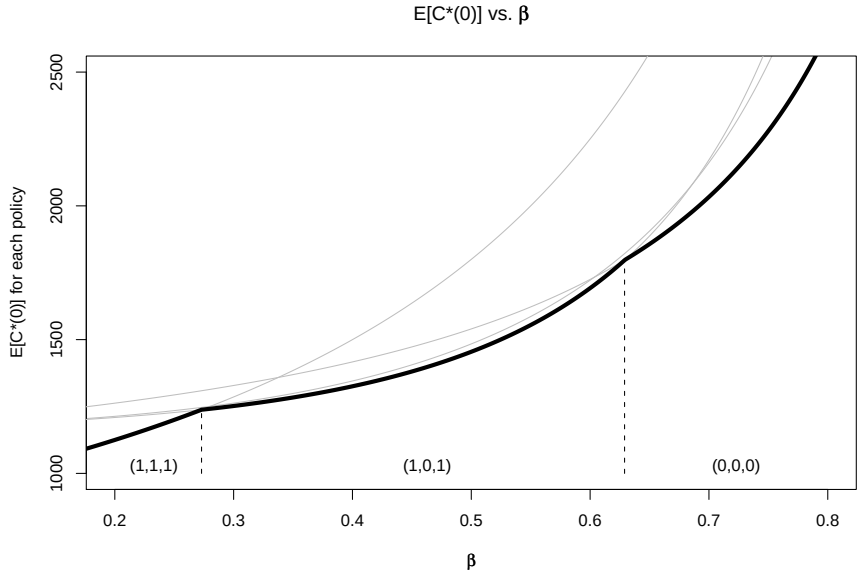

Figure 3.7: Indicative graph of $E[C^*(0)]$ vs β (see C.2.7 for details)

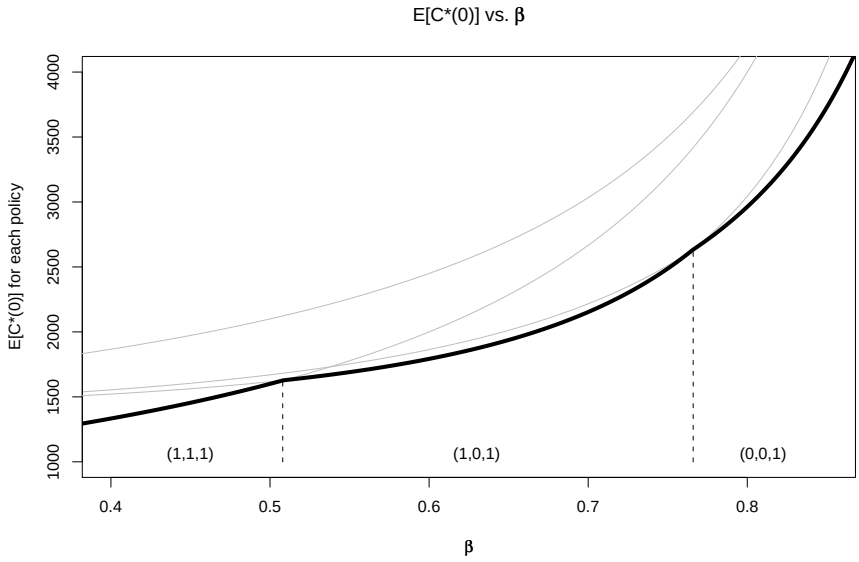

Figure 3.8: Indicative graph of $E[C^*(0)]$ vs β (see C.2.8 for details)

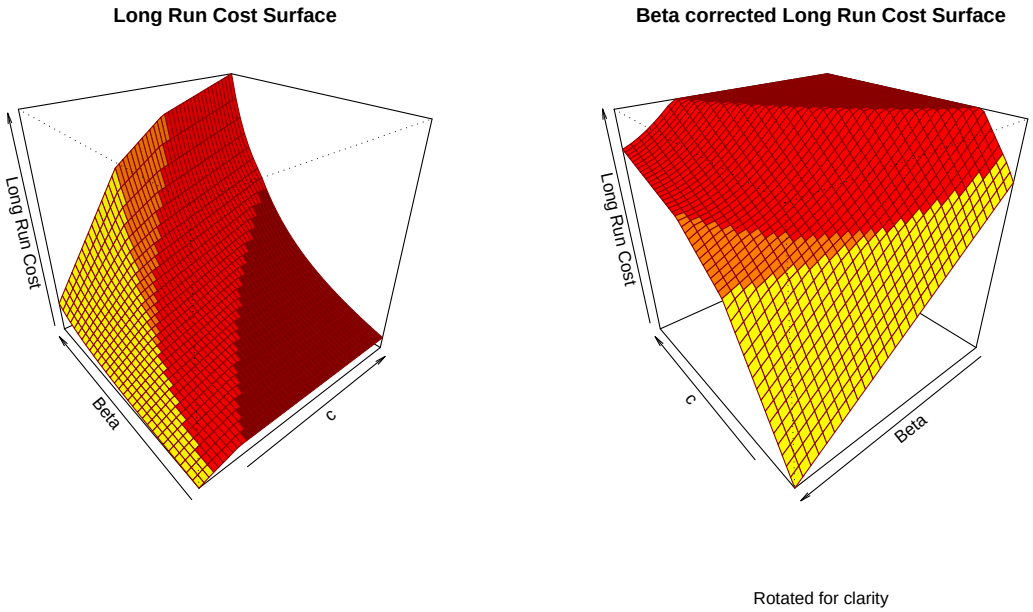

Figure 3.9: $E[C^*(0)]$ and $E[(1-\beta)C^*(0)]$ vs. c and β

it is possible for the discount factor to become significant enough that it will drive a non-compliant firm into perpetual compliance. This makes intuitive sense, as in this model the only punishment for violation today is the threat of a higher levy in the future, therefore a completely myopic firm (one with $\beta = 0$) will have no incentive to comply for any $c > 0$.

A related question is what happens as $\beta \to 1$. Remembering that the discounted cost converges (in some sense) as $\beta \to 1$ to the 'average' cost (the cost in each state multiplied by the steady state probabilities $\boldsymbol{\pi}$), we can determine when this will happen as follows:

The steady state probabilities under the four policies are:

- Under (0,0,0), $\boldsymbol{\pi} = (0, 0, 1)$.
- Under (0,0,1), $\boldsymbol{\pi} \propto (p, pg, g)$.
- Under (1,1,1) or (1,0,1), $\boldsymbol{\pi} = (1, 0, 0)$.

Therefore the optimal policy as $\beta \to 1$ will depend on which is the minimum of

$$F_1 + c,\ \tfrac{p(1+g)}{g+p(1+g)}(F_0 + c) + \tfrac{g}{g+p(1+g)}F_1 \text{ and } F_0.$$

Hence, as $\beta \to 1$,

- Policy (1,1,1) or (1,0,1) will be optimal if $c \geq \max\left\{\tfrac{g}{p(1+g)}(F_0 - F_1), F_0 - F_1\right\}$,
- Policy (0,0,0) will be optimal if $c \leq \min\left\{\tfrac{p(1+g)}{g}(F_0 - F_1), F_0 - F_1\right\}$,
- Policy (0,0,1) will be optimal otherwise.

Some policy implications are immediately clear:

- As in Harrington's model, a larger difference between F_0 and F_1 encourages compliance.
- A low value of g will make firms less likely to violate, and will encourage those firms that do violate partially to do so less of the time. A high value of p will make firms less likely to violate, as they will spend less time in G_1, and are more likely to choose to comply in G_1 in the hope of staying there. However these will also drive firms (particularly those with small values of β) to violate in all states.

- Firms with high costs will be more prone to violate, as will myopic firms.

It is also clear that provided $\beta \gg 0$, it is possible to approximately replicate the compliance profile under Harrington's model using this levy system.

One possible consequence of this analysis is that it would appear that a more directly punitive system (where firms pay dearly after violation) would entail higher levels of compliance than a system based on an insurance levy. However, this may not be possible for political or economic reasons, as it may result in businesses becoming unviable whenever a violation is detected. After all, as pointed out by Guo (1999, p. 47ff), the sanctions imposed by a regulator are preventative in spirit, and penalties should only be viewed as a 'necessary evil'. Firms are often quite vocal about "excessive" penalties, as these may prevent them from devoting adequate financial resources to abate real hazards.

3.4 Capturing Diminishing Returns

Returning to Harrington's (1988) original model, we now raise alternative objections.

- Violations are not binary
- Costs are not linear – there are diminishing marginal returns to effort

We can modify Harrington's model to capture these effects.

We model the firm as having an action space $\{\kappa \in [0, 1]\}$ (Note this set is compact). The cost (c) of pursuing such a strategy is monotonically decreasing[9] with κ, in particular we shall suppose that it is of the form $c = \alpha/\kappa$ for some $\alpha > 0$. (Note α may differ between firms.)

As we shall see in Chapter 6, a geometric distribution appears to be a reasonable approximation to the distribution of a firm's violations. For this reason, we shall model the number of violations V that occur as following the distribution

$$\Pr(V = v) = (1 - \kappa)\kappa^v,$$

which is a form of the geometric distribution with 'failure probability' κ. Thus, a high value of κ implies a higher expected number of violations.

[9] This is because we assume that it costs less to have a higher rate of violation. A possible modification to this model would be to consider c to be decreasing up to a point, after which it is increasing – behaviour which corresponds to a firm having to pay a premium to employees if the risks of working are too high.

This model captures 'diminishing marginal returns to effort in compliance' – it becomes increasingly difficult for a firm to reduce the rate at which it will violate. In other words, if a firm is already at a low level of violation, further preventative measures will be significantly more costly than for a firm at a high level of violation. It also shows that it is impossible for a firm to completely eliminate violations ($\kappa = 0$), as doing so will require an infinite expenditure on safety ($c = \alpha/\kappa = \infty$).

When in G_0 a firm is inspected in such a way that each violation is detected independently with probability ϕ_0, when in G_1 a firm is inspected in such a way that each violation is detected independently with probability ϕ_1. Therefore the number D of violations detected follows the distribution

$$\Pr(D = d|V = v) = \binom{v}{d}\phi_i^d(1-\phi_i)^{v-d},$$

and so

$$\begin{aligned}
\Pr(D = d) &= \sum_v \Pr(D = d|V = v)\Pr(V = v) \\
&= \sum_{v=0}^{\infty}\binom{v}{d}\phi_i^d(1-\phi_i)^{v-d}(1-\kappa(i))\kappa(i)^v \\
&= \frac{(1-\kappa(i))\phi_i^d\kappa(i)^d}{(1-(1-\phi_i)\kappa(i))^{d+1}} \\
&= \frac{1-\kappa(i)}{1-(1-\phi_i)\kappa(i)}\cdot\left(\frac{\phi_i\kappa(i)}{1-(1-\phi_i)\kappa(i)}\right)^d \\
&=: (1-\lambda_i)\lambda_i^d.
\end{aligned}$$

That is, D also follows a geometric distribution.

Now suppose that when the number of detected violations equals d, a firm in G_1 will be moved to G_0 with a probability

$$1-(1-\gamma_1)^d$$

and that a firm in G_0 will be moved to G_1 with a probability

$$(1-\gamma_0)^d.$$

Large values of γ make firms more likely to be placed in G_0, however (provided we define $0^0 = 1$ when needed) a firm that does not violate will invariably be placed in G_1 – this scheme is very forgiving[10]. Also, as we would hope, the higher the value of d, the more likely a firm is to be placed in G_0. We shall assume $0 \le \gamma_1 \le \gamma_0 \le 1$.

[10]In fact, γ_i can be interpreted as the probability that the regulator will not overlook a violation when determining where to place a firm (that was in G_i) for the next period, when violations are considered independently.

If T denotes the event that a firm will be in G_1 in the following period, then

$$\begin{aligned}\Pr(T|D=d) &= (1-\gamma_i)^d \\ \Pr(T) &= \sum_d \Pr(T|D=d)\Pr(D=d) \\ &= \sum_d (1-\gamma_i)^d(1-\lambda_i)\lambda_i^d \\ &= \frac{1-\lambda_i}{1-(1-\gamma_i)\lambda_i} \\ &= \frac{1-\kappa(i)}{1-(1-\phi_i\gamma_i)\kappa(i)}.\end{aligned}$$

Thus we have a transition matrix

$$P_\kappa = \begin{bmatrix} \frac{\phi_0\gamma_0\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)} & \frac{1-\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)} \\ \frac{\phi_1\gamma_1\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)} & \frac{1-\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)} \end{bmatrix}.$$

Finally, suppose that each detected offence is fined an amount F_i when in G_i, that is, DF_i is the total fine issued[11]. Then the expected fine is

$$E[D]F_i = \left(\frac{1-(1-\phi_i)\kappa(i)}{1-\kappa(i)} - 1\right)F_i = \frac{\kappa(i)}{1-\kappa(i)}\phi_i F_i.$$

Therefore the total immediate expected cost is

$$E[C(i)] = \frac{\kappa(i)}{1-\kappa(i)}\phi_i F_i + \frac{\alpha}{\kappa(i)}.$$

This model has a nice symmetry to it, but it is non-linear (this is both its strength and its weakness) and therefore we will not be able to exploit Lemma 3.1.4 in the same way as earlier. However, it is clear that neither vertex ($\kappa \in \{0,1\}$) will give an optimal solution, as at these values the expected costs are infinite[12]. We can therefore immediately claim that an optimal solution must be where the derivative of $E[C^*]$ with respect to $\kappa(i)$ is zero.

[11] More generally, we can assume that the expected amount of a fine is F_i, and that the size of an individual fine is independent of the number of violations detected.

[12] This also ensures that we will not have difficulties regarding manipulation of a 'Geometric distribution with failure probability 1'.

We can now write

$$\begin{aligned}
E[C^*] &= (I-\beta P)^{-1}E[C] \\
&= \begin{bmatrix} 1-\beta\frac{\phi_0\gamma_0\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)} & -\beta\frac{1-\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)} \\ -\beta\frac{\phi_1\gamma_1\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)} & 1-\beta\frac{1-\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)} \end{bmatrix}^{-1} \cdot \begin{bmatrix} \frac{\kappa(0)}{1-\kappa(0)}\phi_0F_0+\frac{\alpha}{\kappa(0)} \\ \frac{\kappa(1)}{1-\kappa(1)}\phi_1F_1+\frac{\alpha}{\kappa(1)} \end{bmatrix} \\
&= \frac{1}{1-\beta}\cdot\frac{1}{1+\beta\left(\frac{\phi_1\gamma_1\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)}-\frac{\phi_0\gamma_0\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)}\right)} \\
&\quad\times\begin{bmatrix} 1-\beta\frac{1-\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)} & \beta\frac{1-\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)} \\ \beta\frac{\phi_1\gamma_1\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)} & 1-\beta\frac{\phi_0\gamma_0\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)} \end{bmatrix} \cdot \begin{bmatrix} \frac{\kappa(0)}{1-\kappa(0)}\phi_0F_0+\frac{\alpha}{\kappa(0)} \\ \frac{\kappa(1)}{1-\kappa(1)}\phi_1F_1+\frac{\alpha}{\kappa(1)} \end{bmatrix}.
\end{aligned}$$

By Proposition 3.1.1, we can justifiably restrict our attention to $E[C^*(1)]$, and so we wish to select $\kappa(0), \kappa(1)$ to minimise

$$E[C^*(1)] = \frac{A}{1-\beta}\cdot\left(\frac{\kappa(0)}{1-\kappa(0)}\phi_0F_0+\frac{\alpha}{\kappa(0)}\right)+\frac{1-A}{1-\beta}\left(\frac{\kappa(1)}{1-\kappa(1)}\phi_1F_1+\frac{\alpha}{\kappa(1)}\right),$$

where $A = \frac{\beta\frac{\phi_1\gamma_1\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)}}{1+\beta\left(\frac{\phi_1\gamma_1\kappa(1)}{1-(1-\phi_1\gamma_1)\kappa(1)}-\frac{\phi_0\gamma_0\kappa(0)}{1-(1-\phi_0\gamma_0)\kappa(0)}\right)}$.

While analytically intractable, this equation is quite easy to solve numerically for given values of ϕ_1, F_1, etc..., and we can plot the long term cost and the optimal policy $(\kappa(0), \kappa(1))$ against α and β. This is done in Figures 3.10 and 3.11.

Some interesting observations can be made here. Generally the results resemble those in Harrington's model, in particular as α increases a firm will violate more, and as β increases firms violate less. On the other hand, in this model we see that this may happen even if it results in the β-adjusted long run cost $E[(1-\beta)C^*]$ to be increasing in β.

3.5 Further extensions

A variety of further extensions to this model are possible.

One of these is by Stafford (2006), who incorporates the effects of self-policing into Harrington's model. Similar ideas are present in Livernois and McKenna (1999), who incorporate self-policing into a simpler enforcement model. A good overview of the economic literature on self policing can be found in Stafford (2005).

Friesen (2003) develops an alternative formulation of the targeted enforcement model, based on random shifts between groups, and shows that this can lead to higher levels of

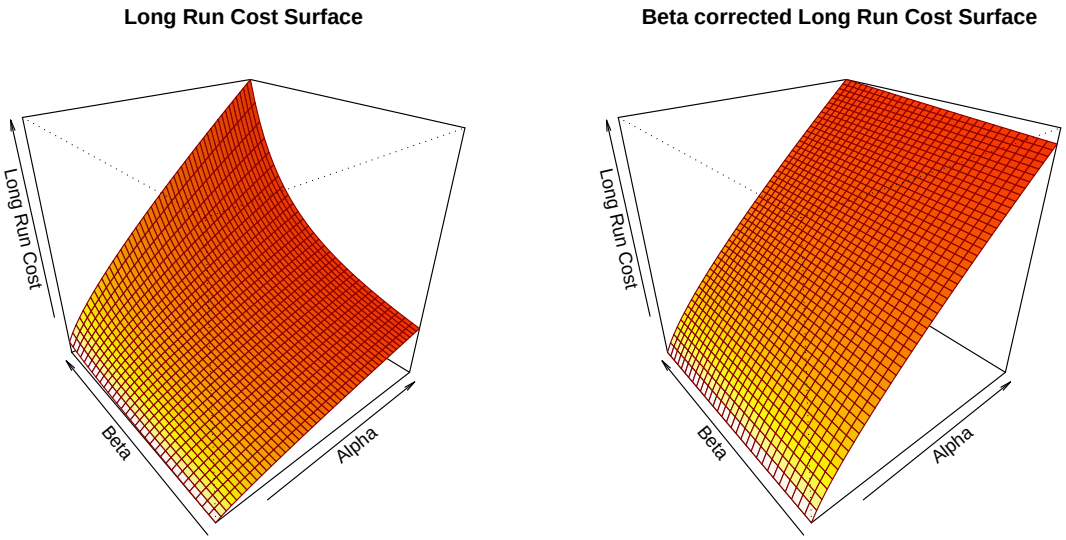

Figure 3.10: $E[C^*(1)]$ and $E[(1-\beta)C^*(1)]$ vs. α and β

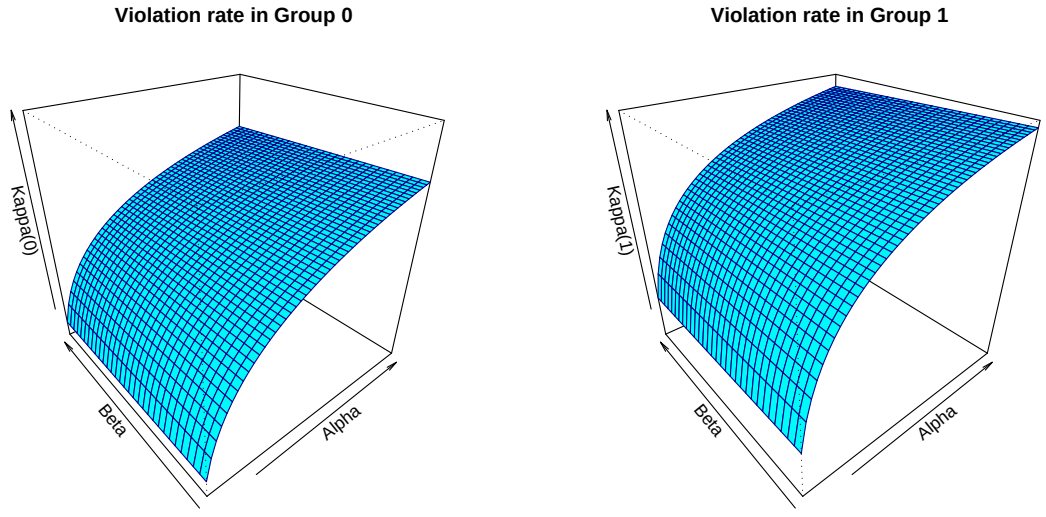

Figure 3.11: Optimal $\kappa(0)$ and $\kappa(1)$ vs. α and β

compliance. Clark, Friesen and Muller (2004) then outline empirical evidence to show the effectiveness of such a scheme in a controlled experiment.

Another extension is for the regulator to not inform the firm which group they are in, and instead allow them to infer this from the regulator's actions. Harrington mentions this issue in passing, where he comments that some time may need to pass before a firm realises which group it is in. In this case, we can model a firm as a Bayesian learner, with some prior opinion as to which state they are in. This is updated in each period depending on the actions of the regulator and results in the estimated probabilities of being in each group following a Markov Chain with countably infinite state space. (Such a situation is often referred to as a 'Partially Observable Markov Chain'.) Unfortunately, the resulting equations are quite difficult to manipulate, and will not be further explored here.

Chapter 4

Estimation with Missing Information

" 'Other maps are such shapes, with their islands and capes!
But we've got our brave Captain to thank'

(So the crew would protest) 'that he's brought us the best –
A perfect and absolute blank!' "

Lewis Carroll
The Hunting of the Snark (1876)

We now move on from these simple economic models to investigating the efficiency of a regulator. In particular we wish to determine empirically what types of firms are, in practice, targeted for higher levels of enforcement, and what factors contribute to a firm violating at a higher rate. In the previous chapter we disregarded these issues, instead assuming that the regulator was not able to differentiate firms with high costs and discount factors from others. Clearly, such information would allow us to determine better models of enforcement, and to ensure that a regulator is acting efficiently.

To estimate parameters for models of regulatory systems is not a trivial task. This is because it is not immediately apparent how to distinguish between firms which truly violate more, and those who are simply detected more frequently. We will approach it using the likelihood principle; as espoused by Fisher (1922) and others. To do this, the following analysis (loosely following Dempster et al. (1977)) is needed.

4.1 Maximum Likelihood Estimation

The general principle is this. Consider a set of data $\mathbf{x}$ with support in some space $\mathcal{X}$. This arises from a distribution with likelihood function $L_c(\boldsymbol{\phi}) := f(\mathbf{x}|\boldsymbol{\phi}, \{\boldsymbol{\xi}_i\})$, for some (unknown) parameters $\boldsymbol{\phi}$ in a space Φ, and known covariates $\{\boldsymbol{\xi}_i\}$ (which we can assume to be fixed). We define the **(complete-data) maximum-likelihood estimator** to be

$$\underset{\boldsymbol{\phi}\in\Phi}{\operatorname{argmax}}\{L_c(\boldsymbol{\phi})\}. \tag{4.1}$$

In general, we will have data which is assumed to be independent and identically distributed (conditional on $\boldsymbol{\phi}$ and $\boldsymbol{\xi}_i$). We can therefore express the (complete-data) likelihood in the form[1]

$$L_c(\boldsymbol{\phi}) = f(\mathbf{x}|\boldsymbol{\phi}, \{\boldsymbol{\xi}_i\}) = \prod_i \tilde{f}(x_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) \tag{4.2}$$

or alternately

$$\ln L_c(\boldsymbol{\phi}) = \ln f(\mathbf{x}|\boldsymbol{\phi}, \{\boldsymbol{\xi}_i\}) = \sum_i \ln \tilde{f}(x_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i). \tag{4.3}$$

Note that the maximum-likelihood estimator in (4.1) is also the value which maximises (4.3).

However, in our case, we do not observe $\mathbf{x}$ directly. Rather we observe some (measurable) function of the data $\mathbf{y} := \mathbf{y}(\mathbf{x}), \mathcal{X} \to \mathcal{Y}$. Therefore, we instead define the **(incomplete-data) maximum likelihood estimator** to be

$$\hat{\boldsymbol{\phi}} := \underset{\boldsymbol{\phi}\in\Phi}{\operatorname{argmax}}\{L_m(\boldsymbol{\phi})\} \tag{4.4}$$

with

$$L_m(\boldsymbol{\phi}) := g(\mathbf{y}|\boldsymbol{\phi}, \{\boldsymbol{\xi}_i\}) = \int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x}|\boldsymbol{\phi}, \{\boldsymbol{\xi}_i\}) d\mathbf{x} \tag{4.5}$$

where $\mathcal{X}(\mathbf{y})$ denotes the preimage in $\mathcal{X}$ of $\mathbf{y}$. (Of course, this is integration with respect to an appropriate 'uniform' measure on $\mathcal{X}$, as will become important later, as much of our data will be discrete[2].)

[1] Where $\tilde{f}$ is the probability (density) function for an individual data point.

[2] This approach is prefered to the alternate approach of considering the missing data as parameters to be estimated. Reasons for this are discussed in Little and Rubin (1983). However, our approach depends on the assumption that there is no relationship between values of the missing data and the *a priori* probability that the data will be missing. As in our example the missing data is simply unobservable *no matter what its value*, this assumption holds trivially – as is pointed out by Little and Schluchter (1985).

Therefore, assuming appropriate continuity conditions, we can express the missing-data likelihood function as

$$\begin{aligned} L_m(\boldsymbol{\phi}) = g(\mathbf{y}|\boldsymbol{\phi}, \{\boldsymbol{\xi}_i\}) &= \int_{\mathcal{X}(\mathbf{y})} f(\mathbf{x}|\boldsymbol{\phi}, \{\boldsymbol{\xi}_i\}) d\mathbf{x} \\ &= \int_{\mathcal{X}(y_1)} \int_{\mathcal{X}(y_2)} \dots \int_{\mathcal{X}(y_n)} \prod_i \tilde{f}(x_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) dx_1 dx_2 \dots dx_n \\ &= \prod_i \int_{\mathcal{X}(y_i)} \tilde{f}(x_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) dx_i, \end{aligned}$$

and so we can write

$$\ln L_m(\boldsymbol{\phi}) = \ln g(\mathbf{y}|\boldsymbol{\phi}, \boldsymbol{\xi}_i) = \sum_i \ln \tilde{g}(y_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) \tag{4.6}$$

with

$$\tilde{g}(y_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) = \int_{\mathcal{X}(y_i)} \tilde{f}(x_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) dx_i. \tag{4.7}$$

Of course, the exact details of how this will appear will vary with the model used. The estimators so defined are not necessarily unbiased or "minimum-variance" estimators, however under reasonable conditions can be shown to be be consistent (converge in probability to the true value) and efficient (their variance converges to the Cramér-Rao lower bound) – see Lehmann (1983, p. 414) for details.

This immediately gives us a method of estimating the parameters – we simply find a maximum of the likelihood function, which can be done using any number of numerical techniques.

4.1.1 The EM Algorithm

The EM (Expectation-Maximisation) Algorithm is a method of finding maximum likelihood estimates without having to calculate the missing data likelihood function (4.5). It was proposed by Dempster et al. (1977), and is widely used in these situations. However, for the following three reasons we will not use it here:

- In our current context, the incomplete data likelihood function often has a simple closed form, and so the use of the EM algorithm is an unnecessary complication.
- As we will mainly be considering generalised linear models, which typically need iterative numerical methods for estimation, we may as well leap directly to the iterative method and avoid further complications.
- The EM algorithm has the drawback that it does not immediately give an indication of the errors associated with the estimates it produces, whereas these can be estimated numerically as outlined in the following section.

4.1.2 Estimation of Errors

Using the above theory, we have a method of estimating the parameters of a general model. We also need to have some idea of the errors associated with these estimates. The fundamental entity that we shall use here is the Observed Information Matrix

$$\mathcal{I}(\hat{\boldsymbol{\phi}}) = -\left[\left.\frac{\partial^2 \ln L_m(\boldsymbol{\phi})}{\partial\phi_i \partial\phi_j}\right|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}\right].$$

As L_m is smooth in the models we consider, this matrix is clearly symmetric (by equality of mixed partial derivatives). We shall see that it can act as an estimate of the inverse variance-covariance matrix of the estimates $\hat{\boldsymbol{\phi}}$. The following is only an outline, a full derivation with technical details can be found in Lehmann (1983, p. 125ff and p. 429ff).

First, we denote by $\boldsymbol{\phi}_0$ the 'true' value of the parameters to be estimated. Provided the model is well identified, i.e.

$$\tilde{g}(y|\boldsymbol{\phi}_1) = \tilde{g}(y|\boldsymbol{\phi}_2) \text{ a.e.} \Rightarrow \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2$$

then the weak law of large numbers implies (as the y_i are independent conditional on $\boldsymbol{\xi}_i$) that

$$\begin{aligned}\frac{1}{n}\ln L_m(\boldsymbol{\phi}) &= \frac{1}{n}\sum_i \ln \tilde{g}(y_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) \\ &\to E[\ln \tilde{g}(y_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i)]\end{aligned}$$

where convergence is with respect to an appropriate probability measure[3] as $n \to \infty$. This is uniquely maximised by the true value of $\boldsymbol{\phi}_0$ as it is with respect to the probabilities defined by these parameters that the expectation is taken (this follows from Jensen's Inequality, see Lehmann (1983, p. 409)). Hence, as we are choosing $\hat{\boldsymbol{\phi}}$ to maximise $\ln L_m$, our estimators are consistent (converge in probability to the true value).

A key quantity is the 'score' function. This is a (vector-valued) function, and is defined as

$$S(\boldsymbol{\phi}) = \frac{d \ln L_m}{d\boldsymbol{\phi}}.$$

We can take a (multivariate) linear Taylor approximation to the score function around $\boldsymbol{\phi}_0$. As our estimate is consistent, this will be a good approximation (at least asymptotically) and has the form (considering $\boldsymbol{\phi}$ as a column vector)

$$S(\boldsymbol{\phi}) \approx S(\boldsymbol{\phi}_0) + S'(\boldsymbol{\phi}_0) \cdot (\boldsymbol{\phi} - \boldsymbol{\phi}_0)$$

[3]It is easier to understand this if we consider a random effects model – i.e. where the $\boldsymbol{\xi}_i$ is also a random variable. Otherwise, we need to be careful in how we define the concept of an infinite sample.

where S' is a matrix of second derivatives of the log-likelihood, that is, $-S'(\hat{\boldsymbol{\phi}}) = \mathcal{I}(\hat{\boldsymbol{\phi}})$.

As our likelihood L_m is smooth, at its maximum value it will have zero derivative[4] (that is, $S(\boldsymbol{\phi}_0) = \mathbf{0}$). Therefore

$$\boldsymbol{\phi} = \boldsymbol{\phi}_0 - \mathcal{I}(\boldsymbol{\phi}_0)^{-1} S(\boldsymbol{\phi}).$$

Taking expectations and variances, this indicates that asymptotically $\boldsymbol{\phi}$ has expected value $\boldsymbol{\phi}_0$ and variance $\mathcal{I}(\boldsymbol{\phi}_0)^{-1} V \mathcal{I}(\boldsymbol{\phi}_0)^{-1}$ where V is the variance of $S(\boldsymbol{\phi})$. This can be determined to be

$$\begin{aligned} V &= E[S(\boldsymbol{\phi})^T S(\boldsymbol{\phi})] \\ &= E\left[\frac{\partial \ln L_m}{\partial \phi_i} \cdot \frac{\partial \ln L_m}{\partial \phi_j}\right] \end{aligned}$$

which under appropriate regularity conditions can be shown to be equal to

$$-E\left[\frac{\partial^2 \ln L_m}{\partial \phi_i \partial \phi_j}\right]$$

which is the 'expected information matrix'. It is now possible to show that, asymptotically, our observed information matrix will converge in some sense to its expectation (through the law of large numbers), and that $\boldsymbol{\phi}$ is asymptotically normally distributed (through the central limit theorem) – see Lehmann (1983) for details. Therefore, it is the case that (asymptotically)

$$\hat{\boldsymbol{\phi}} \sim N(\boldsymbol{\phi}_0, \mathcal{I}(\boldsymbol{\phi}_0)^{-1}) \approx N(\boldsymbol{\phi}_0, \mathcal{I}(\hat{\boldsymbol{\phi}})^{-1}).$$

4.2 Detection Controlled Estimation

Detection Controlled Estimation (DCE) is simply a special case of Missing-Data Estimation. It was proposed by Feinstein (1990) as a means of accounting for imperfect detection of violations of regulations. The model assumes that data is generated in the following way:

- Stage 1: Violations of regulations occur.
- Stage 2: For each violation that has occurred, there is a chance that it will be detected. Only detected violations are known by us.

Mathematically, we presuppose the existence of data $\mathbf{x} = (\mathbf{V}, \mathbf{D})$ where $\mathbf{V}$ represents violations and $\mathbf{D}$ represents detections (at this stage, both binary variables[5]). We know

[4] In general, our parameter space will be a manifold (and therefore open), and our likelihood will be such that it has an attainable maximum with probability one.

[5] A simple generalisation of this model to account for partial detection would allow $\mathbf{D}$ to vary over the range $[0, 1]$. Generalising for $\mathbf{V}$ is done in Section 4.4.

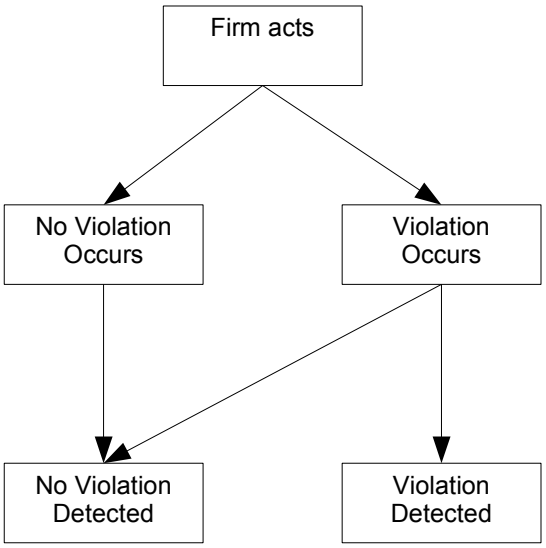

Figure 4.1: The DCE Framework

that there exists a (stochastic) relationship between $\mathbf{V}$ and $\mathbf{D}$. However, we observe only $\mathbf{D}$. Our aim is to determine properties of the random variables $\mathbf{V}$ and $\mathbf{D}$. We proceed to explore this model for a simple case.

4.3 Binary Choice Models

We consider the case where violation and detection is a binary variable – an 'either-or' situation. We shall define

$$V_i := \begin{cases} 1 & \text{if firm } i \text{ has violated} \\ 0 & \text{otherwise,} \end{cases}$$

$$D_i := \begin{cases} 1 & \text{if firm } i \text{ is detected} \\ 0 & \text{otherwise.} \end{cases}$$

As above, we assume that the probabilities of these events depend on some covariates $\boldsymbol{\xi}_i, \boldsymbol{\eta}_i \in \Xi$, where $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ may share values. We propose functions U and W which map $\Xi \to [0,1]$, and therefore propose the general model

$$\begin{aligned} \Pr(V_i = 1) &=: U(\boldsymbol{\xi}_i), \\ \Pr(D_i = 1 | V_i) &=: \begin{cases} W(\boldsymbol{\eta}_i) & \text{if } V_i = 1 \\ 0 & \text{if } V_i = 0. \end{cases} \end{aligned}$$

We then have a series of observations $\{D_i\}$. To find the likelihood function, we note that, given there is no way for a false detection to arise in this framework, we can

determine that

$$\begin{aligned}\mathcal{X}(1) &= \{(1,1)\}\\ \mathcal{X}(0) &= \{(0,0),(1,0)\},\end{aligned}$$

that is, a detection can only arise if a violation occurs and is detected, and no detection can only arise either if a violation occurs and is not detected, or if no violation occurs.

Hence, we can express (4.7) as

$$\tilde{g}(D_i) = \begin{cases} \tilde{f}((1,1)) & \text{if } D_i = 1,\\ \tilde{f}((0,0)) + \tilde{f}((1,0)) & \text{if } D_i = 0.\end{cases} \tag{4.8}$$

Therefore, the (missing-data) maximum-likelihood estimator of parameters[6] ($\hat{\boldsymbol{\phi}}$) is

$$\begin{aligned}\hat{\boldsymbol{\phi}} &= \operatorname*{argmax}_{\boldsymbol{\phi}} \left\{ \sum_i \ln \tilde{g}(D_i|\boldsymbol{\phi}, \boldsymbol{\xi}_i) \right\}\\ &= \operatorname*{argmax}_{\boldsymbol{\phi}} \left\{ \sum_{\{i:D_i=1\}} \ln \tilde{f}((1,1)|\boldsymbol{\phi}, \boldsymbol{\xi}_i) + \sum_{\{i:D_i=0\}} \ln(\tilde{f}((0,0)|\boldsymbol{\phi}, \boldsymbol{\xi}_i) + \tilde{f}((1,0)|\boldsymbol{\phi}, \boldsymbol{\xi}_i)) \right\}.\end{aligned}$$

Now considering $\tilde{f}$, we find

$$\begin{aligned}\tilde{f}((1,1)|\boldsymbol{\phi}, \boldsymbol{\xi}_i) &= \Pr(V_i = 1, D_i = 1)\\ &= \Pr(V_i = 1)\Pr(D_i = 1|V_i = 1)\\ &= U(\boldsymbol{\xi}_i)W(\boldsymbol{\eta}_i),\\ \tilde{f}((0,0)|\boldsymbol{\phi}, \boldsymbol{\xi}_i) &= \Pr(V_i = 0, D_i = 0)\\ &= \Pr(V_i = 0)\Pr(D_i = 0|V_i = 0)\\ &= (1 - U(\boldsymbol{\xi}_i)) \times 1,\\ \tilde{f}((1,0)|\boldsymbol{\phi}, \boldsymbol{\xi}_i) &= \Pr(V_i = 1, D_i = 0)\\ &= \Pr(V_i = 1)\Pr(D_i = 0|V_i = 1)\\ &= U(\boldsymbol{\xi}_i)(1 - W(\boldsymbol{\eta}_i)).\end{aligned}$$

Hence (with $U := U(\boldsymbol{\xi}_i)$ and $W := W(\boldsymbol{\eta}_i)$ for simplicity)

$$\begin{aligned}\hat{\boldsymbol{\phi}} &= \operatorname*{argmax}_{\phi} \left\{ \sum_{\{i:D_i=1\}} \ln UW + \sum_{\{i:D_i=0\}} \ln(1 - U + U(1 - W)) \right\}\\ &= \operatorname*{argmax}_{\phi} \left\{ \sum_{\{i:D_i=1\}} \ln UW + \sum_{\{i:D_i=0\}} \ln(1 - UW) \right\}.\end{aligned}$$

[6] At this stage, $\boldsymbol{\phi}$ represents some properties of U and W.

4.3.1 Single Index Models

For the purposes of estimation, this model is still far too general[7]. A simple way of reducing the generality of the model is to require U and W to be of a particular form. For our purposes, we shall assume $U(\boldsymbol{\xi}_i) = F(\boldsymbol{\xi}_i\boldsymbol{\beta}_U)$ and $W(\boldsymbol{\eta}_i) = G(\boldsymbol{\eta}_i\boldsymbol{\beta}_W)$ for some real-valued vectors $\boldsymbol{\beta}_U$ and $\boldsymbol{\beta}_W$, where $F, G : \mathbb{R} \to [0, 1]$. It is important to note that as we have not yet restricted what can be included in $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_i$, this is still quite general. (For example, a finite Taylor approximation to any function of some variable x_j can be estimated, simply by including appropriate powers of x_j in $\boldsymbol{\xi}_i$ – similarly for Fourier or other approximations.) A model with this assumption is called a single index model, as the probability of violation depends on a single value $\boldsymbol{\xi}_i\boldsymbol{\beta}_U$ (and similarly for detection).

If we then specify F and G (typically these are cumulative distribution functions, particularly logistic or normal) we have a model which we can attempt to estimate. In this case, $\boldsymbol{\phi} = (\boldsymbol{\beta}_U, \boldsymbol{\beta}_W)$, and so we try and find values to maximise

$$\sum_{\{i:D_i=1\}} \ln F(\boldsymbol{\xi}_i\boldsymbol{\beta}_U)G(\boldsymbol{\eta}_i\boldsymbol{\beta}_W) + \sum_{\{i:D_i=0\}} \ln(1 - F(\boldsymbol{\xi}_i\boldsymbol{\beta}_U)G(\boldsymbol{\eta}_i\boldsymbol{\beta}_W)) \tag{4.9}$$

Given certain restrictions on F,G, $\boldsymbol{\beta}_U$ and $\boldsymbol{\beta}_W$, and certain properties of $\boldsymbol{\xi}_i$ and $\boldsymbol{\eta}_i$, this equation has a unique maximum. (See Chapter 5 for further details and discussion.)

As an aside, if we do not wish to restrict F and G to be of a certain form[8], we can follow the ideas of Carroll, Fan, Gijbels and Wand (1997) and assume that their product

$$F(\boldsymbol{\xi}_i\boldsymbol{\beta}_U)G(\boldsymbol{\xi}_i\boldsymbol{\beta}_W) = \text{logit}^{-1}(\mu_F(\boldsymbol{\xi}_i\boldsymbol{\beta}_F))\,\text{logit}^{-1}(\mu_G(\boldsymbol{\eta}_i\boldsymbol{\beta}_G))$$

for some locally linear function[9] μ_F, μ_G. Using local quasi-likelihood methods (see Carroll et al. (1997) for indications as to how this might work), we can then attempt to estimate $\mu_F, \mu_G, \boldsymbol{\beta}_F$ and $\boldsymbol{\beta}_G$. As will become clear through Chapter 5, we will have to be quite careful when applying such techniques, as it is likely that we will not be able to determine the required parameters uniquely.

4.4 Poisson Process Models

As discussed by Guo (1999), in the context of OH&S, this simple binary choice model does not accurately reflect reality. There are typically multiple violations detected on

[7]This is another example of Bellman's (1961) 'curse of dimensionality' – by using a 'single index model', we lose some flexibility but significantly simplify the problem statistically. This is particularly important if we choose to use a nonparametric or semiparametric variation on the themes explored here.

[8]That is, we wish to reintroduce the flexibility of a nonlinear relationship with $\boldsymbol{\xi}_i$.

[9]This obviously restricts F and G to be continuous functions, however, in limit can approach any function.

a single inspection, and the number detected varies considerably. A more appropriate model, in which we can still apply the DCE framework, is therefore a Poisson process model[10], where violations arise as the result of a Poisson process, and are then detected independently.

Mathematically, the new model is as follows

$$\begin{aligned} V_i &\sim \text{Po}(U(\boldsymbol{\xi}_i)) \\ D_i|V_i &\sim \text{Bin}(V_i, W(\boldsymbol{\eta}_i)) \end{aligned}$$

where $U : \mathbb{R} \to [0, \infty)$. As earlier, we will proceed assuming that U and W can be written in the form $U(\boldsymbol{\xi}_i) = \lambda(\boldsymbol{\xi}_i\boldsymbol{\beta}_U)$ and $W(\boldsymbol{\eta}_i) = G(\boldsymbol{\eta}_i\boldsymbol{\beta}_W)$. This construction mimics closely the simpler Binary-Choice model.

We now need to derive the incomplete data likelihood function, as this will allow us to estimate the required parameters without the need for the EM Algorithm.

First we note that with complete data $\mathbf{x} = (\mathbf{V}, \mathbf{D})$, we have

$$\mathcal{X}(i) = \{(j, i) : j \in \{i, i + 1, ...\}\}$$

(that is, the number of violations must be an integer that is at least as large as the number of detected violations).

Hence we can express (4.7) (with $\lambda := \lambda(\boldsymbol{\xi}_i\boldsymbol{\beta}_U)$ and $G := G(\boldsymbol{\eta}_i\boldsymbol{\beta}_W)$ for simplicity) as

$$\begin{aligned} g(D_i) &= \sum_{j \geq D_i} \Pr(V_i = j) \Pr(D_i|V_i = j) \\ &= \sum_{j \geq D_i} \frac{e^{-\lambda}\lambda^j}{j!} \begin{pmatrix} j \\ D_i \end{pmatrix} G^{D_i}(1 - G)^{j-D_i} \\ &= \frac{e^{-\lambda G}(\lambda G)^{D_i}}{D_i!}. \end{aligned}$$

Hence, in the absence of information about V_i, we find D_i is also Poisson distributed with rate

$$\lambda(\boldsymbol{\xi}_i\boldsymbol{\beta}_U)G(\boldsymbol{\eta}_i\boldsymbol{\beta}_W).$$

It is good to note that this product corresponds with the product of F and G in the binary choice model.

[10]It is well known that this model could be derived as a limiting case of our above binary choice model, however the justification for doing so would be tenuous at best here. We therefore propose the following model as a separate, but related, case.

4.5 Overdispersion

One of the key problems associated with using this model is that the data often may not have the properties that are predicted by a Poisson Distribution. An easy way to adapt our model to this is to include a further source of random variation. Typically, the problem is that the variance of the data predicted by the Poisson distribution is smaller than we observe – the data is *overdispersed*.

Guo (1999) follows Gourieroux, Monfort and Trognon (1984a) in using the canonical logarithmic link function $\lambda(x) = e^x$, and then including an error term in the violation process (i.e. replacing $\boldsymbol{\xi}_i\boldsymbol{\beta}_U$ with $\boldsymbol{\xi}_i\boldsymbol{\beta}_U + \epsilon$, where $e^\epsilon \sim \text{Gamma}(1/r, r)$ for some overdispersion parameter r. Alternatively, we could simply multiply the rate λG by $\epsilon \sim \text{Gamma}(1/r, r)$ or assume that the true rate used is sampled from a Gamma distribution with shape parameter r and scale parameter $\lambda G/r$. All of these models lead to the conclusion that the probability function $g(D_i)$ is from a negative binomial distribution with mean λG and 'size' r.

For example, if we model the true rate as being sampled from a Gamma distribution,

$$\begin{aligned}
g(D_i) &= \int_0^\infty \frac{e^{-t}(t)^{D_i}}{D_i!} \times \frac{t^{r-1}e^{-tr/(\lambda G)}}{\Gamma(r)(\lambda G/r)^r} dt \\
&= \frac{1}{D_i!\Gamma(r)} \left(\frac{\lambda G}{r}\right)^{-r} \int_0^\infty t^{D_i+r-1} e^{-t(1+r/(\lambda G))} dt \\
&= \frac{\Gamma(D_i+r)}{D_i!\Gamma(r)} \left(\frac{\lambda G}{r}\right)^{-r} \left(\frac{\lambda G}{\lambda G + r}\right)^{D_i+r} \\
&= \frac{\Gamma(D_i+r)}{D_i!\Gamma(r)} \left(\frac{r}{r+\lambda G}\right)^{r} \left(1 - \frac{r}{r+\lambda G}\right)^{D_i}
\end{aligned}$$

as desired.

We can now incorporate r into our likelihood function, and estimate it in exactly the same way as previously outlined for other parameters.

Chapter 5

Parameter Identifiability

> *«Tous les géomètres seraient donc fins s'ils avaient la vue bonne, car ils ne raisonnent pas faux sur les principes qu'ils connaissent; et les esprits fins seraient géomètres, s'ils pouvaient plier leur vue vers les principes inaccoutumés de géomètrie.»*
>
> **Blaise Pascal**
> ***Pensées: La «Rhétorique» de Pascal (1670)***

As noted earlier, it is not immediately apparent that we will be able to distinguish between the effects of regulation and violation. This problem (of not being able to uniquely determine parameters given any amount of data) is termed non-identifiability.

Following the lead of Koopmans (1949), we shall attempt to further understand this in the following way. We can consider estimation to be performed in two distinct steps: (1) we attempt to infer from our data to the underlying joint distribution which generated it, then (2) we attempt to calculate parameters of our model from the joint distribution. It is this second step which Koopmans refers to as 'identification'[1]. More formally, in this second stage we assume that we know the values of the true probability function, and use it to determine the parameters of the model.

It is this latter step which forms the main study of this chapter, as it is a significant issue within our models. (To see this highlighted, for example, Feinstein's (1990) original paper or Gordon and Smith (2004), where this question is addressed rather through the use of qualitative observations – a technique which is not considered here.)

[1]Koopmans' method is closely related to ergodicity. In stage (1) we assume that our sample-probabilities converge to the true state-probabilities, then in stage (2), we require that these state-probabilities correspond with a unique parameter vector. In practice, stage (1) represents requirements on the data available to us, while (2) represents requirements on the model to be fitted.

In the analysis of this chapter, we assume that any categorical variables have been expressed in a binary form, that is as a number of variables taking only the values zero or one.

5.1 A Formal Definition

We shall first treat the simple binary choice model. We consider a collection of mappings f from a point in the 'data-space' Ξ to the unit interval $(0, 1)$. That is, $f : \Xi \to (0, 1)$. This function gives a probability (of violation, detection or both) $f(\boldsymbol{\xi}, \boldsymbol{\eta})$ for a firm with characteristics $(\boldsymbol{\xi}, \boldsymbol{\eta}) \in \Xi$. We shall term the set of these functions $\mathcal{F}$. Confusingly, each such f *is*, in this context, equivalent to a 'joint distribution' of outcomes as described by Koopmans (1949)[2].

Now we consider a 'parameter-space,' denoted B, coming equipped with a map $m : B \to \mathcal{F}$. This map m is effectively the proposed 'model' for our data, and takes a parameter vector $\boldsymbol{\beta} \in B$ to a function f describing the relationship between a firm's characteristics $(\boldsymbol{\xi}, \boldsymbol{\eta}) \in \Xi$ and the probability of a detected violation.

We assume that there is a (unique) true value $\boldsymbol{\beta}^* \in B$ such that the relationship $f^* = m(\boldsymbol{\beta}^*)$ is what we observe. Our aim is to determine $\boldsymbol{\beta}^*$ from our estimate of f^*. Because of this assumption, without loss of generality we shall assume that $m(B) = \mathcal{F}$, that is, m is a surjection.

The problem of parameter-identifiability can now be formulated rigorously in this context. The parameters of a model are called **point-identifiable** (at f) if the preimage of f contains only a single element. Generalising this statement, we shall call the parameters of a model **globally-identifiable** if the parameters are point-identifiable for every $f \in \mathcal{F}$. (These are my own terms.)

While it is important for the sake of precision, the generality of the above discussion is too great for our purposes, as only certain types of parameter spaces and models are of any practical interest. We shall therefore focus our attention on a very small class of models, namely those formed by the product of two generalised linear functions, where B is an open subset of $\mathbb{R}^k$, m is the obvious map from a parameter vector to the function it generates, etc... A simple but illuminating example of this follows.

[2]This is because there is a bijection between the probability of an event p and the Bernoulli distribution with parameter p. Later, we shall extend this to other distributions, exploiting the fact that there is typically a bijection between a distribution and the parameters which generate it.

5.2 A Small Example

Suppose there are two covariates associated with each firm: ξ and η. Suppose furthermore that a firm violates with probability

$$P(\text{Violates}) = \text{logit}^{-1}(\beta_1 + \beta_2\xi),$$

and is subsequently detected with probability

$$P(\text{Detected}|\text{Violates}) = \text{logit}^{-1}(\beta_3 + \beta_4\eta).$$

Suppose η takes values from $\{0, 1\}$ and ξ takes values from $\{0, 1, 2\}$ (and that any of these values for ξ can be taken when $\eta = 0$). We allow $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4) \in B$ to take any value in $\mathbb{R}^k$. Hence, the functions f are of the form

$$f(\xi, \eta) = \text{logit}^{-1}(\beta_1 + \beta_2\xi)\,\text{logit}^{-1}(\beta_3 + \beta_4\eta)$$

where $\text{logit}^{-1} = (1 + e^{-x})^{-1}$, the standard logistic cumulative distribution function.

Under the above paradigm for determining identification, the question that must be answered is: "Given the known values for f, can the parameters $\boldsymbol{\beta}$ be uniquely determined?"

To answer this we note that as $\text{logit}^{-1}(x) \neq 0$ for any x, we can take a logarithmic transformation – this simply works to make the algebra easier to follow, as if we write $\phi := -\log(1 + e^{-x}) = \log(\text{logit}^{-1}(x))$, then we know

$$\log f_{\boldsymbol{\beta}}(\xi, \eta) = \phi(\beta_1 + \beta_2\xi) + \phi(\beta_3 + \beta_4\eta).$$

In particular, as under Koopmans' approach we can claim to *know* the value of f (and hence of $\log f$) at each of its values, we can claim to know

$$\begin{bmatrix} \log f(0,0) \\ \log f(1,0) \\ \log f(2,0) \\ \cdots \end{bmatrix} = \begin{bmatrix} \phi(\beta_1) + \phi(\beta_3) \\ \phi(\beta_1 + \beta_2) + \phi(\beta_3) \\ \phi(\beta_1 + 2\beta_2) + \phi(\beta_3) \\ \cdots \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ & & \cdots & & \end{bmatrix} \begin{bmatrix} \phi(\beta_0) \\ \phi(\beta_1 + \beta_2) \\ \phi(\beta_1 + 2\beta_2) \\ \phi(\beta_3) \\ \phi(\beta_3 + \beta_4) \end{bmatrix}.$$

Now as we know each of these values, we can take any linear combination of them – in other words we can do row operations on the matrix in the last of these terms. From this, we can determine

$$\phi(\beta_1 + \beta_2) - \phi(\beta_1) = k_1,$$
$$\phi(\beta_1 + 2\beta_2) - \phi(\beta_1) = k_2,$$

for some known values k_1, k_2. As ϕ is invertible, this pair of equations implies:

$$0 = 2\phi^{-1}(k_1 + \phi(\beta_1)) - \phi^{-1}(k_2 + \phi(\beta_1)) - \beta_1 =: \Gamma(\beta_1).$$

If this uniquely defines β_1, then we can easily determine $\beta_2 = \phi^{-1}(\phi(\beta_1) + k_1) - \beta_1$. From here, we can determine β_3, β_4 by simply subtracting $\phi(\beta_1 + \beta_2\xi)$, which is now known, from the known value of $\log f_{\boldsymbol{\beta}}(\xi, \eta)$ and fitting using these points.

Therefore, we claim that:

Proposition 5.2.1. *For this model to be locally identifiable it is sufficient that $\Gamma(\beta_1) = 0$ have a unique solution. Furthermore, if this is the case for all $k_1, k_2 \neq 0$ (with $k_2/k_1 > 1$ – required for a solution to exist), then our model is globally identifiable (apart from the case where $\beta_2 = 0$).*

As here $\phi = -\log(1 + e^{-x})$,

$$\begin{aligned} 0 &= \Gamma(\beta) \\ &= 2\phi^{-1}(k_1 + \phi(\beta_1)) - \phi^{-1}(k_2 + \phi(\beta_1)) - \beta_1 \\ &= -2\log[(1 + e^{-\beta_1})e^{-k_1} - 1] + \log[(1 + e^{-\beta_1})e^{-k_2} - 1] - \beta_1 \\ \Rightarrow e^{\beta_1} &= \frac{(1 + e^{-\beta_1})e^{-k_2} - 1}{[(1 + e^{-\beta_1})e^{-k_1} - 1]^2}. \end{aligned}$$

From here, it is easy to follow through the algebra to find that

$$(1 + e^{-\beta_1})^2 e^{-2k_1} - 2(1 + e^{-\beta_1})e^{-k_1} + 1 = (1 + e^{-\beta_1})e^{-\beta_1 - k_2} - e^{-\beta_1}$$

and thus (provided $k_2/k_1 > 1$)

$$\begin{aligned} (1 + e^{-\beta_1})e^{-2k_1} - 2e^{-k_1} &= e^{-\beta_1 - k2} - 1 \\ e^{-\beta} &= \frac{e^{-2k_1} - 2e^{-2k_1} + 1}{e^{-k_2} - e^{-2k_1}} \end{aligned}$$

therefore

$$\beta_1 = \log\left(\frac{e^{2k_1} - e^{k_2}}{[e^{k1} - 1]^2}\right) - k_2$$

is the unique solution to $\Gamma(\beta_1) = 0$. Therefore, in this situation our model is globally identifiable[3].

While this would seem to solve the problem of identification, even for the relatively straightforward case where our link functions are inverse normal distribution functions, showing that Γ is injective is not a completely trivial problem. However it is easy to plot Γ numerically for any pair k_2, k_1. Therefore, we recommend this plot as a diagnostic when using other link functions in this context – if it is clear that there will be only one solution, then an applied statistician can be confident that the model is well identified.

[3]It is possible to extend this result to when the known points are for any three distinct values of ξ. See Appendix A.2 for details.

5.3 Ein Gedankenexperiment

This above example leads us to the following gedankenexperiment ('thought-experiment'), which will highlight how we can distinguish between violation and detection.

Suppose we have a covariate η which contributes to detection but not to violation. By comparing firms which differ only in η, we can determine the effect that η has, and can attribute that effect to the detection process.

We now consider firms differing in other variables. The marginal impact of these variables on our 'control' firm types can be estimated. We exploit the fact that if the link function (logit^{-1} in the above example) is curved, then the marginal impact of other variables on detection will differ depending where on the curve a firm is, and so we can use this fact to determine a firm's location along the curve.

Provided we have appropriate data and link functions, this should allow us to estimate the parameters of the detection process. Given these, finding the parameters of the violation process is straightforward.

5.4 A General Response

We will formalise this above idea, to allow us to determine if a model is identifiable. To do so, we note the useful representation of the known points in the matrix above. In general we suppose $\phi(\boldsymbol{\beta}_1\boldsymbol{\xi})$ gives the log of the probability of violation for a given $\boldsymbol{\xi}$, and $\psi(\boldsymbol{\beta}_2\boldsymbol{\eta})$ the log of the probability of detection for a given $\boldsymbol{\eta}$. Here the $\boldsymbol{\xi}, \boldsymbol{\eta}$ are vectors of covariates (which may share components), and the $\boldsymbol{\beta}_i$ are real coefficient vectors such that $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \in B$ an open subset of $\mathbb{R}^k$.

We can then express our knowledge, for each possible point $(\boldsymbol{\xi}_i, \boldsymbol{\eta}_i)$, in the form

$$\begin{bmatrix} \log f(\boldsymbol{\xi}_1, \boldsymbol{\eta}_1) \\ \log f(\boldsymbol{\xi}_2, \boldsymbol{\eta}_2) \\ \vdots \end{bmatrix} = M \begin{bmatrix} \phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_1) \\ \phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_2) \\ \vdots \\ \psi(\boldsymbol{\beta}_1\boldsymbol{\eta}_1) \\ \psi(\boldsymbol{\beta}_1\boldsymbol{\eta}_2) \\ \vdots \end{bmatrix}$$

for some matrix M of 0's and 1's.

In general, M will not be left-invertible[4], as M will generally have more rows than

[4]By M being left-invertible, we mean that there exists a matrix N for which $NM = I$.

columns, and the columns will be linearly dependent[5]. We will therefore consider working on determining only one of our processes, and then using these results to determine the other. (In the following discussion, without loss of generality we look at determining the process denoted by ϕ.) We therefore consider performing row operations on M (and possibly column operations, exchanging the variables in the final vector), to obtain a matrix of the form

$$\left[\begin{array}{ccccc|c} \mathbf{-1} & \mathbf{I}_{l_1} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{-1} & \mathbf{I}_{l_2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \hline & & & \vdots & & \end{array}\right].$$

Here $l_j + 1$ is, in some sense, the number of values $\boldsymbol{\xi}$ can take for a given $\boldsymbol{\eta}_j$, and so $\sum_j (l_j + 1)$ is the number of covariate values determining the ϕ process. (For an example of such a matrix, see the example of Section 5.6.) This matrix has sufficient zeros down the right hand side to eliminate the ψ process from our equations, and allowing us to focus our attention on estimation of the ϕ process only.

From this, it is clear that we can obtain an expression for every $\phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_i)$ as the sum of $\phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_j)$ and a known value (denoted k_i). If ϕ is invertible[6], then we can obtain the equation

$$\boldsymbol{\beta}_1\boldsymbol{\xi}_i = \phi^{-1}(\phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_j) + k_i).$$

Expressing this in terms of a matrix of $\boldsymbol{\xi}$ values (and omitting the trivial value $\boldsymbol{\xi}_j$),

$$\Xi\boldsymbol{\beta}_1 = \begin{bmatrix} \boldsymbol{\xi}_2^T \\ \boldsymbol{\xi}_3^T \\ \vdots \end{bmatrix} \boldsymbol{\beta}_1 = \begin{bmatrix} \phi^{-1}(\phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_j) + k_2) \\ \phi^{-1}(\phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_j) + k_3) \\ \vdots \end{bmatrix}.$$

Now provided we have points which are not perfectly multicollinear, and at least one variable which can take on at least three possible values[7] which was not part of $\boldsymbol{\eta}$, then Ξ is left-invertible. Hence we know

$$\boldsymbol{\beta}_1 = (\Xi^T\Xi)^{-1}\Xi^T \begin{bmatrix} \phi^{-1}(\phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_j) + k_1) \\ \phi^{-1}(\phi(\boldsymbol{\beta}_1\boldsymbol{\xi}_j) + k_2) \\ \vdots \end{bmatrix}.$$

[5] An exception to this is if we allow one of our covariates to become unbounded, as then we may have a point (or a limiting sequence of points) such that one of ϕ and $\psi \to 0$ for any β. In this case we may be able to exclude this value from our analysis, leaving M to be of full column rank. In such a case, the question of identification becomes vastly simpler, as we can take a left-inverse of M and obtain values of ϕ, ψ directly. However this assumption is often not a necessary one, as our models are typically identifiable without such a requirement, and obtaining unbounded data is difficult in practice.

[6] If ϕ is not invertible, it may still be possible to consider the preimage of each point and draw some conclusions. This however is not a case we will discuss here.

[7] More generally a combination of a variables not in $\boldsymbol{\eta}$ which can collectively take on $a + 2$ values is often enough, however problems arise if data is not available for every combination of covariates. See Appendix A.3 for a discussion of this.

Now, we need to show that this equation has only one solution. If we scale our data such that $\boldsymbol{\xi}_j = (1, 0, 0, ..., 0)$ we can write the equation

$$\boldsymbol{\beta}_1 = (\Xi^T\Xi)^{-1}\Xi^T \begin{bmatrix} \phi^{-1}(\phi(\beta_{11}) + k_1) \\ \phi^{-1}(\phi(\beta_{11}) + k_2) \\ \vdots \end{bmatrix}.$$

If we consider the first line of these vectors, we obtain a simple enough equation in one variable. We simply wish to show that this has a unique solution β_{11} in B. Subtracting β_{11} from both sides, we obtain (a scalar multiple of) the function denoted Γ in the above example. Plotting this function on an arbitrary domain is easy enough (for 'nice' functions ϕ) and it is often quite clear by inspection that only one solution exists. We can do this for each j to obtain a collection of functions Γ_j, each of which must have a unique root.

Once we have a handle on our solution for the ϕ process, we can simply write

$$\log f(\boldsymbol{\xi}, \boldsymbol{\eta}) - \phi(\boldsymbol{\beta}_1\boldsymbol{\xi}) = \psi(\boldsymbol{\beta}_2\boldsymbol{\eta}).$$

We then proceed using these differenced data to ensure that we get identifiability. Provided ψ is invertible, it is now sufficient that the data making up $\boldsymbol{\eta}$ are not perfectly multicollinear.

Note: We do not here require ψ to satisfy all the above conditions on ϕ – this will be important when considering a Poisson process model, as the canonical link function in this case is exponential, which fails the requirements above.

5.4.1 Extending to Non-Binary Models

When dealing with a Poisson or Negative Binomial[8] model, we no longer wish to consider $f(\boldsymbol{\xi}, \boldsymbol{\eta})$ as the probability of an event, but now rather as the rate at which events occur. Of course f no longer gives values in $(0, 1)$ but now in $(0, \infty)$. However, as there is no part of this analysis which depends on the requirement that our functions be bounded below one, the above carries directly over to this new setting.

5.4.2 A Geometric Observation

It is interesting to note that this problem boils down to a problem which is fairly straightforward to state in terms of geometry. We consider the curve defined by the points

[8]We will not prove that the overdispersion parameter r is well identified, as there is no difference between a DCE model and a standard Generalised Linear Model in this regard.

$(x, \phi(x) : x \in \mathbb{R})$. For this curve, is there a unique set of points such that the 'vertical' differences between the points are the values $(k_1, k_2, ...)$ and the 'horizontal' differences between the points are in a given ratio, determined by the matrix Ξ? A diagram of this is given in Figure 5.1. In other words, we ask the question, "Is it possible through translation and horizontal stretching to place the rectangle shown on the curve in more than one way?". If not, then our model is identifiable.

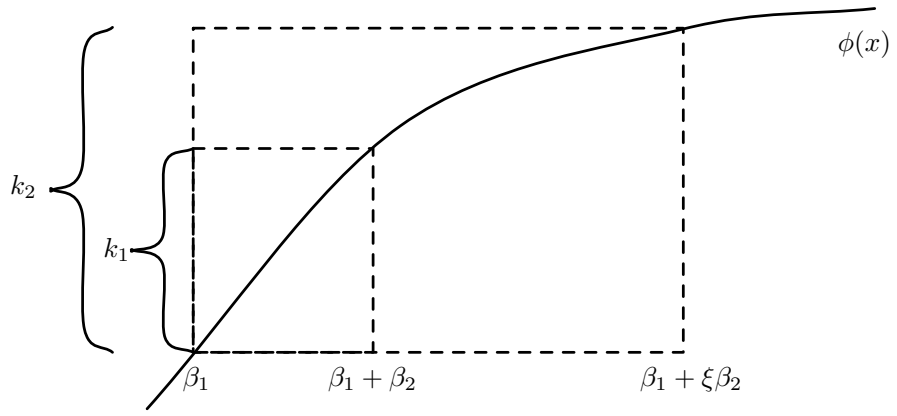

Figure 5.1: The geometric problem of identification

A solution to this geometric problem (that is, a list of the classes of curves ϕ for which a unique solution exists) would immediately determine the class of functions for which our parameters will be most often identifiable.

5.5 Examples of failure

There are a few simple examples where the identification fails. Feinstein (1990) claims that of those functions representing the distribution functions of typical parametric families, only the exponential case fails. In particular he states the following theorem:

Theorem 5.5.1 (Feinstein's Theorem A1 (pp271))**.** *Assume that $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ contain only continuous components (apart from intercepts) and that they have some elements in common. If identification fails, the link functions must each belong to the exponential family.*

This claim is, in the absence of modification, incorrect[9]. We here present a few counterexamples, including one where all the assumptions present in the theorem are satisfied, but identification fails (for a non-exponential case) even up to a constant and

[9]See Appendix A.1 for a discussion of this theorem and the appropriate modifications which need to be made.

scalar multiple. Feinstein also fails (in our opinion) to appropriately outline the requirements on the data to be used in the parametric context, instead immediately moves to semi-parametric estimation.

We need to distinguish between those cases where we can show that parameter identifiability will fail and those where this analysis simply cannot determine that parameter identifiability holds. In general this section will illustrate the former, for this reason the fact that they are not identifiable will be quite clear compared with the application of the above analysis.

5.5.1 Uniform Distributions

Suppose we use a linear function as our link function, or even one which is linear on an adequately large subsection (e.g. the uniform distribution distribution function). In particular, consider the simple example where ξ and η take values from $\{-1, 0, 1\}$ 'independently'[10] of each other. We now write the probability as

$$f(\xi, \eta) = (\beta_1 + \beta_2\xi)(\beta_3 + \beta_4\eta).$$

Suppose now that $0 < |\beta_2| < \beta_1$ and $0 < |\beta_4| < \beta_3$. Then let $c^* = \max\{\beta_1 \pm \beta_2, \beta_3 \pm \beta_4\}$. For any $c \leq c^*$, we can write

$$f(\xi, \eta) = (c\beta_1 + c\beta_2\xi)\left(\frac{\beta_3}{c} + \frac{\beta_4}{c}\eta\right)$$

and this function will represent identical probabilities. Hence this model is not identified in this case.

From the perspective of our analysis, we note that on the relevant range, $\phi(x) = \log(x)$, and hence if we let $\boldsymbol{\xi} = (1, 0)$ we have:

$$B = \{\beta_1, \beta_2, \beta_3, \beta_4 : 0 < |\beta_2| < \beta_1, 0 < |\beta_4| < \beta_3\}$$

and also,

$$\phi^{-1}(\phi(\beta_1) + k_i) = e^{\log(\beta_1)+k_i} = \beta_1 e^{k_i},$$
$$\Xi = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \Rightarrow \Xi^{-1} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

and similarly if we consider ψ.

Then we obtain the equation

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} \beta_1 e^{k_2} \\ \beta_1 e^{k_3} \end{bmatrix}.$$

[10] In the sense of any combination of them being possible.

Hence we wish to show that there is a unique solution to

$$\Gamma(\beta_1) = \frac{\beta_1}{2}[e^{k_1} + e^{k_2}] - \beta_1 = 0.$$

This is clearly not the case, as we have assumed that $\beta_1 > 0$ and so either no β_1, or any β_1, will satisfy $\Gamma(\beta_1) = 0$.

5.5.2 Exponential Functions

Suppose that our probability (or our rate) is of the form

$$f(\xi, \eta) = e^{\beta_1 + \beta_2 \xi} e^{\beta_3 + \beta_4 \eta}$$

and $B = \mathbb{R}^4$.

It is clear that we can rewrite this as

$$f(\xi, \eta) = e^{\beta_1 + \beta_3 + \beta_2 \xi + \beta_4 \eta}$$

and so for any $c \in \mathbb{R}$

$$f(\xi, \eta) = e^{\beta_1 + c + \beta_2 \xi} e^{\beta_3 - c + \beta_4 \eta}$$

and therefore our parameters are not identifiable[11].

In terms of our model, $\phi(x) = x$ and so

$$\phi^{-1}(\phi(\beta_1) + k_i) = \beta_1 + k_i$$

and similarly for ψ. Using the same data points as above we have:

$$\Xi^{-1} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$$

and so wish to show that there is a unique solution in β_1 to:

$$0 = \frac{1}{2}(\beta_1 + k_2 + \beta_1 + k_3) - \beta_1 = \frac{k_2 + k_3}{2}$$

which is obviously not the case.

[11] This case points out an interesting problem of determining what types of models have only some parameters identified, a subtlety which we will not explore here. Feinstein's (1990) paper discusses the question of identification up to a constant and scalar multiple at length (particularly in the context of semiparametric estimation), however makes further assumptions as to the nature of the data available.

5.5.3 Insufficient Points

Consider the case when

$$f(\xi, \eta) = \text{logit}^{-1}(\beta_1 + \beta_2\xi)\,\text{logit}^{-1}(\beta_3 + \beta_4\eta)$$

but ξ and η can only take the values $\{0, 1\}$ while $B = \mathbb{R}^4$. It is clear that we can choose $\beta_1, \beta_2, \beta_3, \beta_4 \in \mathbb{R}$ such that the four points

$$\text{logit}^{-1}(\beta_1), \quad \text{logit}^{-1}(\beta_1 + \beta_2), \quad \text{logit}^{-1}(\beta_3), \quad \text{logit}^{-1}(\beta_3 + \beta_4)$$

will take on any specified values in $(0, 1)$. Taking a log transform, we claim to know

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \log(\text{logit}^{-1}(\beta_1)) \\ \log(\text{logit}^{-1}(\beta_1 + \beta_2)) \\ \log(\text{logit}^{-1}(\beta_3)) \\ \log(\text{logit}^{-1}(\beta_3 + \beta_4)) \end{bmatrix}.$$

As the matrix here is clearly not invertible, and we have four free variables, our solution space must be of dimension >0. Hence, it is impossible to identify the values of $\text{logit}^{-1}(\beta_1), ...,$ and hence it must be impossible to identify the values of $\beta_1, ..., \beta_4$.

In terms of our above approach, this corresponds to

$$\Xi = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

which is clearly not left-invertible (it has non-full rank). A similar result will occur if our data is perfectly multicollinear.

5.5.4 Insufficient Differences in data

Consider the case when

$$f(\xi) = \text{logit}^{-1}(\beta_1 + \beta_2\xi)\,\text{logit}^{-1}(\beta_3 + \beta_4\xi)$$

with $B = \mathbb{R}^4$. Then it is clear that we could simply swap the two processes over

$$f(\xi) = \text{logit}^{-1}(\beta_3 + \beta_4\xi)\,\text{logit}^{-1}(\beta_1 + \beta_2\xi)$$

and so we can never distinguish between the β_1 and β_3 values and the β_2 and β_4 values. Except in the special case where $\beta_1 = \beta_3, \beta_2 = \beta_4$, i.e. when our detection and violation processes are identical, this implies that our parameters are not identifiable.

In the context of the above model, this is because when we write out our M matrix, we find that it is of the form $[A|A]$ for some matrix A. This implies that no number of row operations will be sufficient to separate out the two processes, therefore that our above construction is impossible.

5.5.5 A Peculiar Counterexample

We here present an example which does not allow for the parameters to be identified even up to a constant and scalar multiple, when the link functions involved are not exponential, and all the assumptions of Feinstein's theorem (without any further strengthening) are satisfied. This example is clearly pathological, however it outlines the problems that can be encountered – particularly if we allow our link functions to vary.

Let our rate of detected violation be of the form

$$f(\xi, \eta) = F(\beta_1 + \beta_2\xi + \beta_3\eta)G(\beta_4 + \beta_5\xi)$$

where $F(x) = e^x$ and $G(x) = e^{x+\sin(\log x)}$ (which are not both exponential), $\xi \in (0, 1)$ and $\eta \in \mathbb{R}$ (which are both continuous) and

$$B = \{(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5) \in \mathbb{R}^5 : \beta_4 > 0.2, |\beta_5| < \beta_4\}.$$

(Note, for $x > 0.175$ approx., $G(x)$ is invertible.)

In this case, we can rewrite f in the form

$$f(\xi, \eta) = \exp\{(\beta_1 + \beta_4) + (\beta_2 + \beta_5)\xi + \beta_3\eta + \sin(\log(\beta_4 + \beta_5\xi))\}$$

and so if $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ is a solution, then so is

$$\left(\beta_1 + (1 - e^{2n\pi})\beta_4, \beta_2 + (1 - e^{2n\pi})\beta_5, \beta_3, e^{2n\pi}\beta_4, e^{2n\pi}\beta_5\right),$$

for any $n \in \mathbb{Z}$ (and there are possibly other solutions as well). It is clear that the constants are not identified here, but also that there is a rather peculiar relationship between the values of β_2 and β_5 (in particular, they are not even identified up to scalar multiples).

To apply the above analysis to this situation is difficult, as even though $G(x)$ is invertible, its inverse has no simple form. However it can be inverted numerically[12], and a plot of Γ is shown in Figure 5.2.

It is immediately clear that Γ has more than one root, and also that each of these roots is isolated from the others. Hence our model is not well identified, even if we have an invertible information matrix at this point.

5.6 And An Example of Success

To present a nontrivial example where identification does not fail, let us consider the case where violations occur at a rate of $e^{\beta_4+\beta_5\xi_1+\beta_6\xi_3}$, and these violations are detected with

[12]The code to do this is in Appendix C.2.1, and is easily modifiable for any other invertible link functions.

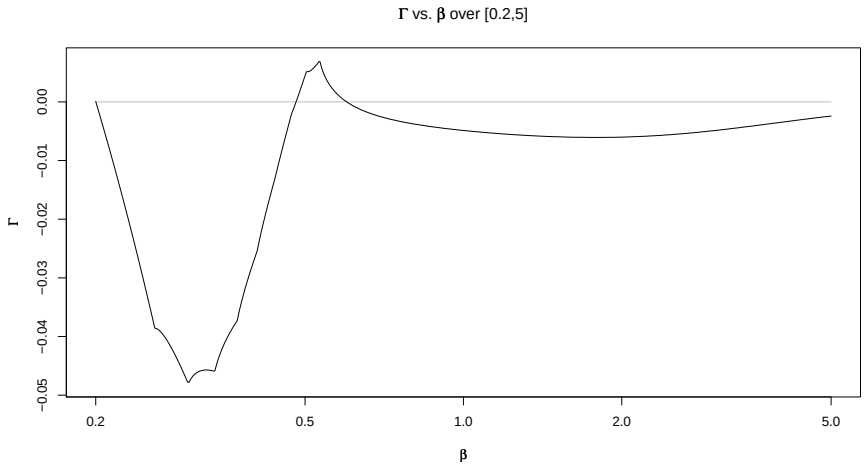

Figure 5.2: An approximate graph of Γ vs β_4 (Horizontal axis is log-scale)

probability $\Phi(\beta_1+\beta_2\xi_1+\beta_3\xi_2)$. Here Φ is the normal distribution function, $\xi_2 \in \{0,1,2\}$ and $\xi_1, \xi_3 \in \{0,1\}$ independently, and $B = \mathbb{R}^6$. Hence the rate of detected violations is given by

$$f(\xi_1,\xi_2,\xi_3) = \Phi(\beta_1+\beta_2\xi_1+\beta_3\xi_2)e^{\beta_4+\beta_5\xi_1+\beta_6\xi_3}.$$

Of course, if our data can take more values than this, that will not cause any problems – we can simply ignore them, as these points are sufficient.

With $\phi(x) := \log\Phi(x)$, we know the points:

$$\begin{bmatrix} 1&0&0&0&0&0&1&0&0&0\\ 1&0&0&0&0&0&0&0&1&0\\ 0&1&0&0&0&0&0&1&0&0\\ 0&1&0&0&0&0&0&0&0&1\\ 0&0&1&0&0&0&1&0&0&0\\ 0&0&1&0&0&0&0&0&1&0\\ 0&0&0&1&0&0&0&1&0&0\\ 0&0&0&1&0&0&0&0&0&1\\ 0&0&0&0&1&0&1&0&0&0\\ 0&0&0&0&1&0&0&0&1&0\\ 0&0&0&0&0&1&0&1&0&0\\ 0&0&0&0&0&1&0&0&0&1 \end{bmatrix} \begin{bmatrix} \phi(\beta_1)\\ \phi(\beta_1+\beta_2)\\ \phi(\beta_1+\beta_3)\\ \phi(\beta_1+\beta_2+\beta_3)\\ \phi(\beta_1+2\beta_3)\\ \phi(\beta_1+\beta_2+2\beta_3)\\ \beta_4\\ \beta_4+\beta_5\\ \beta_4+\beta_6\\ \beta_4+\beta_5+\beta_6 \end{bmatrix}$$

Performing row & column operations we get:

$$\begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ & & & & \vdots & & & & & \end{bmatrix} \begin{bmatrix} \phi(\beta_1) \\ \phi(\beta_1+\beta_3) \\ \phi(\beta_1+2\beta_3) \\ \phi(\beta_1+\beta_2) \\ \phi(\beta_1+\beta_2+\beta_3) \\ \phi(\beta_1+\beta_2+2\beta_3) \\ \beta_4 \\ \beta_4+\beta_5 \\ \beta_4+\beta_6 \\ \beta_4+\beta_5+\beta_6 \end{bmatrix}$$

Now suppose we estimate this model using some method, and obtain estimates $\beta_1 = 1$, $\beta_2 = 2$ and $\beta_3 = 1$. Hence we have implicitly estimated the values

$$\begin{aligned} \phi(\beta_1+\beta_3) - \phi(\beta_1) &= 0.1497409, \\ \phi(\beta_1+2\beta_3) - \phi(\beta_1) &= 0.1714030, \\ \phi(\beta_1+\beta_2+\beta_3) - \phi(\beta_1+\beta_2) &= 0.001319138, \\ \phi(\beta_1+\beta_2+2\beta_3) - \phi(\beta_1+\beta_2) &= 0.001350523. \end{aligned}$$

And so we must consider the two functions

$$\begin{aligned} \Gamma_1(\beta_1) &= 2\phi^{-1}(\phi(\beta_1)+0.1497409) - \phi^{-1}(\phi(\beta_1)+0.1714030) - \beta_1, \\ \Gamma_2(\beta_1^*) &= 2\phi^{-1}(\phi(\beta_1^*)+0.001319138) - \phi^{-1}(\phi(\beta_1^*)+0.001350523) - \beta_1^*, \end{aligned}$$

where $\beta_1^* := \beta_1 + \beta_2$. Proving that these have unique roots is difficult analytically, however plotting them is fairly straightforward[13], as in Figure 5.3.

It is clear that both of these functions will have a unique root, and therefore that the estimated values of β_1 and $\beta_1 + \beta_2$ are the only ones which gives this relationship. From this it follows that the estimated value of β_3 is unique, and hence that the estimated values of β_4, β_5 and β_6 are also unique.

In other words, this relationship is (point) identified.

5.7 Some Heuristics

In practice, it would be useful to know *a priori* which models are at least likely to be identifiable, and so the following rules of thumb are useful. In general, if these assumptions are satisfied, identification will follow.

[13] See Appendix C.2.2 for the code used.

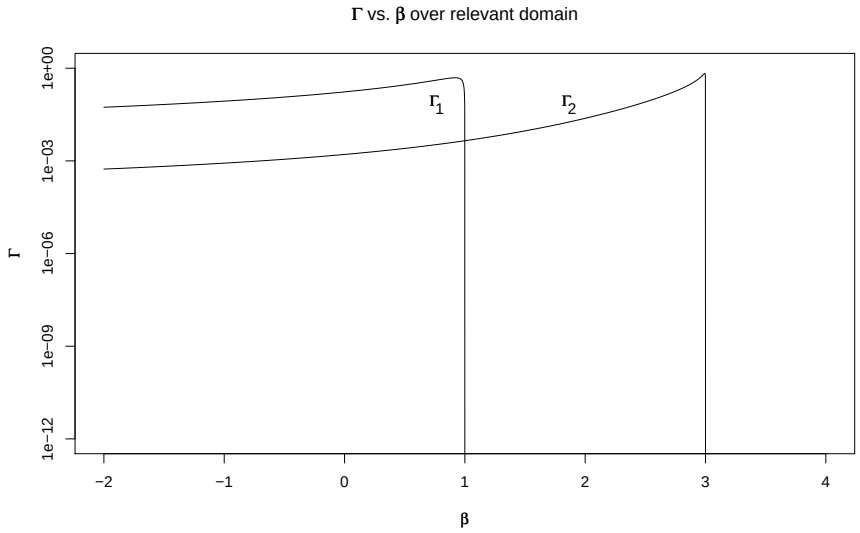

Figure 5.3: A graph of Γ_1, Γ_2 vs β_1, β_1^*. For this example, a logarithmic scale is used on the vertical axis to highlight the positive nature of the graph.

1. We shall assume that at least one of our processes has a 'nice' link function – for example those similar to logistic or probit regression work well, exponential and linear functions do not work well.

2. We wish this process to be dependent on data which contains at least one covariate which takes at least 3 distinct values (with non-zero probability). If higher-order terms/piecewise linear terms in this covariate are also included in the model, then we will require more than 3 points. Overall, we require one more point than we would for 'classical' regression[14].

3. The coefficient of this covariate must be non-zero.

4. This covariate must be independent of the other process.

5. We also require our covariates not to be perfectly multicollinear.

5.8 Semi-parametric estimation

We have looked here at using parametric methods for estimating the proposed relationship between detection, violation and observed covariates. Often these are quite sufficient, as the distinction between parametric models is limited, and often our data is quite discrete. On the other hand, we do not want to simply have our models being

[14]As noted earlier, if no such covariate is available, we can use generally a covariates which can take on $a + 2$ values, but this often causes our *data* to be insufficient for estimation.

estimable purely because of our parametric assumptions. For an exploration of semiparametric regression methods that could very well be used here, see Horowitz and Hardle (1996), or for an alternate approach, see Carroll et al. (1997). As mentioned earlier, Feinstein (1990) discusses the question of parameter identifiability in this context at length, albeit his results assume that continuous covariates are available.

5.9 Data Requirements

In this chapter we have focused on the question of 'identifiability' – the second stage of Koopmans' approach. Given this, it is appropriate to make a few comments about the requirements on the data available for this to work.

Basically, the requirements on the data will look quite similar to the requirements for identifiability above, where Ξ is replaced by the available covariate matrix. In the optimal setting, this means that we would like at least one observation with each combination of covariates. However, we also know that if our model is not well identified, then (by the rank theorem of differential geometry) there will often exist a lower degree manifold of parameters which explain the data equally well. In this case, we will essentially have a 'ridge' in our likelihood function, and moving along such a ridge we will have zero derivative. Conversely, the existence of such a ridge indicates that our data either (1) fails to uniquely identify the joint distribution of the outcomes (a data-based problem) or (2) that our model is unidentifiable and will always have a continuum of equally plausible estimates (a model-based problem).

Such a ridge can be found empirically by looking at the eigenvalues of the observed information matrix $\mathcal{I}$. A negative eigenvalue indicates that we have not found the maximum likelihood estimates, a zero eigenvalue (possibly up to numerical error) indicates a 'ridge', with the associated eigenspace giving the direction of the ridge from our current estimates. This acts as another useful tool in determining if our parameter estimates are unique, particularly for a model which is identifiable in theory, but for which we are unsure if we have sufficient data.

Given sufficient data, any identifiable model should have unique parameter estimates, however such data may not be available, particularly not in sufficient quantities to give accurate, stable estimates for our models.

Chapter 6

Data and Analysis

«J'avais entres tes mains déposé la justice,
«De peur que l'homme n'erre et ne se pervertisse
«Comme au temps de Japhet,
«Des âmes des vivants j'avais fait ton domaine,
«Je t'avais confié la conscience humaine.
«Réponds, qu'en as-tu fait? »

Victor Hugo
La Vision de Dante* from *Choix de Poèmes (1853)

The U.S. Department of Labor Occupational Safety and Health Administration (hereafter OSHA) publishes the results of all inspections freely on their website (OSHA, 2007). Similar data for Australia not being available, we shall investigate the relationships between violation, detection and a firm's profile within the Paper and Allied Products sector in the U.S. The dataset used consists of the 6673 finalised inspections performed on industries in the SIC Major Group 26 classification (Paper And Allied Products), for which the inspection was started between 1 January 1995 and 31 December 2005. Of these, 6132 inspections were actually carried out, and so this data set was used for consideration (541 cases did not result in an inspection, the reasons for this included the business no longer being active, entry being denied, the business having 10 or fewer employees, etc...). The details of how these inspections were carried out are outlined in the document OSHA (1993).

This industry classification (SIC Major Group 26) includes a variety of distinct types of factories. These sub-categories can be seen in Table 6.1.

Other details were available for the firms inspected, namely the street address of the site inspected and whether the workers were members of a union. Also given was

SIC	Description
2611	Pulp Mills
2621	Paper Mills
2631	Paperboard Mills
2652	Setup Paperboard Boxes
2653	Corrugated and Solid Fiber Boxes
2655	Fiber Cans, Tubes, Drums, and Similar Products
2656	Sanitary Food Containers, Except Folding
2657	Folding Paperboard Boxes, Including Sanitary
2671	Packaging Paper and Plastics Film, Coated and Laminated
2672	Coated and Laminated Paper, Not Elsewhere Classified
2673	Plastics, Foil, and Coated Paper Bags
2674	Uncoated Paper and Multiwall Bags
2675	Die-Cut Paper and Paperboard and Cardboard
2676	Sanitary Paper Products
2677	Envelopes
2678	Stationery, Tablets, and Related Products
2679	Converted Paper and Paperboard Products, Not Elsewhere Classified

Table 6.1: SIC Codes for Industries within the '26' Major Classification

- the open and close dates of the case,
- whether the scope of the inspection was partial or complete,
- whether the inspection was focused on safety or health issues,
- whether the inspection was planned, based on an accident, based on a referral or various other categories (and the planning guide used if appropriate),
- any particular emphasis to the inspection,
- whether advanced notice was given and
- whether the corporation was public or privately owned.

In terms of outcomes, most details were given, in particular the initial, current and final numbers of violations within different categories (serious, repeat, willful, other and unclassified) and the fines given within each category. If the inspection was due to an accident, further details were also given.

Unfortunately, this does not include all significant details of the firms in question. The empirical work of Gray and Shadbegian (2005) indicates that some of the most crucial details in predicting the rate at which a firm violates air pollution regulations are its size, age and technology, and it is reasonable to assume that these would also affect its occupational hazard levels. We do not have these details – in fact Gray and

Shadbegian (2005) were only able to access them using confidential survey data from the U.S. Census Bureau.

Nevertheless, we will use this data to investigate whether different firm characteristics are significant in determining the effectiveness of enforcement and violation. Of course, our results are subject to possibly significant omitted variable bias, which must be taken into account when drawing conclusions.

6.1 A Model of Violations

To prevent problems of 'Reverse Causation', we look at the number of serious violations cited in random ('Planned') inspections (as opposed to investigations arising from complaints or accidents), where the inspection was actually carried out. From our data set, we have 1963 such inspections, of which 612 had no serious violations cited, the mean number of serious violations cited was 2.75 and the standard deviation was 3.87. A histogram of this data can be seen in Figure 6.1.

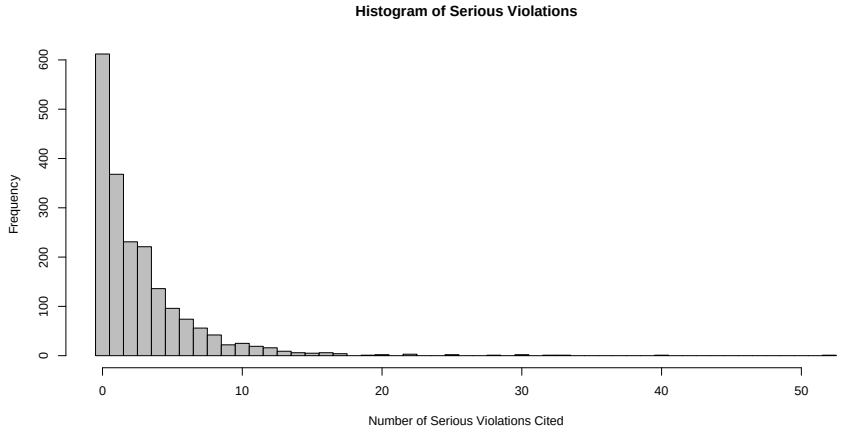

Figure 6.1: Numbers of Serious Violations cited in inspections 1995-2005

As an initial model, we used

$$D \sim \text{NBinom}(d, \lambda G),$$

where D is the number of violations, d is an overdispersion parameter, λ is the rate of violation and G is the probability of a violation being detected. (This uses the 'n, μ' parameterisation of the negative binomial distribution, that is, λG is the expected number of detected violations.)

The functional forms assumed for λ and G were

$$\log \lambda = \boldsymbol{\xi}\boldsymbol{\beta}_1,$$
$$\operatorname{logit} G = \boldsymbol{\eta}\boldsymbol{\beta}_2,$$

where $\boldsymbol{\xi}$ contains (indicator variables for)

- x_1: Industry SIC Code
- x_2: Inspection Scope (Complete/Partial)
- x_3: Inspection Focus (Safety/Health)
- x_4: Unionisation of Workplace
- x_5: Whether advanced warning was given

and $\boldsymbol{\eta}$ contains x_1, x_2,..., x_5 and

- x_6: the length of time the case took to resolve.

In other words, we assume that violation depends on various factors, including the inspection type and whether advanced warning was given (as these may be correlated with underlying risk factors, and may indicate whether a firm is 'cleaning up its act' when it knows an inspection is coming). On the other hand we assume that detection of violations does depend on whether advanced warning was given, and depends linearly on the duration of the case[1].

As the duration of the case can take more than 3 distinct values and we are including it in a process which depends on the logistic link function, by the analysis of Chapter 5, this model will hopefully be identifiable (we still do not know that its coefficient will be nonzero).

This model was fitted in `R`, using a combination of the `nlm` and `optim` commands. Code for this is given in Appendix C.1. These also allow a numerical Hessian matrix for our likelihood function, i.e. the negative of the observed information matrix, to be given directly. Estimated coefficients and their estimated standard errors (in parentheses) are given in Table 6.2.

The overall log-likelihood of this data under this model was -3974.437. Looking at Table 6.2, it is clear that many of the estimated parameters are insignificant, and also

[1] We also investigate the effects of changing this assumption from linearity to something else, however the fundamental question of causation remains. Further discussion of this can be found in Section 6.2 below.

Covariate	Violation		Detection	
(Intercept)	0.40386	(0.29801)	-1.58110	(1.00978)
x_1=2621	0.80518	(0.30371)	0.41176	(1.00488)
x_1=2631	0.45240	(0.33300)	1.26932	(1.06819)
x_1=2652	0.32519	(0.39086)	2.43362	(1.20760)
x_1=2653	0.31846	(0.30414)	1.80888	(1.00846)
x_1=2655	-0.00258	(0.38030)	2.05759	(1.16534)
x_1=2656	0.02209	(0.52810)	1.55868	(1.38559)
x_1=2657	0.36380	(0.32430)	2.22429	(1.06458)
x_1=2671	0.55970	(0.35400)	1.73142	(1.08820)
x_1=2672	0.41755	(0.33882)	1.10096	(1.05849)
x_1=2673	0.19483	(0.33148)	2.45957	(1.10553)
x_1=2674	0.13923	(0.37332)	2.22313	(1.14232)
x_1=2675	0.37309	(0.35929)	1.57518	(1.07942)
x_1=2676	0.55338	(0.38821)	1.11874	(1.15924)
x_1=2677	0.47976	(0.34586)	1.92858	(1.08364)
x_1=2678	0.56911	(0.44693)	0.40959	(1.17624)
x_1=2679	0.40428	(0.31210)	1.88396	(1.02590)
x_2=Partial	-0.69602	(0.14469)	-0.48583	(0.31465)
x_3=Safety	0.78895	(0.08585)	-0.79726	(0.29320)
x_4=Unionised	0.23080	(0.08125)	-0.27975	(0.18898)
x_5=Notice Given	-0.85195	(0.46181)	1.07665	(1.22735)
x_6			1.29194	(0.27152)
d	1.228149	(0.06799)		

Table 6.2: Estimated Coefficients and Standard Errors for Full Model

that many of the SIC categories are not significantly different from each other. To this end, a model was fitted using only the first 3 digits of the SIC code, giving a slightly more generic grouping of industries. The results of doing so are in Appendix B.1. This model had a total of 21 parameters to be estimated, and gave a log-likelihood of -3988.568. Comparing this with the 45 parameters above, and using a Likelihood ratio test (and the asymptotic χ^2_{45-21} approximation) this indicated that there is no evidence of distinction within the SIC subgroupings (P-value of 0.25).

Some qualitative results that can be seen from this analysis are that pulp mills appear to have lower levels of serious violations than worksites further down the production line, and also that pulp, paper and cardboard mills appear to have lower levels of violations detected. Also, it appears that firms undergoing inspections with a 'Safety' emphasis violate at a considerably higher rate (and are detected less) than those selected for 'Health' inspections. A Unionised workforce also corresponds to higher levels of violation and lower levels of detection. A possible explanation for these results is that in these workplaces, violations are more likely to be detected through accidents and complaints, rather than through the results of random inspections.

Advanced notice of an inspection appears to be associated with lower levels of violations (but possibly with higher rates of detection), it is unclear from this data if this is due to a selection bias in which firms are given notice or due to firms cleaning up possible violations prior to the inspection taking place[2]. A partial inspection is associated with considerably lower levels of violations, probably indicating that it is firms which behave well which are selected for partial inspections. Nevertheless, there is weak evidence that partial inspections are less effective at detecting violations than complete inspections, as we would intuitively expect.

Also, the estimated dispersion parameter $d = 1.228149$ is quite small. This indicates that the occurrence and detection of violations does not follow a Poisson process, and indicates that a geometric distribution would possibly be better for modelling the number of violations detected in a single inspection – supporting the model in Section 3.4.

Our key variable x_6 has a strongly nonzero coefficient, and therefore the model is well identified. However one matter of concern is that our estimates may depend too much on the proposed functional form. In particular we would expect that the difference between an inspection lasting zero days and one day would be far more significant than between an inspection lasting 100 and 101 days. To model this, we can apply a nonlinear transformation to the durations and use these transformed values to fit the model. Using the transformation given by $\log(1 + x)$, we fit this model using the grouped SIC data. The results of this can be seen in Appendix B.2. The qualitative results commented on above are still clear under this analysis.

[2]Care must be taken with these estimates, as only 17 planned inspections with advanced notice are actually recorded in this dataset.

The estimated log-likelihood of these results is considerably higher, at -3976.193 (vs. -3988.568 for the untransformed data), indicating a better fit overall.

Some interactive effects were also fitted to this model, however the estimates were all insignificant. (See Appendix B.3 for details.)

6.2 Modelling Inspector Suspicion

In the analysis of our data, we use the duration of an inspection as a covariate in our model. There are clear problems in doing this, as a long duration is possibly related to significant numbers of violations being detected early on. Another problem is that the duration of the case is not equivalent to the time spent on the inspection, but rather involves administrative time and the possibility of appeals by the firm. We might then decide that it is inappropriate to include duration as a covariate, however dropping it results in our model ceasing to be identifiable.

We therefore propose the following adjustment to our model. An inspector, when reaching a worksite, forms an opinion about the safety at the worksite, which will dictate how much effort he will put into the case. This decision is made before any violations are observed. We call this variable 'suspicion', and denote it by S. This suspicion is clearly not observable, however we shall assume that it is linearly related to the true expected number of violations λ and the 'modified duration[3]' of the case $\log(1 + x_6)$. Mathematically

$$S = \gamma_1 + \gamma_2\lambda + \gamma_3 \log(1 + x_6) + \epsilon.$$

For simplicity, we shall assume that e^ϵ follows a Gamma distribution, as it can then be incorporated into the scale factor in our negative binomial model. There are still some questions regarding causation in this model, however we hope that these will not impact too significantly on our results.

We clearly cannot observe S, however we can observe $\log(1 + x_6)$ and estimate λ. Therefore if our violation process depends linearly on S in some sense, then we can include $\hat{\lambda}$ and $\log(1 + x_6)$ in our models[4] in the place of S. The parameters estimated for $\hat{\lambda}$ and $\log(1 + x_6)$ will then be the same as those for S (up to a scalar multiple), and will allow us to have a well identified model.

Qualitatively this implies that a firm's characteristics can affect detection in two ways: by directly making violations easier to detect, and by raising an inspector's suspicion

[3]We use the transformed duration rather than the duration to incorporate the fact that extremely long inspections are probably more due to administrative time than to an inspector being suspicious. This also gives a better fitting model empirically.

[4]This is the approach to missing data that is criticised in Little and Rubin (1983), however we shall use it here simply for the sake of tractibility.

about a worksite.

Fitting this model in R is fairly straightforward, and the estimated coefficients can be found in Table 6.3. The maximum log-likelihood is considerably higher than in the models above, at -3958.363.

Covariate	Violation		Detection	
(Intercept)	1.21394	(0.35722)	-3.63892	(1.15873)
x_1=2621	0.19969	(0.34194)	0.90809	(1.20088)
x_1=2631	-0.41099	(0.43047)	3.31043	(1.41716)
x_1=2652	-0.62668	(0.50174)	4.70187	(1.63623)
x_1=2653	-0.40963	(0.36960)	3.46873	(1.24931)
x_1=2655	-0.94796	(0.45820)	4.68683	(1.49504)
x_1=2656	-0.99357	(0.56361)	4.56433	(1.77182)
x_1=2657	-0.55136	(0.40657)	4.35631	(1.33044)
x_1=2671	-0.20680	(0.48458)	3.18581	(1.66932)
x_1=2672	-0.15635	(0.45083)	2.26545	(1.51906)
x_1=2673	-0.55741	(0.38216)	4.09633	(1.31785)
x_1=2674	-0.61976	(0.58621)	4.22333	(1.79366)
x_1=2675	-0.54914	(0.52213)	3.97976	(1.59508)
x_1=2676	0.28024	(0.48262)	0.65588	(1.85475)
x_1=2677	-0.39178	(0.40986)	3.78952	(1.36236)
x_1=2678	-0.38314	(0.47354)	2.73703	(1.50284)
x_1=2679	-0.36454	(0.38162)	3.51306	(1.25966)
x_2=Partial	-0.83042	(0.18746)	0.94168	(0.62141)
x_3=Safety	1.07501	(0.14994)	-2.37574	(0.64005)
x_4=Unionised	0.30643	(0.11120)	-0.85386	(0.46773)
x_5=Notice Given	-1.39948	(0.36903)	17.2514	(2097.16)
Suspicion:				
$\log(1+x_6)$			1.02060	(0.16229)
$\hat{\lambda}$			0.22072	(0.11355)
d	1.281470	(0.0726924)		

Table 6.3: Estimated Coefficients and Standard Errors For Full Model with Suspicion

The qualitative results outlined above remain, however there is a problem with the estimate of the effect of advanced notice on the detection process. Due to the very small sample size, this parameter is not well identified under this model[5]. We therefore drop it from the detection process – which is equivalent to assuming that inspectors do not alter their behaviour based on whether a firm has been given advanced notice, but that advanced notice will only be issued to firms with low levels of violations, or will result in firms 'cleaning up' before the inspector arrives.

[5]The observed information matrix has a very small eigenvalue, with associated eigenvector in the direction of the parameter in question.

Another clear result is that, once again, many of the industries are very similar, and so we can fit the model using only the first 3 digits of the SIC code, for a more generic grouping. Doing so results in the parameters in Table 6.4 and a log-likelihood of -3973.99. Testing if these models are significantly different (using a χ^2_{24} approximation) results in a P-value of 0.1466, indicating no significant difference.

Covariate	Violation		Detection	
(Intercept)	1.09435	(0.40525)	-3.20129	(1.28831)
x_1=262	0.22598	(0.40109)	0.92687	(1.34314)
x_1=263	-0.24727	(0.47427)	2.73880	(1.57327)
x_1=265	-0.50065	(0.41804)	3.80959	(1.38904)
x_1=267	-0.33247	(0.41359)	3.34419	(1.36791)
x_2=Partial	-0.64932	(0.18526)	0.16737	(0.58655)
x_3=Safety	1.08017	(0.13088)	-2.35215	(0.72622)
x_4=Unionised	0.37646	(0.11180)	-1.00421	(0.48938)
x_5=Notice Given	-0.28346	(0.29484)		
Suspicion:				
$\log(1 + x_6)$			1.06433	(0.16233)
$\hat{\lambda}$			0.20291	(0.10488)
d	1.242076	(0.06959422)		

Table 6.4: Estimated Coefficients and Standard Errors For Simplified Model with Suspicion

As a final note, these estimates of the violation and detection probabilities indicate that, over our dataset, the mean rate of violation was 5.9805 with a standard deviation of 3.047975, while the mean probability of detection was only 0.513835, with a standard deviation of 0.2812096.

Chapter 7

Conclusions

" Wir müssen lernen, die Menschen weniger auf das, was sie tun und unterlassen, als auf das, was sie erleiden, anzusehen."

Dietrich Bonhöffer
Nach zehn Jahren (1943)

"One man's death is a tragedy. The death of a million is a statistic."

Anonymous (commonly attributed to Josef Stalin)

We now come to the end of this thesis, and we will summarise the results obtained in three sections: Economic modelling results, Statistical results and finally some Practical results.

7.1 Economic Models

We have seen that the presence of a regulator has a significant impact on a firm's compliance with OH&S regulations. Regulators can act in a variety of ways, and a targeted enforcement system allows for higher levels of compliance with less effort. At the same time, we also note that a regulator's goals may go beyond compliance, as political and macroeconomic concerns are also present.

We have shown that the results of Harrington (1988) depend on the distribution of costs c and discount factors β through the population. In situations where compliance is not a binary choice and where there are diminishing returns to effort, we see that firms may not choose whether to 'violate' or 'not violate', but some intermediate policy. Even so, some of the qualitative results of Harrington's work appear to remain, namely that a

firm will typically comply more when it is being targeted than otherwise. We have also extended Harrington's analysis to encompass the effects of the discount factor on a firm's decision making process, and have seen that under his model, a high discount factor can only drive a firm from complete non-compliance to partial compliance, but never to full compliance. We have also investigated this model in a nonlinear setting, and shown that a high discount factor may entail higher long term costs, even after correcting for the inevitable difference due to the discount factor.

We have also created a new model, in the spirit of Harrington's, to investigate the effects of using an unavoidable insurance levy as a regulatory tool. In such a situation we can often still achieve very good levels of compliance. Also, under this model the discount factor plays an even larger role in determining a firm's actions, as it is only the threat of future penalties that leads a firm to comply.

Comparing these models also allows us to see that increased costs of compliance do lead to higher long term costs, however the ability to choose non-compliance implies that the effects are less when costs are already high.

From a practical viewpoint, under the assumption that firms are expected discounted cost minimisers, our results confirm that a system which is more punitive is more effective at enforcing compliance. At the same time, it is quite reasonable for the fine/levy to be very low for firms in the 'good' group, as this will not discourage compliance, but will prevent regulation costs becoming too onerous, provided a firm complies.

7.2 Statistical Methods

We have looked at the DCE framework, under which we can attempt to distinguish between a violation and a detection process, using only data on detected violations. We have seen that this depends heavily on the existence of a covariate which contributes to one of these processes and not the other. If violations are not binary, then it is critical that the link function chosen to relate the covariates with the rate of detection is not pathological, and we have developed a method to test this.

In this regard our results improve on the methods available in the literature, as they do not depend on the existence of continuous covariates, let alone continuous covariates with unbounded support. This makes these methods far more applicable to real data problems, and can help a researcher in determining appropriate models that can be used.

We have also outlined a method by which the suspicion of violation can be incorporated into a DCE model.

7.3 Practical Questions

We have applied this statistical methodology to a dataset obtained from OSHA (the U.S. Occupational Safety and Health Administration), and can see that it indicates some interesting phenomena. We find that firms undergoing safety inspections have significantly higher underlying rates of violation than otherwise, but that these violations are considerably less likely to be detected. Similarly, we notice that workplaces with unionised employees generally have higher levels of violation – however causation is unclear in this case.

Given more extensive data, these conclusions could be considerably strengthened. In particular it would be good to have access to other non-binary variables, for example plant size, age and mean employee income, as these may help to lessen our estimates' dependence on inspection duration. Such data is available in certain circumstances (as is discussed by Gray and Shadbegian (2005)), however is difficult to access outside of the U.S.

It would be interesting to apply this methodology to data from other economies – for example similar data from the U.K. is also available online, through the Health and Safety Executive records websites `http://www.hse.gov.uk/notices/` and `http://www.hse.gov.uk/prosecutions/`. Data for Australia is not easily available to the public, however this methodology could quite reasonably be applied 'in house' by a regulator.

Overall, there are still a variety of issues in the economics and statistics of regulation that have not been addressed. This is a large and active area of economic thought, and we have merely scratched the surface of the complexities that underlie the workings of this system.

Bibliography

Bartel, A. P. and Thomas, L. G. (1985). Direct and indirect effects of regulation: A new look at OSHA's impact, *Journal of Law and Economics* **28**(1): 1–25.

Becker, G. S. (1968). Crime and punishment: An economic approach, *The Journal of Political Economy* **76**(2): 169–217.

Bellman, R. E. (1957). *Dynamic Programming*, 2003 edn, Dover Publications.

Bellman, R. E. (1961). *Adaptive Control Processes: A Guided Tour*, Princeton University Press.

Bonhöffer, D. (1943). Nach zehn jahren, *in* C. Gremmels, E. Bethge and R. Bethge (eds), *Widerstand Und Ergebung: Briefe und Aufzeichnungen aus der Haft*, 1998 edn, Vol. 8 of *Dietrich Bonhöffer Werke*, Chr. Kaiser Verlag, Gütersloh.

Boyer, M., Lewis, T. R. and Liu, W. L. (2000). Setting standards for credible compliance and law enforcement, *Canadian Journal of Economics/Revue canadienne d'économique* **33**(2): 319–340.

Bunyan, J. (1678). *The Pilgrim's Progress*, Penguin Classics, 1987 edn, Penguin Books Ltd., London.

Carroll, L. (1876). *The Hunting of the Snark*, Penguin Classics, 1995 edn, Penguin Books Ltd., London.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partially linear single-index models, *Journal of the American Statistical Association* **92**(438): 477–489.

Clark, J. E., Friesen, L. and Muller, R. A. (2004). The good, the bad, and the regulator: An experimental test of two conditional audit schemes, *Economic Inquiry* **42**(1): 69–87.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1): 1–38.

Dickens, C. (1837). *Oliver Twist*, Penguin English Library, 1979 edn, Penguin Books Ltd., Harmondsworth, Middlesex.

Feinstein, J. S. (1989). The safety regulation of U.S. nuclear power plants: Violations, inspections, and abnormal occurrences, *The Journal of Political Economy* **97**(1): 115–154.

Feinstein, J. S. (1990). Detection controlled estimation, *Journal of Law and Economics* **33**(1): 233–276.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics, *Philosophical Transactions of the Royal Society, A* **222**: 309–368.

Friesen, L. (2003). Targeting enforcement to improve compliance with environmental regulations, *Journal of Environmental Economics and Management* **46**: 72–85.

Gordon, S. C. and Hafer, C. (2005). Flexing muscle: Corporate political expenditures as signals to the bureaucracy, *American Political Science Review* **99**(2): 245–261.

Gordon, S. C. and Smith, A. (2004). Quantitative leverage through qualitative knowledge: Augmenting the statistical analysis of complex causes, *Political Analysis* **12**: 233–255.

Gormley, Jr., W. T. (1979). A test of the revolving door hypothesis at the FCC, *American Journal of Political Science* **23**(4): 665–683.

Gourieroux, C., Monfort, A. and Trognon, A. (1984a). Pseudo maximum likelihood methods: Applications to Poisson models, *Econometrica* **52**(3): 701–720.

Gourieroux, C., Monfort, A. and Trognon, A. (1984b). Pseudo maximum likelihood methods: Theory, *Econometrica* **52**(3): 681–700.

Gray, W. B. (1987). The cost of regulation: OSHA, EPA and the productivity slowdown, *The American Economic Review* **77**(5): 998 –1006.

Gray, W. B. and Shadbegian, R. J. (2005). When and why do plants comply? paper mills in the 1980s, *Law & Policy* **27**(2): 238–261.

Guo, S. (1999). *The influence of OSHA Inspectors' Detection Capabilities on OSHA's Effectiveness: Evidence from a Panel Data 1979-1985*, PhD thesis, Department of Economics: Clark University, Worcester, Massachusetts.

Harrington, W. (1988). Enforcement leverage when penalties are restricted, *Journal of Public Economics* **37**: 29–53.

Horowitz, J. L. and Hardle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates, *Journal of the American Statistical Association* **91**(436): 1632–1640.

Horwitz, T. (1994). *9 to Nowhere*, The Wall Street Journal, December 1.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes*, The MIT Press (Technology), Cambridge, Massachusetts.

Hugo, V. (1853). *Choix de Poèmes*, 1968 edn, Manchester University Press.

Innes, J. (2002). Origins of the factory acts, *in* N. Landau (ed.), *Law, Crime and English Society, 1660-1830*, Cambridge University Press, pp. 230–233.

Koopmans, T. C. (1949). Identification problems in economic model construction, *Econometrica* **17**(2): 125–144.

Laffont, J.-J. and Tirole, J. (1986). Using cost observation to regulate firms, *The Journal of Political Economy* **94**(3 Part 1): 614–641.

Laffont, J.-J. and Tirole, J. (1991). The politics of government decision-making: A theory of regulatory capture, *The Quarterly Journal of Economics* **106**(4): 1089–1127.

Laffont, J.-J. and Tirole, J. (1993). *A Theory of Incentives in Procurement and Regulation*, The MIT Press, Cambridge, Massachusetts.

Lehmann, E. L. (1983). *Theory of Point Estimation*, John Wiley & Sons.

Little, R. J. A. and Rubin, D. B. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood, *The American Statistician* **37**(3): 218–220.

Little, R. J. A. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values, *Biometrika* **72**(3): 497–512.

Livernois, J. and McKenna, C. (1999). Truth or consequences: enforcing pollution standards, *Journal of Public Economics* **71**: 415–440.

Marx, K. (1867). *Capital: A Critique of Political Economy*, tr. Moore, S. and Aveling, E., 1903 edn, Charles H. Kerr and Co., Chicago.

Marx, K. and Engels, F. (1848). *The Communist Manifesto*, Great Ideas, 2004 edn, Penguin Books Ltd, London.

Mendeloff, J. and Gray, W. B. (2005). Inside the black box: How do OSHA inspections lead to reductions in workplace injuries?, *Law & Policy* **27**(2): 219–237.

Milne, A. A. (1927). *Now We Are Six*, The Pooh Gift Box, 1980 edn, Methuen Childrens Books Ltd, London.

OSHA (1993). *Changes to the Field Operations Manual (FOM)*, 13 December 1993 edn, U.S. Dept of Labor, Occupational Safety and Health Administration: Office of General Industry Compliance Assistance.

OSHA (2007). Inspection data, `http://www.osha.gov/oshstats/index.html`, Accessed 15/5/07.

Pascal, B. (1670). *Pensées et Opuscules*, 1949 (10^e) edn, Classiques Larousse, Paris (VI^e).

Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley & Sons, New York.

Sappey, R., Burgess, J., Lyons, M. and Buultjens, J. (2006). *Industrial Relations in Australia: Work and Workplaces*, Pearson-Prentice Hall, pp. 370– 388.

Smith, A. (1776). *The Wealth of Nations*, Pelican Classics, 1978 edn, Pelican Books.

Stafford, S. (2005). Does self-policing help the environment? EPA's audit policy and hazardous waste compliance, *Vermont Journal of Environmental Law* **6**: Online.

Stafford, S. L. (2006). Self-policing in a targeted enforcement regime, *Working Papers 26*, Department of Economics, College of William and Mary. Available at http://ideas.repec.org/p/cwm/wpaper/26.html.

Stigler, G. J. (1971). The theory of economic regulation, *The Bell Journal of Economics and Management Science* **2**(1): 3–21.

Weil, D. (1996). If OSHA is so bad, why is compliance so good?, *The RAND Journal of Economics* **27**(3): 618–640.

Workcover SA (2006a). Levy information, Available at `http://www.workcover.sa.gov.au/`, Adelaide.

Workcover SA (2006b). Supplementary levy program 2006-07, Available at `http://www.workcover.sa.gov.au/`, Adelaide.

Workcover SA (2007a). 2007-2008 industry levy rates, Available at `http://www.workcover.sa.gov.au/`, Adelaide.

Workcover SA (2007b). A guide to the 2007-08 bonus/penalty scheme, Available at `http://www.workcover.sa.gov.au/`, Adelaide.

Workcover SA (2007c). A guide to the 2007-08 safework incentive for large employers, Available at `http://www.workcover.sa.gov.au/`, Adelaide.

Appendix A

Further Derivations

"The method employed I would gladly explain,
While I have it so clear in my head,
If I had but the time and you had but the brain —
But much yet remains to be said. "

Lewis Carroll
The Hunting of the Snark (1876)

A.1 Feinstein's Theorems

Rephrased into the parametric case, Feinstein (1990) begins by claiming that as long as each of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ both contain at continuous components (and are not the same), the condition for identification to fail is that there exists a point $(\boldsymbol{\xi}^*, \boldsymbol{\eta}^*)$, a neighbourhood of which posesses positive density, for which

$$F(\boldsymbol{\xi}^*\boldsymbol{\beta}_1)G(\boldsymbol{\eta}^*\boldsymbol{\beta}_2) = F(\boldsymbol{\xi}^*\boldsymbol{\beta}_1^*)G(\boldsymbol{\eta}^*\boldsymbol{\beta}_2^*)$$

for some $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_2^*$, where $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ are the true values.

He states the following two theorems:

Theorem A.1.1 (Feinstein's Theorem A1 (p. 271))**.** *Assume that $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ contain only continuous components (apart from intercepts) and that they have some elements in common. If identification fails, the link functions must each belong to the exponential family.*

Theorem A.1.2 (Feinstein's Theorem A2 (p. 271))**.** *Assume that each of $\boldsymbol{\xi}$ and $\boldsymbol{\eta}$ posesses at least one continuous component with unbounded support and that at least one of each of these components enters only into $\boldsymbol{\xi}$ and only into $\boldsymbol{\eta}$. Then our parameters are identified up to a constant and scalar multiple, but the scalar multiples may differ.*

Without further strengthening, Theorem A1 is false. We can remedy this if we first specify that $\boldsymbol{\xi}$ is not a linear function of $\boldsymbol{\eta}$ nor vice versa (this excludes the example presented in Section 5.5.5) which implies that *both* our violation and detection processes must contain (continuous) covariates that do not affect the other process. We then change 'identifiable' to 'identifiable up to a constant and scalar multiple' (which excludes the example presented in Section 5.5.1). With these changes, the theorem appears to be valid. The interested reader is referred to Feinstein's paper for a proof – there is little point reproducing it here.

Even then, the requirement that all our data must be continuous is impracticable for this context. Feinstein's Theorem A2 loosens this requirement somewhat, requiring that there must be at least one continuous component with unbounded support. This has merit in that it also applies to a semiparametric estimation approach, however the method proposed in our discussion may still be preferable, as typically a variable taking only 3 values is sufficent.

A.2 Identifiability with Logistic regression

We wish to show that given any three distinct values of ξ, denoted ξ_1, ξ_2, ξ_3, if we know the values of

$$\begin{aligned}\phi(\beta_1 + \beta_2\xi_2) - \phi(\beta_1 + \beta_2\xi_1) &= k_1,\\ \phi(\beta_1 + \beta_2\xi_3) - \phi(\beta_1 + \beta_2\xi_1) &= k_2,\end{aligned}$$

where $\phi(x) := -\log(1 + e^{-x})$, then we can uniquely determine the value of β_1.

To do so, we first note that it is sufficient to show that this is true for the points $\xi = (0, 1, x)$ for any $x > 1$. The general case then follows by scaling and translation. We know from the analysis surrounding the example in Section 5.2 that when $x = 2$ this can be done. (Provided $k_2/k_1 > 1$, which is required for a solution to exist.)

We now note that we are trying to show that for each value of x, the equation

$$0 = x\phi^{-1}(k_1 + \phi(\beta_1)) - \phi^{-1}(k_2 + \phi(\beta_1)) - (x - 1)\beta_1$$

has a unique solution for β_1. We can now take the total differential of this function (with respect to x and β) to find

$$\begin{aligned}0 &= \left[\phi^{-1}(k_1 + \phi(\beta_1)) - \beta_1\right] dx + \left[x\frac{\phi'(\beta_1)}{\phi'(\phi^{-1}(k_1 + \phi(\beta_1)))} - \frac{\phi'(\beta_1)}{\phi'(\phi^{-1}(k_2 + \phi(\beta_1)))} - x + 1\right] d\beta_1\\ &= \left[\beta_1 + \beta_2 - \beta_1\right] dx + \left[x\frac{\phi'(\beta_1)}{\phi'(\beta_1 + \beta_2)} - \frac{\phi'(\beta_1)}{\phi'(\beta_1 + \beta_2 x)} - x + 1\right] d\beta_1,\end{aligned}$$

and so

$$\frac{d\beta_1}{dx} = -\frac{\beta_2}{x\frac{\phi'(\beta_1)}{\phi'(\beta_1+\beta_2)} - \frac{\phi'(\beta_1)}{\phi'(\beta_1+\beta_2 x)} - x + 1}.$$

For $x > 1$, this is continuous and finite. Continuity is clear, to see finiteness we note $\phi'(x) = \frac{1}{1+e^x}$, and therefore the denominator is

$$x\frac{1+e^{\beta_1+\beta_2}}{1+e_1^{\beta}} - \frac{1+e^{\beta_1+\beta_2 x}}{1+e_1^{\beta}} - x + 1.$$

Multiplying through by $1 + e^{\beta_1}$ (which is positive), we obtain

$$x(1+e^{\beta_1+\beta_2}) - (1+e^{\beta_1+\beta_2 x}) - (x-1)(1+e^{\beta_1}) = xe^{\beta_1+\beta_2} - e^{\beta_1+\beta_2 x} - (x-1)e^{\beta_1}.$$

Now dividing through by e^{β_1},

$$xe^{\beta_2} - e^{\beta_2 x} - (x-1) = x(e^{\beta_2} - 1) - (e^{\beta_2 x} - 1),$$

which is clearly zero for $x = 1$, and which has derivative (with respect to x)

$$e^{\beta_2} - 1 - \beta_2 e^{\beta_2 x},$$

which is less than zero for all $x > 1$ and β_2. (Evaluate at $x = 1$, show this is at most zero and then consider the fact that the derivative with respect to x is clearly negative.)

We therefore know that $\frac{d\beta_1}{dx}$ is continuous and is finite for all $x > 1$. Hence for any $x > 1$, we can consider a compact interval I (the interior of which contains x and 2) on which $\frac{d\beta_1}{dx}$ is clearly Lipschitz continous, and so we can apply Picard's uniqueness theorem to the differential equation given by $\frac{d\beta_1}{dx}$ with boundary condition given by the unique solution for $x = 2$. Hence, there is a unique function relating β_1 and x, and therefore for any value of $x > 1$, we have shown there is a unique value of β_1.

NB. Some modifications may be needed if $k_2/k_1 \leq 1$, namely to show that we can 'piece together' solutions connecting $x = 2$ and $x = \xi$, each of which shows that locally we have uniqueness for the appropriate collection of k_2 and k_1. Nevertheless, the general argument should follow *mutatis mutandis*.

A.3 Left-invertibility and Covariates

Suppose that the vector of covariates $\boldsymbol{\xi}$ contains at least a variables which can take on at least $a + 2$ values, which are not included in $\boldsymbol{\eta}$. (See Section 5.4 for a definition of this notation in context.) We assume that these values can be taken given any value of $\boldsymbol{\eta}$, and that they are linearly independent of the remaining covariates in $\boldsymbol{\xi}$.

For each value of $\boldsymbol{\eta}$, we define our matrix of interest Ξ to be the matrix with rows given by the values of $\boldsymbol{\xi}$, omitting one reference case $\boldsymbol{\xi}_l$ and all covariates which are also contained in $\boldsymbol{\eta}$ (or more generally are functions of $\boldsymbol{\eta}$). By defining our matrix in this way, we are essentially conditioning on the value of $\boldsymbol{\eta}$, which corresponds with the logic of Section 5.3. We wish to show that Ξ has a left-inverse.

We recall from the properties of linear regression, for a model of the form $Y = X\beta$, a necessary and sufficient condition for X to have a left-inverse is that our predictors are not perfectly multicollinear. This corresponds to the columns of X being linearly independent, and in this case the left-inverse is given by $(X^TX)^{-1}X^T$. We can apply the same requirement to our matrix Ξ.

Consider the matrix A defined by only considering the a variables mentioned above and an intercept term. These variables take on at least $a + 2$ values, and so even after we have removed one, we have $a + 1$ linearly independent rows in our $(a + 1) \times (a + 1)$ matrix A, one corresponding to each remaining combination of covariates. Therefore the columns of A cannot be linearly dependent.

Our matrix Ξ can then be decomposed into submatrices $\Xi = [A|B]$, where A is the matrix above and B corresponds to those covariates not in η and not in A. We assumed linear independence between the covariates in A and B, and so we know that the columns of A are not linearly dependent on those of B. Therefore we must still have linearly independent columns in Ξ and so Ξ has a left-inverse.

Practically, we shall often take $a = 1$, and therefore the requirement is that the process of interest (typically detection) has at least one covariate which takes at least three values, and that this covariate does not affect the other process (typically violation). Alternatively, we may take $a = 2$, in which case we simply need two binary variables with no interactive effect, which affect the process of interest but not the other. This is a considerably weaker requirement than requiring *both* processes to have *continuous* covariates which do not affect the other process.

Appendix B

Further Results

"Errors using inadequate data are much less than those using no data at all."

Charles Babbage

B.1 Grouping Industries

If only the first three digits of the SIC code are used, the estimated coefficients are as in Table B.1.

Covariate	Violation		Detection	
(Intercept)	0.38108	(0.30031)	-1.52745	(1.01655)
x_1=262	0.80133	(0.30659)	0.40500	(1.01380)
x_1=263	0.44214	(0.33589)	1.29271	(1.07606)
x_1=265	0.29237	(0.30328)	1.93680	(1.01197)
x_1=267	0.38876	(0.30191)	1.80386	(1.01160)
x_2=Partial	-0.67766	(0.14360)	-0.54823	(0.30926)
x_3=Safety	0.81832	(0.08472)	-0.86702	(0.29366)
x_4=Unionised	0.24472	(0.08085)	-0.26658	(0.18455)
x_5=Notice Given	-0.86214	(0.45049)	0.93458	(1.17289)
x_6			1.28919	(0.26786)
d	1.196393	(0.06560338)		

Table B.1: Estimated Coefficients and Standard Errors

The minimum log-likelihood is -3988.568.

B.2 Using Transformed Durations

Again to reduce the impact of very long inspections, an alternative model was fitted using the transformed values

$$x'_6 = \log(1 + x_6)$$

The results of this can be seen in Table B.2

Covariate	Violation		Detection	
(Intercept)	1.08676	(0.54068)	-2.86769	(0.97164)
x_1=262	0.27948	(0.53008)	1.59345	(0.96748)
x_1=263	-0.19704	(0.56939)	2.49056	(1.05188)
x_1=265	-0.40900	(0.54085)	3.17134	(0.96913)
x_1=267	-0.28067	(0.53662)	2.99798	(0.96609)
x_2=Partial	-0.60400	(0.19593)	-0.60492	(0.35184)
x_3=Safety	1.00109	(0.11313)	-1.01612	(0.29315)
x_4=Unionised	0.24477	(0.10169)	-0.22328	(0.20221)
x_5=Notice Given	-0.99409	(0.48321)	1.12510	(1.19456)
$x'_6 = \log(1 + x_6)$			1.20853	(0.17955)
d	1.235340	(0.06892)		

Table B.2: Estimated Coefficients and Standard Errors

The minimum log-likelihood is -3976.193.

B.3 Incorporating Interactions

The model in Appendix B.2 was also fitted with interactions between some of the variables included, the results were as in Table B.3

The minimum log-likelihood is -3970.086.

Covariate	Violation		Detection	
(Intercept)	0.95055	(0.50696)	-2.33475	(1.02066)
x_1=262	0.35019	(0.49162)	1.49519	(0.93313)
x_1=263	-0.11649	(0.52845)	2.41553	(1.02781)
x_1=265	-0.34782	(0.49685)	3.16549	(0.93789)
x_1=267	-0.22247	(0.49650)	2.97290	(0.94304)
x_2=Partial	-0.59916	(0.40484)	0.26201	(1.02627)
x_3=Safety	1.04754	(0.13413)	-1.53779	(0.48839)
x_4=Unionised	0.37080	(0.15522)	-1.19223	(0.59512)
x_5=Notice Given	-0.88939	(0.47508)	1.23368	(1.13542)
x_2 and x_3	0.21170	(0.43897)	-0.85286	(1.02832)
x_2 and x_4	-0.33328	(0.37532)	-0.65205	(0.71801)
x_3 and x_5	-0.14091	(0.19553)	1.11806	(0.63913)
x_6'			1.28783	(0.19447)
d	1.235340	(0.06892)		

Table B.3: Estimated Coefficients and Standard Errors

Appendix C

Computer Code

"What of my dross thou findest there, be bold
To throw away; but yet preserve the gold.
What if my gold be wrapped up in ore?
None throws away the apple for the core."

John Bunyan
The Pilgrim's Progress (1678)

The methods used in this Thesis were all implemented using the `R` statistical computing environment.

C.1 Code to implement DCE methods

We first read all the data, then reformat and extract the variables of interest.

```
#Further extract only Planned inspections where the
# inspection was completed, put the covariates into a matrix

(type=='Planned')&(x2!='No Insp')->RelDat
Xi<-model.matrix(yFull~x1+x2+x3+x4+x5+x6)[RelDat,]
y<-yFull[RelDat]
```

Next, some generic methods to be used as link functions were installed

```
ilogit<-function(x){1/(1+exp(-x))}
logit<-function(x){log(1/((1/x)-1))}
```

Then vectors to indicate which covariates are included in which process are generated

```
#Process 1 = Violation, does not depend on x6
#Process 2 = Detection, can depend on anything
m1<-c(rep.int(1,dim(Xi)[2]-1),0)
m2<-c(rep.int(1,dim(Xi)[2]-1),1)
```

We now define our log likelihood function:

```
llik<-function(y,Xi, beta1, beta2, disp, m1, m2){
  rate<-exp(Xi%*%(beta1*m1))*ilogit(Xi%*%(beta2*m2))
  sum(dnbinom(y,disp, mu=rate, log=T))
}
```

To simultaneously fit all the required parameters, we will create a vector `beta` which contains the parameters for the first process, then for the second process, and then the dispersion parameter. We will also create a version of the log likelihood which depends only on these:

```
#Initialise beta to c(0's,0's,1)
betas<-c(m1*0, m2*0,1)

#Negative log likelihood function, based only on betas
fixednllik<-function(betas){
  -llik(y, Xi, betas[1:((length(betas)-1)/2)],
    betas[((length(betas)+1)/2):(length(betas)-1)],
    betas[length(betas)], m1, m2)
}
```

We can now fit the model using the `nlm` and `optim` commands (two methods are used to hopefully improve numerical stability)

```
nlm(fixednllik, betas, print.level=2,
  iterlim=5000)->results
optim(results$estimate, fixednllik, control=list(trace=2,
  reltol=1e-8, maxit=50000), hessian=T)->results
```

This gives estimates of the parameters and also of the observed information matrix, from which we can extract standard errors

```
results$par->betas
results$hessian->I

#Invert the observed information matrix, but only for 'real'
# parameters, i.e. where the covariate is actually used

temp<-solve(I[mod==1, mod==1])
V<-matrix(0,nrow=length(betas), ncol=length(betas))
V[mod==1, mod==1]<-temp
s<-sqrt(diag(V))
```

Finally, we can output these results quickly using the following script

```
b1<-betas[1:((length(betas)-1)/2)]
b2<-betas[((length(betas)+1)/2):(length(betas)-1)]
s1<-sqrt(diag(V))[1:((length(betas)-1)/2)]
s2<-sqrt(diag(V))[((length(betas)+1)/2):(length(betas)-1)]
disp<-betas[length(betas)]
matrix(c(labels(Xi)[[2]],round(b1,5),paste("(",round(s1,5),")", sep="")
  ,round(b2,5),paste("(",round(s2,5),")", sep="")), nrow=length(b1))
```

and this final matrix can be directly written to LaTeXusing the `quantreg` package.

C.2 Code used to generate graphs

C.2.1 Code used to generate Figure 5.2

To generate this figure, we first specify the link functions - note that these can be replaced with any (invertible) link function, and the code will still function.

```
F<-function(x){exp(x)}
G<-function(x){exp(x+sin(log(x)))}
```

and define a function to invert the log of one of them

```
ilogG<-function(x){
temp<-function(y){sum(abs(log(G(y))-x))}
optimise(temp,c(0.2,1000))$minimum}
```

We now fix some points at which we will assume we have data, and some estimates of $\beta_1, ...\beta_5$

```
b1<-0.3; b2<-0.9; b3<-0.4; b4<-0.6; b5<--0.3
xi<-(0:9)/9; eta<-0:2
Xi<-matrix(nrow=30, ncol=3)
for(i in 1:30){
 Xi[i,1]<-1
 Xi[i,2]<-xi[(i-1)%%10+1]
 Xi[i,3]<-eta[((i-1)-(i-1)%%10)/10+1]}
XX<-solve(t(Xi)%*%Xi)%*%t(Xi)
```

From here, we can find the k_i values, and hence define the Γ function

```
k<-0
for(i in 1:30){
  k[i]<-log(G(c(b4,b5,0)%*%Xi[i,]))-log(G(c(b4,0,0)%*%Xi[i,]))}
Gamma<-function(b){
 temp<-0
 for(i in 1:length(k)){temp[i]<-ilogG(log(G(b))+k[i])}
 (XX%*%temp)[1]-b}
```

And so finally can plot Γ vs β as desired. (A small smoothing term is included to dissipate numerical errors introduced by the numerical inversion procedure.)

```
b<-c((200:1500)/1000, (151:500)/100);q<-0
for(i in 1:length(b)){q[i]<-Gamma(b[i])}
plot(c(0.2,5), c(0,0), col='grey','l', xlab=expression(beta),
  ylab=expression(Gamma), main=expression(Gamma * " vs. " * beta *
  " over [0.2,5]"), ylim=c(min(q), max(q)), xlim=c(0.2,5), log='x')
lines(b,smooth(q))
```

C.2.2 Code used to generate Figure 5.3

To generate this figure, the following code was used. Changing the code to account for different estimates of the β's would be a trivial matter.

```
iphi<-function(x)
qnorm(x, 0,1, log.p=T)
```

```
phi<-function(x)
pnorm(x, 0,1, log.p=T)

beta_1<-1; beta_2<-2; beta_3<-1

k1<-phi(beta_1+beta_3)-phi(beta_1)
k2<-phi(beta_1+2*beta_3)-phi(beta_1)
k3<-phi(beta_1+beta_2+beta_3)-phi(beta_1+beta_2)
k4<-phi(beta_1+beta_2+2*beta_3)-phi(beta_1+beta_2)

x<-(-2000:4000)/1000
G1<-2*iphi(phi(x)+k1)-iphi(phi(x)+k2)-x
G2<-2*iphi(phi(x)+k3)-iphi(phi(x)+k4)-x
plot(x, G2,'l', ylim=c(1e-12,1), xlab=expression(beta),
  ylab=expression(Gamma), main=expression(Gamma
  * " vs. " * beta * " over relevant domain"), log='y')
```

C.2.3 Code used to generate Figure 3.2

To generate this figure, the following variables were set:

$$\beta = 0.5, \phi_0 = 0.75, \phi_1 = 0.5, F_1 = 400, F_0 = 700, p = 0.25, g = 0.8,$$

The following code was then used:

```
c<-0:1000; b<-0.5; phi0<-0.75; phi1<-0.5; F1<-400;
F0<-700; g<-0.8; p<-0.75
C1<-c/(1-b)
A<-1-b*(1-phi1*p)
B<-b*phi0*g
C2<-(1/(1-b))*(A*c+B*phi1*F1)/(A+B)
C3<-rep(phi0*F0/(1-b), length(c))
L0<-phi1*F1
L1<-phi0*F0+b*phi0*g*(phi0*F0-phi1*F1)/(1-b*(1-phi1*p))
plot(c,C1,'l', xlab="c = Cost of Compliance", ylab="E[C*(0)] for
  each policy", main="E[C*(0)] vs. c", col="grey",
  ylim=c(-50,2000))
```

C.2.4 Code used to generate Figure 3.3

To generate this figure, the following parameters were set:

$$c = 700, \phi_0 = 0.75, \phi_1 = 0.5, F_1 = 0, F_0 = 700, p = 0.25, g = 0.8,$$

The code used was similar to that in C.2.3, details are available from the author.

C.2.5 Code used to generate Figure 3.4

The parameters used were

$$\phi_0 = 0.75, \phi_1 = 0.5, F_1 = 400, F_0 = 700, g = 0.9, p = 0.3$$

and the surface is generated for c between 20 and 400, and β between 0.2 and 0.8.

The code used was similar to that in C.2.3, details are available from the author.

C.2.6 Code used to generate Figure 3.6

The parameters used were:

$$\beta = 0.8, F_1 = 100, F_0 = 700, g = 0.9, p = 0.3$$

The code used was similar to that in C.2.3, details are available from the author.

C.2.7 Code used to generate Figure 3.7

The parameters used were:

$$c = 270, F_1 = 100, F_0 = 900, g = 0.5, p = 0.3$$

The code used was similar to that in C.2.3, details are available from the author.

C.2.8 Code used to generate Figure 3.8

The parameters used were:

$$c = 600, F_1 = 100, F_0 = 800, g = 1, p = 0.2$$

The code used was similar to that in C.2.3, details are available from the author.

C.2.9 Code used to generate Figure 3.9

The parameters used were

$$\phi_0 = 0.75, \phi_1 = 0.5, F_1 = 100, F_0 = 900, g = 0.9, p = 0.3$$

and the surface is generated for c between 20 and 400, and β between 0.2 and 0.8.

The code used was similar to that in C.2.3, details are available from the author.

C.2.10 Code used to generate Figures 3.10 and 3.11

The parameters used were

$$\phi_0 = 0.75, \phi_1 = 0.5, F_1 = 400, F_0 = 700, \gamma_0 = 0.7, \gamma_1 = 0.3$$

and the surface is generated for α between 20 and 400, and β between 0.2 and 0.8.

The code used was similar to that in C.2.3, details are available from the author.