The Role of HRT Exposure in Breast Cancer Etiology: a Causal Inference Approach

Dr. Federico Andreis

Supervisor: Prof. Rino Bellocco
Co-Supervisor: Prof. Keith Humphreys

Dottorato di Ricerca in Statistica, XXIII Ciclo
Università degli Studi di Milano-Bicocca

to my family,
with all my love

Acknowledgements

Finding the right words to thank everybody who has had a part in this work and in my last three years' life turns out to be almost as tough as writing this thesis itself, I hope not to forget anybody, nor to be unfair in any sense! On both a personal and academic level, the biggest thanks goes to my supervisor, Prof. Rino Bellocco, who has been an extremely valuable mentor, for both research and life matters. Thanks to him I had the opportunity to see what an high level international research environment is like, through the Summer School he organizes every year and during the time I spent at MEB Department of Karolinska Institutet in Stockholm. I have the deepest respect for him and his work, as a teacher, as a researcher, as a man.
Thanks to Prof. Juni Palmgren, if not for her, I would not have had the opportunity to spend the wonderful year I spent at MEB Department, thanks for her encouragements and infinite kindness.
Thanks to Prof. Keith Humphreys from Karolinska Institutet, my co-supervisor, for the endless patience and professionality he devoted to my doubts, mistakes and ideas during my work for this thesis in Stockholm.
Thanks to Arvid Sjölander for the valuable discussions on his and my work, he for sure has one of the brightest minds I have ever had the luck to deal with.

Thanks to my family for supporting me in any way (and warning me when I take the wrong route!). I honestly believe I have been given the greatest gift a person can receive: a loving, sincere, unite and wonderful family.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of cancer etiology is one of the most interesting and difficult tasks in medical-related disciplines: unravelling the biology of such severe diseases has been one of the primary aims of medicine. Understanding the biological mechanisms that lead to the development of a tumor is, though, often not trivial, due either to our lack of knowledge on the matter or to the extreme complexity of the relationships among the involved factors.

Statistical methodology can provide useful tools to help improving our understanding of such intricate connections, thus making epidemiology a powerful instrument to integrate medical research.
In this project we will target the problem of correctly assessing and quantifying the causal effect of an exposure (Hormone Replacement Therapy, HRT) on the onset and the progression of breast cancer in post-menopausal women.

Evidence has been found in the literature that sex hormones play a central role in the etiology of breast cancer. Treatments such as HRT (a combination of estrogen/progestin) are known to be a possible cause for this kind of tumors, but confusion arose in the field when some studies reported women who underwent HRT and developed breast cancer to be having a better prognosis than those who did not take the treatment. Possible reasons have been investigated (see for example [1] for a summary) and hypotheses have been formulated, even though still lacking a strong evidence in their support. One confirmative analysis proposal is the one by Sjölander [10], who developed a model to assess differences in prognosis by contrasting HRT-induced (i.e. caused by hormon replacement therapy) cancer cases against non HRT-induced ones; his methods, despite being formally and mathematically convincing, are quite complicated for a non-statistician who is interested in applying them, moreover the conclusions on the strength of the link between HRT and prognosis do not have a causal interpretation. Furthermore the cancer sub-type distinction is not observable, thus requir-

ing many working assumptions.

In this project another kind of analysis is proposed, which may offer more flexibility and clarity, shifting from the semi-parametric methodology employed in [10] to a fully parametric setting. The problem of sub-type distinction is tackled through a clustering technique relying on additional information (i.e. covariates) on the women in the sample. Our method allows also, under proper assumptions, to approximately estimate the strength of the causal effect of HRT on breast cancer prognosis.
The need for distributional assumptions (not present in [10]) may here be a strength or a weakness, depending on our level of previous knowledge of the phenomenon under investigation. It induces more computational weight through more parameters we have to estimate, but may also help providing a more flexible and adherent-to-reality explanation if we can motivate our choices.

This work is structured as follows:

- **Chapter 2** deals with breast cancer epidemiology; notions on the biology of this kind of tumor are given, and the role of sex hormones is presented.
- In **Chapter 3** we summarize the topic of principal stratification in causal inference, first introducing potential outcomes and counterfactuals, then reviewing the seminal paper by Frangakis and Rubin [2] and the work by Gilbert [3]. A short introduction to a very useful graphical instrument, the DAGs, is given.
- **Chapter 4** contains a review of the approach developed by [10] to evaluate the association between HRT and breast cancer aggressiveness. In section 4.1 we summarize the basic concept behind the approach and in section 4.2 we describe the assumptions which where made to address the subgroups identifiability issue.
- In **Chapter 5** we develop an alternative approach to the problem, which aims at estimating the causal effect of HRT on the outcome of interest, reformulating the estimation procedures from a semi-parametric (as in [10]) into a full parametric setting.
- Applications of our methodology are presented in **Chapter 6**. We analyze data from the case-cohort study of swedish post-menopausal breast cancer "CAHRES", which was also considered by [10]. We analyze both a continuous (the Nottingham Prognostic Index, NPI) and a dichotomous outcome (5-years survival).

- In **Chapter 7** we describe the simulations we have carried out to compare our approach with the one by [10] and with the standard analysis, focusing on the issue of estimates' bias and how interpretation of results may go wrong if the model we decide to use is misspecified.
- **Chapter 8** contains the conclusive remarks about this work.

Chapter 2

Breast Cancer and HRT

2.1 Breast Cancer Epidemiology

Breast cancer is the most commonly occurring cancer in women worldwide, accounting for more than 20% of all cancer diagnoses among female individuals. During the last decades thousands of research papers have been published describing risk factors for this tumor.
There are at least twelve well established risk factors that influence breast cancer risk, plus many that either have been reported only inconsistently in the literature or have received only limited study to date; the following table summarizes the established factors, reporting direction and qualitative strength of the effect:

Risk factor	Direction of effect
Family history in first-degree relative	++
Height	++
Benign breast disease	++
Mammographically dense breasts	++
Age at first birth > 30 years vs < 20	++
Menopause at > 45 years vs < 45	++
High endogenous estrogen levels	++
Postmenopausal hormone use	+
Ionizing radiation exposure	++
Menarche at < 12 years vs > 14	+
High body mass index (postmenopausal)	+
High body mass index (premenopausal)	-

Table 2.1: Well-confirmed risk factors for breast cancer. + indicates a slight to moderate increase in risk, ++ a moderate to large increase in risk and - a slight to moderate decrease in risk.

Age incidence curves for breast cancer are generally similar in shape across countries, but there are big differences in terms of absolute rates at every age. Overall, rates substantially increase with age, the diagnosis is rare in women less than 40 years old and there is a slowing of the rate growths near the age of menopause, strongly suggesting a role of reproductive hormones in the etiology of the disease. The highest rates are observed in Europe and North America, whereas the lowest mainly in Asia (China and Japan in particular); studies have been carried out which indicates that international differences in breast cancer rates are due, at least in part, to environmental and lifestyle differences.

Moreover, a steady increase in breast cancer rates during the last decades (with a peak in the 80s) has been observed, the reasons being not completely clear, but likely attributable also to changes in reproductive patterns, increasing obesity in postmenopausal women, use of post-menopausal hormones and improvements in tumor detection for small sized (< 2 cm) cancers. Together with this came a decrease in cancer mortality, for which the growing use of screening mammography (leading to earlier detection of preexisting cancers) most certainly played a major role. Earlier causes do also exist, but we will focus on post-menopausal breast cancer only.

2.2 The Role of Hormones and HRT

Evidence has been found that sex hormones play a central role in the etiology of breast cancer. Several reproductive factors are consistently associated with breast cancer risk (for example age at menarche, age at first birth and parity, brest feeding, spontaneous and induced abortions and age at menopause) and higher levels of endogenous estrogen are reported to be associated with a higher risk. The effect of hormones intake has also been studied in detail, contraceptives and postmenopausal hormone use being the principal sources.

Improved survival among females with breast cancer has meant that more women are going through menopause, which for some women can cause severe symptoms such as vaginal atrophy, skin drying, hot flashes, night sweats, and loss of sexual desire. For many years, HRT (Hormon Replacement Therapy, usually a combination of the hormones estrogen and progestin) was widely prescribed to women to relieve these menopausal symptoms. It was also thought that HRT might reduce the risk of breast cancer, heart disease, and other conditions.

However, since more than half of breast cancers are fueled by estrogen, several trials were started in the 1990s in order to evaluate the potential risk

of breast cancer relapse in women using HRT.
A notable case was the HABITS trial that, although designed to have a follow-up time of five years, was stopped after only two years: recurrent or de novo breast cancer had developed in a number of women in the HRT group and in some in the non-HRT group. All women with a breast cancer event in the HRT group and two of those assigned to the no-HRT group had received HRT, and most had their tumor event while receiving treatment.

In July 2002, a large randomized clinical trial of estrogen and progestin in healthy postmenopausal women (part of the Women's Health Initiative) was stopped early when researchers found that women who took the hormones had an increased risk of developing breast cancer and heart disease.
The U.S. Food and Drug Administration has since recommended that women discuss with their doctors whether the benefits of taking estrogen and progestin outweigh the risks and that, if used, the hormones should be taken "at the lowest doses for the shortest duration to reach treatment goals."

2.3 HRT and Prognosis

Several studies have reported that women who undergo hormonal therapy and develop breast cancer tend to have a better prognosis than women with breast cancer who do not.
One reason for this could be that HRT-induced cancers are less aggressive than those caused by other factors, although this is debated [1].

This topic poses an interesting (and difficult) task: quantifying differences in prognostic factors across subtypes of breast cancer, a problem which turns out to be of non trivial solution.
In the first place there are identifiability issues: due to the fact that tumors which occur among women who are treated with HRT are a mixture of HRT-induced and other tumors, the sub-types are not identifiable without additional information or without making additional assumptions. To rule this out, a framework based on principal stratification has been proposed [10], which can lead to identifiable population's *strata* reflecting the aforementioned sub-types.
In the second place a "naïve" methodology (i.e. classic conditional linear or logistic regressions) is likely to incur in bias problems with respect to the estimand of interest, i.e. the direct effect of HRT on prognosis or survival. We will show how to approximately reach the situation where no such bias occur, thanks to the principal stratification framework.

Summarizing:

- Quantifying differences in prognosis and being able to assess to which factors such differences may be due is difficult, handling this problem is the topic of this thesis.
- The "naïve" approach consider cases only and regress prognosis on HRT duration, i.e. completely ignores heterogeneity among the subjects.
- We employ principal stratification (a technique developed in causal inference studies) to tackle the identifiability problem.

Before going further with the problem of estimating the HRT-prognosis relationship, we first review principal stratification in Chapter 3.

Chapter 3

Causal Inference and Principal Stratification

3.1 Review

In 2002 Frangakis and Rubin (FR) [2] proposed a new framework to deal with the problem of how to compare treatments effects adjusting for a post-treatment variable, which is known to be inducing the so called "posttreatment selection bias". Avoiding such bias is important if the focus of the analysis is the estimation of the causal effect of the treatment on an outcome of interest; the authors present as a solution the creation of a stratification based upon the posttreatment variables, in such a way that effect estimates within these "strata" will alway have a causal interpretation. FR make use of the notions of "potential outcome" and "counterfactual", originally introduced by Neyman (1923) [5] and later extended by Rubin in the '70s [7], [8] and [9].
Another remarkable work that is of interest to us is the 2003 article by Gilbert, Bosch and Hudgens (GBH) [3], in which the authors address the problem of evaluating the impact of vaccination on HIV viral load and other surrogate endpoint measures of infection; observing that a standard test that compares the distribution of viral load between the infected subgroups of vaccine and placebo recipients does not assess a causal effect of vaccine (because the comparison groups are selected after randomization), they make use of FR principal stratification method to obtain causal estimands. They introduce, moreover, a class of logistic selection bias models in order to identify the estimands in a correct way.
We quickly review the basic causal inference concepts and then summarize FR and GBH's articles.

3.2 Potential Outcome and Counterfactuals

Imagine that we have information on a binary (for simplicity) exposure X and an outcome of interest Y; for example X is assignment to either treatment (1) or placebo (0) and Y is survival after 5 years (1 alive, 0 deceased). The definition of a causal effect of X on Y for a certain subject requires a comparison between the outcome for that individual if treated and the outcome if not treated.
We use the notation $Y(x)$ to indicate the potential value that the outcome Y would assume if X were forced to x, so $Y(1)$ is the survival $(0, 1)$ we would observe were the subject treated and $Y(0)$ the survival were the subject assigned to placebo. These quantities are called *potential outcomes* and we obviously cannot simultaneously observe both $Y(1)$ and $Y(0)$ for the same individual, what we observe is the *factual* realization of the variable Y, i.e. $Y(1)$ if the subject has been treated and $Y(0)$ if not; the complementary potential outcome is then called *counterfactual*. Individual causal effects are not computable except under extremely strong (and often unreasonable) assumptions, because we cannot observe the same subject under both exposure levels.

If we want to analyze a population of individuals, rather than only one, the same reasoning applies: the definition of a population causal effect calls for a comparison between the whole population under exposure and the whole population under non-exposure. But as for separate individuals, those assigned to treatment yeld a factual outcome $Y(1)$ and for them we do not observe $Y(0)$, whereas those assigned to placebo yeld the value $Y(0)$ and we do not observe their $Y(1)$.
A tentative approach could be to use association as a surrogate for causation, i.e. using the group of those who have been factually exposed as a surrogate for the hypothetical population "had everybody been exposed" and thus considering $\Pr(Y = 1|X = 1)$ as a proxy for $\Pr[Y(1) = 1]$; similarly for the "had nobody been exposed" hypothetical population we may use the data on those subjects with $X = 0$, i.e. $\Pr(Y = 1|X = 0)$ for $\Pr[Y(0) = 1]$. Comparisons on such distributions does not, in general, yeld a causal effect (although randomization to treatment could simplify estimation), due to the problem of *non-exchangeability* (see [6] for details); a formal definition of such a concept may be given using potential outcomes, assessing that "exposed and unexposed are exchangeable if $Y(1)$ and $Y(0)$ are jointly independent of X", which implies the following equalities:

$$\begin{aligned} &\Pr[Y(1) = 1|X = 0] = \Pr[Y(1) = 1|X = 1] = \Pr[Y(1) = 1] \\ &\Pr[Y(0) = 1|X = 1] = \Pr[Y(0) = 1|X = 0] = \Pr[Y(0) = 1] \end{aligned} \tag{3.1}$$

or, in short, $[Y(0), Y(1)] \amalg X$. This means that if exchangeability holds, contrasting, for example, $E(Y|X = 1)$ and $E(Y|X = 0)$ yelds the same results as contrasting $Y(1)$ and $Y(0)$, which is precisely what required to obtain a causal effect. It is important to note that observed data can never tell us whether exposed and unexposed are exchangeable or not, in order to judge the plausibility of such assumption we have to rely on subject matter knowledge.

There are different reasons why the exchangeability condition may not hold, the most important of which is the presence of a third factor which affects both X and Y. If so then there will be an association between X and Y *even if* X *has no causal effect on* Y, we call this common cause a *confounder* and show in this work a way to properly take into account such variables in order to avoid the bias they induce in an analysis.

3.3 Principal Stratification - Frangakis and Rubin

A crude analysis such as contrasting the two groups of treated and not treated does only, in general, yeld an association measure because we are comparing different sets of individuals, while the definition of causal effect requires a comparison to be made on a common set (i.e. the same individuals under both treatment and no treatment). FR propose to structure the analysis focusing on subgroups of units defined by measured pretreatment variables values, thus creating a stratification such that a comparison within each of this subgroups yeld a causal effect.
Let us consider a binary treatment $X = 0, 1$, an outcome Y and a post-treatment variable S; an example could be the HIV vaccine trial, where X is assignment to either vaccine or placebo, Y survival and S is compliance to the trial.

Definition

- *The basic principal stratification* P_0 *with respect to posttreatment variable* S *is the partition of units* $i = 1, ..., n$ *such that, within any set of* P_0*, all units have the same vector* $[S_i(0), S_i(1)]$.
- *A principal stratification* P *with respect to posttreatment variable* S *is a partition of the units whose sets are unions of sets in the basic principal stratification* P_0.

For clarity, a possible principle stratification P is the partition of individuals into the set whose posttreatment variable is unaffected by treatment in

this study (i.e. with $S_i(0) = S_i(1)$) and the remaining subjects (i.e. with $S_i(0) \neq S_i(1)$); given the HIV trial example, the first set would be formed by the "always compliers and never compliers regardless of treatment", and the second would include those whose decision to comply or not is influenced by the treatment they have been assigned to (here we assume the trial is not blinded).
FR observe that, generally, it is impossible to directly observe the principal stratum to which a subject belongs, because we can observe only either $S_i(0)$ or $S_i(1)$ for each individual, nonetheless they proceed by assuming the strata belongings as known to present their method.

Definition *Let P be a principal stratification with respect to the posttreatment variable S and let S_i^P indicate the stratum of P to which unit i belongs. Then a principal effect with respect to that principal stratification is defined as a comparison of potential outcomes under $X = 0$ versus $X = 1$ (in this case placebo vs vaccine) within a principal stratum ς in P, i.e. a comparison between the ordered sets*

$$\{Y_i(0) : S_i^P = \varsigma\} \qquad \textit{and} \qquad \{Y_i(1) : S_i^P = \varsigma\}. \tag{3.2}$$

The key is in the fact that the value of the ordered pair $[S_i(0), S_i(1)]$ is, by definition, not affected by treatment (although the potential variable $S_i(0)$ generally differs from $S_i(1)$). Therefore the following properties are valid:

1. *The stratum S_i^P, to which unit i belongs, is unaffected by treatment for any principal stratification P.*

2. *Any principal effect, as defined in* (3.2)*, is a causal effect.*

Which is to say, if memberships to strata were known, stratification of the individuals by S_i^P would adjust for personal characteristics reflected in the posttreatment variable without inducing treatment selection bias for any principal stratification P.

As stated already, belonging to a stratum is something which is in general not directly observable, nevertheless FR remark that it is often possible to build plausible restrictions for such an assignment, for example using covariates to predict each individual's membership, and using sensitivity analysis techniques for the causal effects, exploring possible ranges of unobserved quantities.

3.4 Principal Stratification - Gilbert, Bosch and Hudgens

GBH develop a method, based on FR's principal stratification, which aims at obtaining a causal effect of HIV vaccine on viral load that adjusts for the posttreatment selection bias. They define the estimand *causal vaccine effect* on viral load as a comparison of potential viral loads under the two randomization assignments (either vaccine or placebo) for a subgroup of subject with a common pair of potential infection status outcomes (i.e. a *principal stratum* in the sense of FR), in particular they consider those subjects in the "always infected" stratum, i.e. those who would be infected regardless of randomization to vaccine or placebo. The authors state that drawing inference on such stratum addresses a a practical question for individuals vaccinated in a public health program: "If I acquire HIV despite vaccination, what is the viral load compared to if I had forgone vaccination?".

The causal estimands are not identified, because membership of an infected placebo recipient to the "always-infected" principal stratum is unknown (we do not observe the counterfactual variable "infection status had the subject been randomized to vaccine"). To address this problem GHB make use of models for the probability that an infected placebo recipient is in the "always-infected" stratum as a function of the potential viral load under randomization to placebo. Earlier works (see for example [4]) implicitly took this approach, by defining two selection models that express bounds for the maximum plausible levels of selection bias and allow for identification of the estimands. GBH remark that it is important to also consider selection models that reflect intermediate degrees of selection bias, which may be more realistic and will allow for more powerful statistical tests, and develop a method for sensitivity analysis that explores a continuous range of possible selective effects betweem the two extreme situations of no bias and maximal plausible bias that [4] consider.

Under Rubin's [9] *SUTV* (stable unit treatment value) and FR [2] assumptions, GBH derive a testing procedure that contrasts the distribution $F_{(v)}^{alw.inf}(y)$ (potential viral load for the "always-infected" had all of them been randomized to vaccine) with $F_{(p)}^{alw.inf}(y)$ (the same distribution had all of them been randomized to placebo), which always yelds a causal effect. This way, the null hypothesis of no causal effect of vaccination on viral load in the "always-infected" principal stratum can be expressed as:

$$H_0 : F_{(v)}^{alw.inf}(y) = F_{(p)}^{alw.inf}(y), \forall y. \tag{3.3}$$

Neither of the two distributions in (3.3) is identifiable, hence requiring GBH to make two further assumptions

1. treatment assignment of each subject is independent of his/her potential outcome
2. no subject would be simultaneously infected if randomized to vaccine or uninfected if randomized to placebo

which they find justifiable due to the nature of the trial (randomized and blind) and to subject matter considerations.

Let $S_i(\cdot)$ denote the potential outcome for subject i of the variable indicating whether the subject would be infected ($S_i(\cdot) = 1$) or not ($S_i(\cdot) = 0$) given treatment $\cdot$ (either vaccine (v) or placebo (p)). Consider moreover a partition of the whole population into three principal strata:

- *always-infected*, for which $\{S_i(v) = S_i(p) = 1\}$
- *never-infected*, for which $\{S_i(v) = S_i(p) = 0\}$
- *protected*, for which $\{S_i(v) = 0, S_i(p) = 1\}$.

A fourth possible stratum would be formed by subjects having the pair $\{S_i(v) = 1, S_i(p) = 0\}$, but this set is, by assumption **2**, empty. A subject with $S(p) = 1$ may then belong to either the always-infected or the protected stratum, thus making such distinction unidentifiable. GBH define the level of *vaccine efficacy* (VE) against infection, determining the proportion of subject with $S(p) = 1$ in each of such strata, as

$$VE = 1 - RR = 1 - \frac{\Pr\{S_i(v) = 1\}}{\Pr\{S_i(p) = 1\}}, \tag{3.4}$$

and remark that it is a causal estimand measuring the relative reduction in infection risk conferred by randomizing to vaccine versus placebo. Under all the assumptions the authors write then the density $f_{(p)}(y)$ of potential viral load in subject infected under randomization to placebo as a mixture of the densities for the protected ($f_{(p)}^{prot}(y)$) and the always-infected ($f_{(p)}^{alw.inf}(y)$) strata as follows:

$$f_{(p)}(y) = VE * f_{(p)}^{prot}(y) + (1 - VE) * f_{(p)}^{alw.inf}(y). \tag{3.5}$$

With some calculations, the density in (3.5) can be rewritten as a biased sampling model as follows:

$$f_{(p)}^{alw.inf}(y) = W^{-1} w(y) f_{(p)}(y), \tag{3.6}$$

where $w(y) = \Pr\{S_i(v) = 1, Y_i(p) = y, S_i(p) = 1\}$ and $W = \int_{-\infty}^{\infty} w(y) f_{(p)}(y) dy$ is a normalizing constant equal to $1 - VE = RR$. The weight function $w(y) = RR(y) = 1 - VE(y)$ is the probability that a subject infected with viral load y if randomized to placebo would be infected if randomized to vaccine. GBH show that it follows that testing (3.3) is equivalent to testing

$$H_0 : F_v(y) = (1 - VE)^{-1} \int_{-\infty}^{y} w(z) dF_p(z), \forall y. \tag{3.7}$$

By assumption **1**, VE is identified from the observed data, so if $w(\cdot)$ were known, then both $F_{(v)}^{alw.inf}(\cdot)$ and $F_{(p)}^{alw.inf}(\cdot)$ would be identified, and the hypothesis (3.3) could be tested. Being $w(\cdot)$ unknown, and being not possible to test whether the chosen $w(\cdot)$ is correctly specified, the authors assume it as known, and test (3.7) for a variety of fixed choices of $w(\cdot)$. In particular they choose a logistic model, indexed by an interpretable bias selection parameter β; for a finite β, e^{β} is the odds ratio of infection under randomization to placebo with viral load y versus viral load $y - 1$. GBH remark the importance of choosing an interpretable model for $w(\cdot)$, so to be guided by beliefs about plausible degrees of selection bias in the choice of β.
The authors conclude deriving testing procedures which would allow to adjust for selection bias, if the model for $w(\cdot)$ has been correctly specified, and presenting a simulation study in which they investigate the power of different tests for a variety of correct and incorrect choices of the β parameter, i.e. what and how much can go wrong if the selection bias model is improper.

3.5 A Short Introduction to DAGs

As stated already, exchangeability may not hold; knowing this we may consider it more plausible if we adjust for some additional (set of) known variable(s) L in our analysis. Nonetheless, even conditional exchangeability remains untestable (we may have unmeasured/unadjusted confounders) and must be justified by subject matter knowledge. It is possible to obtain

exchangeability by adjusting, but it is also possible to "destroy" it, therefore it is of primary importance to epidemiologists involved in observational studies to determine which the correct set of variable to adjust for is. An important tool that can be useful in such a process are the *DAG*s, namely *Directed Acyclic Graph*s.

In mathematics, a *graph* is an abstract representation of a set of objects where some pairs of objects are connected by links. The interconnected objects are represented by mathematical abstractions called vertices, and the links that connect some pairs of vertices are called edges. Typically, a graph is depicted in diagrammatic form as a set of dots for the vertices, joined by lines or curves for the edges. In causal inference (and in statistical modelling in general) the vertices are variables, and the edges represent causal links between pairs of such variables. It is worth noting that formal relationships exist with counterfactuals through non-parametric structural equations, even if this is beyond the subjects of this work.

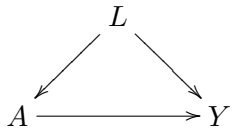

Figure 3.1: A Simple DAG With 3 Variables

Each of the arrow in Fig.3.1 represents a causal effect: A causes Y, L causes both A and Y. This graph is *directed* since each connection between two variables is an arrow, and *acyclic*, because it contains no *directed cycles*, i.e. following the arrows it is not possible to come back to the starting point, whichever it was.
Direction of arrows has a precise meaning, giving the direction of the (causal) link from one variable to another: in Fig.3.1 for example, A affects Y but *not the other way around*. Presence of arrows encodes *assumptions*, if an arrow from, say, L to A is present then we believe L may or may not affect A, while if the same arrow is absent then we believe that L *does not affect* A.

A set of graphical rules allows us to assess whether two variables on a graph are independent (*d-separated* in graph terminology) or not, and this will help us determine what variables we need to adjust for; we first introduce some notation.

- The **ancestors** of a variable, V, are all other variables which affects V either directly or indirectly. In the DAG in Fig.3.1 A has one ancestor, L, which in turn has no ancestors.

- The **descendants** of a variable, V, are all other variables affected by V, either directly or indirectly. A has a single descendant Y, L has two descendants, A and Y.
- A **path** is a route between two variables passing through the arrows (not necessarily following the direction of them). For example, there are two possible paths from A to Y in the DAG above: $A \rightarrow Y$ and $A \leftarrow L \rightarrow Y$.
- A path can be either **blocked** or **open** according to two very simple rules.

I A path is **blocked** if it contains a "chain" $\rightarrow L \rightarrow$ or a "fork" $\leftarrow L \rightarrow$ *and* we have conditioned on the middle variable.

II A path is **blocked** if it contains an "inverted fork" $\rightarrow L \leftarrow$ *and* we have *not* conditioned on the middle variable *or on any of its descendants*

else, a path is *open*; the middle variable in an inverted fork is usually called a *collider*. If all paths between, say, A and Y are blocked by conditioning on a (set of) variable(s) L, then we say that A and Y are *conditionally independent* given L or, using graph terminology, that A and Y are **d-separated** by L. If at least one path is open, then A and Y are (most likely) conditionally associated. Note that L may be the empty set, in this case then A and Y are unconditionally independent.

Consider again the DAG in Fig.3.1 and assume it describes the true causal structure. Suppose we desire to test the existence of the arrow from A to Y, i.e. the existence of a causal effect of A on Y, should we adjust for L in our analysis or not? The first step is to draw the graph under the *null hypothesis* of no causal effect, that is, we redraw it deleting the arrow from A to Y.

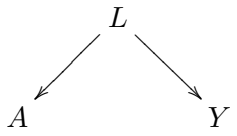

Figure 3.2: Simple DAG Under the Null Hypothesis of no Causal Effect

Making use of the graphical rule **I** we verify that conditioning on L **blocks** all possible paths from A to Y. Hence, any evidence of an association between A and Y given L would prove that **there is** a causal effect of A on Y, i.e. that the arrow we have deleted is actually there. In this simplified (and unrealistic) situation exposed and unexposed are conditionally

exchangeable, given L, and we can not only test the presence of a causal effect, but also estimate it using an observed (conditional) association, for example employing a simple linear or logistic regression model (clearly depending on the nature of the variables and the estimand of interest). Now consider a different situation, depicted in Fig.3.3.

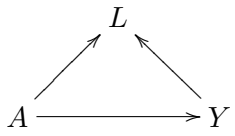

Figure 3.3: Another DAG Axample

Here L is jointly caused by A and Y, suppose, again, that this describes the true causal structure; we draw the same graph under the null hypothesis of no causation of A on Y and apply the graphical rules to verify if and when A and Y are *d-separated*.

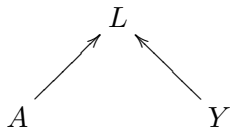

Figure 3.4: Another DAG Axample, Null Hypothesis

It is immediate to note that, by rule **II**, conditioning on the collider L would **open** the path $A \rightarrow L \leftarrow Y$. Evidence of an association between A and Y conditional on L would then **not** prove the existence of a causal effect of A on Y, and an estimate obtained via a simple conditional analysis would be biased. The only way to obtain exchangeability, and then being able to both test the existence of a causal effect and estimate its magnitude, would here be **not** to condition on L; this way the path $A \rightarrow L \leftarrow Y$ would be, by rule **II**, blocked.

Chapter 4

A P.S. Approach to HRT and Breast Cancer

4.1 Doomed vs Healthy

Consider, for simplicity, a scenario where each woman is classified either as treated or untreated (thus ignoring, for the time being, the possible duration of the estrogen/progestin therapy); suppose that we can think of women as belonging to one of three categories:

- those who do not develop cancer in any case, regardless of whether they are treated or not, we call them "healthy"
- those who develop cancer if treated, but would not have had they not been treated, we call them "sensitive"
- those who develop cancer in any case, regardless of treatment status, we call them "doomed"

According to the literature we exclude the possibility of a fourth strata, i.e. women that would only develop cancer if not treated and would not if treated. This classification can be easily summarized in a table for clarity:

cancer	healthy	sensitive	doomed
untreated	no	no	yes
treated	no	yes	yes

Table 4.1: Classification into "healthy", "sensitive" or "doomed"

Upon development of breast cancer we find ourselves in a difficult situation: has the woman been treated with HRT? if the answer is no, then there is no doubt (in this simplified scenario) that she belongs to the "doomed" group. But what if the woman has been treated? we find ourselves in an unclear

situation, since she could belong either to the "doomed" or to the "sensitive" group; if she is "sensitive" then the tumor is HRT-induced (i.e. caused by the hormonal therapy), if she is "doomed" then the tumor could be either HRT-induced or have other causes. We are therefore not able to distinguish between HRT-induced and other subtypes of cancer, since we cannot "assign" a woman who developed breast cancer to one of the two groups.

If we want to consider (and eventually maybe quantify) an association between HRT and breast cancer prognosis, we need then to compare the prognosis for untreated women with tumors not caused by HRT (doomed) against a mixture of treated women with HRT-induced tumors (sensitive) and not caused by HRT tumors (doomed). This is not possible unless we make some assumptions in order to overcome the unidentifiability issue we mentioned; the framework of potential outcomes and principal stratification offers a way to address this problem and set up a model which is suitable for estimating the quantities of interest [10].

4.2 Assumptions and Estimands of Interest

Methods have been developed to solve the issue of unidentifiability, relying on the assumptions typical to the framework of potential outcomes. Two of these methods are presented in [10] as an application of the principal stratification techniques to the HRT problem. A scenario where HRT is defined on a continuous scale (previous literature focused mainly on binary covariates) and with additional information on covariates is considered, the aim being to estimate, under a set of reasonable assumptions, the difference in prognosis for women with different tumor subtype.

Following [10] and the potential outcomes notation in FR [2]:

- Z denotes disease status, $Z = 1$ for cancer and $Z = 0$ for no cancer
- X denotes duration of past HRT use, we assume X continuous in $[0, \infty)$, where $X = 0$ indicates no past HRT use
- C denotes baseline covariates, allowed to be measured on any mixture of scales
- Y is a prognostic factor, allowed to be measured on any scale, and for convenience defined as 'not defined' when $Z = 0$
- $Z(x)$ and $Y(x)$ denote the potential outcomes of Z and Y at HRT level $X = x$

- $Z(\cdot)$ denotes the entire potential outcome function $\{Z(x), \forall x\}$, and we say that two women belong to the same principal stratum if they have the same $Z(\cdot)$
- we call *non HRT-induced* a breast cancer which is not caused by HRT

[10] make the following assumptions:

I $Z(X) = Z; Y(X) = Y$
This *consistency* assumption states that the potential outcomes corresponding to the factual (observed) HRT level are equal to the observed outcomes Z and Y; those corresponding to other (counterfactual) levels are unobserved, thus considered as missing.

II $Z(x) \geq Z(x')$ if $x \geq x'$
HRT could hypothetically prevent breast cancer for some women, but this is highly unlikely. Assumption II states therefore that a woman who develops cancer at one level of HRT, would also have developed it had she been treated longer; $Z(\cdot)$ is then a monotonically increasing step function, thus (given the dichotomous nature of Z and II) straightforwardly summarizable into a scalar R, defined as the minimum level of x for which $Z(x) = 1$ (i.e. the cancer occurs).

III $\{Y(x), R\} \amalg X|C, \forall x$
This is a *conditional independence* statement, and assumes that X can be considered randomized within levels of C.

For this analysis, we consider the principal strata as representing women with different types of cancer, i.e. we distinguish between HRT-induced and non HRT-induced cancers. We can 'label' these strata using the scalar R (which summarizes the potential outcome function $Z(\cdot)$) as follows:

cancer subtype	$R = 0$	$R > 0$
$X = 0$	non HRT-induced	no cancer
$X > 0$	HRT-induced/non HRT-induced	HRT-induced

Table 4.2: Cancer subtypes strata

R could be seen as some sort of 'resistance' each woman may have against breast cancer, keeping into account HRT. $R = 0$ would mean 'no resistance' at all, whereas $R > 0$ would mean not to develop the tumor until a certain duration of the hormonal therapy has been reached. Again we can see

how there is an 'overlapping' of tumor subtypes among treated women: if $(R = 0, X > 0)$ then the tumor could either be caused by HRT or by other factors. [10] makes the additional assumption that since for most women (moderate levels of) HRT does not cause cancer, then most of the observed cancer within $R = 0$ should be bound to be non HRT-induced, thus reducing the partition into:

cancer subtype	$R = 0$	$R > 0$
$X = 0$	non HRT-induced	no cancer
$X > 0$	non HRT-induced	HRT-induced

Table 4.3: Cancer subtypes strata with the non HRT-induced assumption

or, similarly:

- $Z = 1, R > 0 \rightarrow$ HRT-induced tumor
- $Z = 1, R = 0 \rightarrow$ non HRT-induced tumor.

The target estimand of interest is the difference in prognosis between women with HRT-induced tumors and women with non HRT-induced tumors, which [10] formalizes as

$$m(x, C) := g[E\{Y(x)|0 < R \leq x, C\}] - g[E\{Y(x)|R = 0, C\}], x > 0 \quad (4.1)$$

where $g(\cdot)$ is a known, smooth, monotone link function, and $m(0, C) := 0$. Note that $m(x, C)$ is not a measure of the HRT effect on prognosis (but this does not mean it cannot be influenced by an existing HRT effect). One last assumption is made in this general setting:

IV $Pr\{Y(x)|R = 0, C\} = Pr\{Y(0)|R = 0, C\}, \forall x$
If a woman develops an HRT-induced cancer, then it's likely that her prognosis depends to some extent on the duration of the therapy (for example, a larger dosage could cause a more aggressive cancer). If a woman, instead, develops a non HRT-induced cancer, then there is no reason to believe that her prognosis can be affected by her dosage (duration) of HRT, which is what this assumption states (for a justification of this see [10]).

4.3 Identification and Estimation

Based on assumptions I-IV [10] develops a structure which permits to model, through a mixture distribution, the prognostic factor Y both in the cohort and in the case control settings (for the case control studies one more assumption will be needed).

Cohort study
A cohort study generates an *i.i.d.* sample from $Pr(Y, Z, X, C)$.
Define $\pi(X, C) := Pr(Z = 1|X, C)$. Under I-III it can be shown that

$$Pr(R \leq x|C) = \pi(x, C) \tag{4.2}$$

In particular $Pr(R = 0|C) = \pi(0, C)$ holds. Equation (4.2) implies that $Pr(R|C)$ (i.e. the conditional distribution of the principal strata) is identified under cohort sampling. Moreover it is possible to write the conditional distribution of Y, given $(Z = 1, X, C)$ as a mixture of potential outcomes for women with HRT-induced and non HRT-induced cancers. Let $\pi_H = \frac{\pi(0,C)}{\pi(x,C)}$, then:

$$\begin{aligned} \Pr(Y|Z = 1, X = x, C) =& \pi_H * Pr\{Y(x)|R = 0, C\} + \\ &+(1 - \pi_H) * Pr\{Y(x)|0 \leq R \leq x, C\} \end{aligned} \tag{4.3}$$

Assumption IV allows identifiability of the mixture components. Combining it with (4.3) and (4.1) yelds:

$$\begin{aligned} m(x, C) =& g\left\{\frac{E[Y|Z = 1, X = x, C)] - E[Y|Z = 1, X = 0, C)] * \pi_H}{1 - \pi_H}\right\} + \\ &-g\{E[Y|Z = 1, X = 0, C)]\} \end{aligned} \tag{4.4}$$

Equation (4.4) shows that $m(x, C)$ is identified if $\pi_H \neq 1$ or, equivalently, if $\pi(x, C) \neq \pi(0, C)$.

Case control study
A case control study generates two samples: one *i.i.d* sample of size n_1 from

$Pr(Y, X, C|Z = 1)$ and one *i.i.d.* sample of size n_0 from $Pr(Y, X, C|Z = 0)$. Assumptions I-IV alone do not guarantee identifiability in the case control setting, since under this sampling scheme $\pi(X, C)$ is not identified (hence neither is $P(R|C)$). We need then to introduce:

V $\pi(x, c) \simeq 0, \forall x, c$
this is the 'rare disease' assumption, which is often reasonable as it forms the basis for choosing case controls designs in practice. Since case control studies don't follow patients over time, a relative risk can't be evaluated; it is possible, however, to calculate the exposure-odds ratio, which approximates the RR when prevalence is close to zero.

Define the odds ratio

$$\eta(X, C) := \frac{\pi(X, C)[1 - \pi(0, C)]}{\pi(0, C)[1 - \pi(X, C)]} \tag{4.5}$$

which by Bayes rule is trivially identified from case-control sampling, and equal to $\frac{Pr(X|Z=1,C)Pr(X=0|Z=0,C)}{Pr(X|Z=0,C)Pr(X=0|Z=1,C)}$.
Given V (i.e. when $\pi(x, c) \simeq 0$), it is easy to show that $\pi_H \simeq \eta^{-1}(X, C)$. This way we can obtain an 'approximate' identification in this setting as well.

Two approaches are proposed in [10] in order to estimate $m(x, C)$:

- The **implicit method** specifies models for $E(Y|Z = 1, X, C)$ and $\pi(X, C)$ ($\eta^{-1}(X, C)$ for the case control study), and use the relation in (4.4) to obtain a model for $m(x, C)$. A quite straightforward choiche can be to model the mean level with a linear relationship and the $\pi(X, C)$ ($\eta^{-1}(X, C)$) with a simple logistic regression, both employing the additional covariates C as regressors. The implicit model turns out to be computationally light, but could be quite hard to interpret in terms of the involved parameters, which may make it difficult to formulate scientifically relevant questions about $m(x, C)$. Standard errors for $m(x, C)$ can be obtained from the delta method.

- The **explicit method** models $m(x, C)$ directly (and works when the $g(\cdot)$ function is the identity link or the log-link) and involves fitting one main model, for example $m(x, C) = \psi_0 + \psi_0 x$, together with other nuisance sub-models involving covariates and the mixture weights π_H. This procedure makes use of semi-parametric estimating techniques

(namely the GEE, Generalized Estimating Equations), yelding estimators which, under regularity conditions, have good asymptotical properties and possess an analytical expression for their variance-covariance matrix. One drawback is some 'rigidity' this model implies, not being possible to extend it to handle more than two strata (which could be of interest in other settings).

Appropriateness of assumptions I-III can be tested to a certain extent, thanks to restrictions which are directly obtainable from such statements (see [10]).

4.4 Different approaches comparison

Some remarks can be made about what, in this particular setting, could drive the choiche of using a semi-parametric approach rather than a full parametric one for estimation. Our intent is to propose a different way of modelling (call it **B**) parametrizing unobservable quantities (i.e. potential outcomes) as opposed to the procedure [10] developed (call it **A**).

A		**B**
$\surd$	**Testing**	$\surd$
$\surd$	**Computational Weight**	$\times$
$\times$	**Flexibility**	$\surd$
$\times$	**Understanding**	$\surd$

Table 4.4: **A** vs **B** Estimation Methods

- Testing: both methods allow a quite straightforward parameters testing
- Computational Weight: the **B** method could turn out to be computationally heavier than the **A**, depending on our choiche of distributional assumptions
- Flexibility: **B** allows a good flexibility without impacting too much neither on the estimation procedure complexity nor restricting the choiche of the functional form of the estimand of interest
- Understanding: semi-parametric procedures can be quite difficult to be fully understood and implemented for non-statisticians, whereas full parametric ones may be more intuitive to many.

Chapter 5

A Modeling Proposal

5.1 Focusing on Causality

The models reviewed in 4.3 deal with the question "how can we assess the differences in prognosis between women with cancers caused by HRT and women with cancers caused by other factors?" and try to quantify such differences through the estimand in equation (4.1). If we are interested in inferring into causality, then we need a different approach and two important steps are to be considered:

- assessment of the presence (or absence) of a causal link between variables (here in particular from X to Y)
- estimation of the magnitude of such connection.

The first step can be achieved with a simple analysis after some reasoning which can be supported by the use of DAGs, whereas the second will require more assumptions to be feasible.

Let the following DAG represent the simplest possible situation, still keeping in mind that in a realistic setting the presence of unmeasured confounders U is unavoidable:

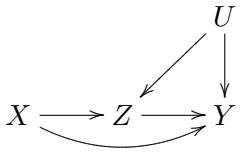

Figure 5.1: Simplest Possible Structure with Unmeasured Confounders

Fig.5.1 graphically represents what follows: X may affect Y directly or through Z, U summarizes supposed unmeasured confounders which affect

both Z and Y but not X. If we think of X as HRT, Z as the disease status and Y as the prognosis then U could represent some sort of personal characteristics, possibly genetic traits, that are likely to influence the onset of the tumor and the prognosis, but do not have any impact on the physician's decision to prescribe HRT or not to a particular woman.

Following what presented in Section 3.5, in order to assess the presence (or absence) of a causal link from X to Y we should first draw the graph under the null hypothesis of no effect of X on Y:

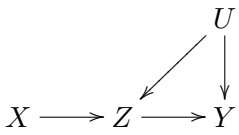

Figure 5.2: Null Hypotesis: No Effect of X on Y

If no unmeasured confounders were present, evidence of association between X and Y through a naïve analysis (regressing Y on X given Z) would both confirm the presence of the causal link $X \to Y$ *and* yeld an estimate of such effect (it would be equal to the regression coefficient for X). The reason for this being that, in this case, conditioning on Z would block every possible path from X to Y.
This is however, as said, not realistic, since we will always be in the presence of unmeasured confounding factors which may affect the involved variables opening paths ($X \to Z \leftarrow U \to Y$ in this case, Z being a collider) and thus inducing a component of heterogeneity which we would not be able to explain through the informations at our disposal, and eventually resulting in a biased (*confounded*) estimation of the true effects. In such a situation, then, a simple naïve model (a linear or a logistic regression, for example) is not indicated if our aim is to estimate a causal effect.

Let us now consider a more complex scenario, depicted in the DAG in Fig.5.3. R indicates the unknown "resistance" to developing breast cancer with respect to HRT; if both R and X were known, then we would know whether a particular woman were to be classified as either "healthy", "doomed" or "sensitive", i.e. which *stratum* she belongs to, thus uniquely determining that woman's disease status Z and the cause of a cancer event ($Z = 1$).
Again, if we draw the DAG under the null hypotesis (i.e. deleting the arrow from X to Y), it is easy to see that conditioning on Z *and* R would block every possible path from X to Y, allowing us to assess the presence of a causal effect if we find evidence of association between such variables even

after having adjusted for the disease status and the strata belonging.

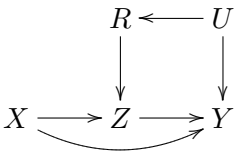

Figure 5.3: Strata Structure

If we could observe R then we may still want to use a simple linear or logistic regression to perform the analysis and include strata belonging as a covariate. Unluckily R is an *unobservable* variable, thus requiring some more sofisticated modelling; our choice has been to build a selection model which could cluster women who developed breast cancer (i.e. cases) into two groups: "doomed" and "sensitive". This will be presented in detail in Section 5.2. The rationale behind this choice is that if we build a good enough selection model, then we may think we have *approximately* correctly placed each woman into the right subgroup, thus obtaining an approximately correct adjustment for R, which in turn could allow to consider the $X \to Z \leftarrow R \leftarrow U \to Y$ path as blocked.

Under assumptions I-IV (Section 4.2) and assuming that women with $R = 0$ experience non HRT-induced cancers only, the following DAG (Fig.5.4) is mathematically redundant with respect to the one in Fig.5.3, but may be more useful to explain the biology which we believe drives the situation we are exploring. The tumor subtypes (TS) (i.e. no tumor, HRT-induced tumor or non HRT-induced tumor) appears then explicitely in the graph; in this interpretation Z becomes a sort of proxy for TS (which actually incorporates the information on disease status together with the tumor subtype distinction).

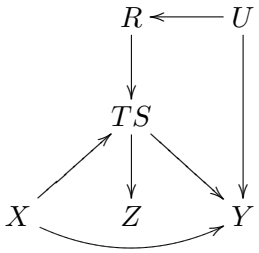

Figure 5.4: Tumor Subtypes Structure

5.2 Model Formulation in a Case-Control Study Design

Our aim is to directly parametrize unobservable variables, making a normality assumption on the potential outcomes of Y distribution in the two groups of cases "doomed" ($Z = 1, R = 0$) and "sensitive" ($Z = 1, R > 0$); this is the main difference of our approach as compared to the one proposed by [10]. The normality assumption we are going to make is supported mainly by exploratory data analysis in real datasets from the CAHRES study: the prognostic index has an approximately normal distribution in subpopulations which may be considered somewhat representative of the "doomed" and the "sensitive" groups.

Let $K = K_0 \cup K_1$ be the set of all individuals, formed by the disjoint union of those who are cases (K_1) and those who are not (K_0). Suppose that our sample is the entire population.

Then the joint likelihood (now for simplicity without any covariate) is:

$$L = \prod_{K_1} \Pr(Y|Z = 1, X) \prod_{K} \Pr(Z|X) \tag{5.1}$$

In a more realistic setting, where we do not observe the entire population but only a sub-group of it, we must keep into account the way this sample has been drawn from it. This means we need to know the so called *ascertainment* probabilities, i.e. the probability with which each person entered the sample. Our approach will assume known ascertainment probabilities.

This results in a further step in modeling $\Pr(Z|X)$, which should now be written as $\Pr(Z|X, A = 1)$, where A is a dichotomous variable whose value 1 indicates that a particular subject from the population was chosen to belong to the sample. It is easy to show that:

$$\Pr(Z|X, A = 1) = \frac{P(Z|X)P(A = 1|Z)}{\sum_{z=0,1} P(Z|X)P(A = 1|Z)}. \tag{5.2}$$

Suppose we know the proportions with which cases ($Z = 1$) and controls ($Z = 0$) have been randomly drawn out of a certain population, say p_1 for the cases and p_0 for the controls. Then we have that

$$P(A=1|Z) = \begin{cases} p_1 & Z=1 \\ p_0 & Z=0. \end{cases} \tag{5.3}$$

Substituting in (5.2) yelds:

$$\Pr(Z=z|X,A=1) = \frac{zp_1P(Z=1|X) + (1-z)p_0P(Z=0|X)}{p_1P(Z=1|X) + p_0P(Z=0|X)} \tag{5.4}$$

Let now $\bar{K} = \bar{K_0} \cup \bar{K_1}$ denote the sample, where $\bar{K_0}$ are the controls and $\bar{K_1}$ the cases, i.e. $\bar{K} = K \cap \{A=1\}$.

We assume a normal distribution for the outcome Y in both subgroups, allow for equal variance but model the mean level in a different way, i.e.

$$\begin{aligned} &Y|Z=1,X,R=0 \sim N(\mu_0,\sigma) \\ &Y|Z=1,X,R>0 \sim N(\beta_0+\beta_1X,\sigma). \end{aligned} \tag{5.5}$$

Such choice also reflects the assumption that HRT duration (X) does not affect breast cancer prognosis if a woman belongs to the "doomed" group.

Let $\pi_{\underline{\alpha},X} := \frac{expit(\alpha_0)}{expit(\alpha_0+\alpha_1X)}$, then the likelihood function results in:

$$\begin{aligned} \ln L = &\sum_{\bar{K_1}} \ln \left[\pi_{\underline{\alpha},x} \frac{e^{-\frac{(y-\mu_0)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} + (1-\pi_{\underline{\alpha},x}) \frac{e^{-\frac{(y-\beta_0-\beta_1x)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \right] + \\ &+ \sum_{\bar{K}} \ln \frac{zp_1expit(\alpha_0+\alpha_1x) + (1-z)p_0[1-expit(\alpha_0+\alpha_1x)]}{p_1expit(\alpha_0+\alpha_1x) + p_0[1-expit(\alpha_0+\alpha_1x)]}. \end{aligned} \tag{5.6}$$

This expression can be easily extended to the case when we have covariates if we assume that A is independent of such covariates, i.e. $P(A|Z,C) = P(A|Z)$ and equation (5.2) can be rewritten as

$$\Pr(Z|X, A = 1, C) = \frac{P(Z|X, C)P(A = 1|Z)}{\sum_{z=0,1} P(Z|X, C)P(A = 1|Z)}. \quad (5.7)$$

The full joint likelihood, including covariates can then be expressed as:

$$L = \prod_{\bar{K}_1} \Pr(Y|Z = 1, X, C) \prod_{\bar{K}} \frac{\Pr(Z|X, C) \Pr(A = 1|Z)}{\sum_{z=0,1} \Pr(Z|X, C) \Pr(A = 1|Z)}. \quad (5.8)$$

Equation (5.7) holds as long as the choice of including a subject in the sample is made completely at random or depends on factors that do not enter the model as covariates; if this were not the case, then we would have to consider such factors and build a more complicated model for $P(A = 1|Z, C)$.

5.3 Remarks

There is an important remark that must be done: the estimated HRT effect β_1 for the group of sensitives (as defined in equation (5.6)) may not, in this case, be a causal effect, even if we have blocked all the paths by conditioning on the correct variables and thus we are within the correct principal stratum. The reason for this is that the sensitive stratum belonging is dependent on X, being defined as "those women with $0 < R \leq X$"; under all our assumptions β_1 represents what follows:

$$E[Y(x + 1)|Z = 1, 0 \leq r \leq x + 1, C] - E[Y(x)|Z = 1, 0 \leq r \leq x, C], \quad (5.9)$$

i.e., the variation in the potential outcome $Y(\cdot)$ for a one unit variation of X, at the same level of C. The problem is that if X changes then the definition for the sensitive group does as well, and the assessment of a causal effect requires comparison of the *same group of units under different interventions*. Assumption **II** in Section 4.2 ensures that a woman who would develop an HRT-induced cancer at a certain level of HRT would also do if treated for a longer period (i.e. for a larger value of X). This grants that if we consider *increments* in HRT duration, then the group of already sensitive women does not change, but, still, women who previously did not develop the tumor ($R > X$) could enter the group as cases: for example, a particular

woman has a value of $R = 5$, is treated for $X = 4.5$ years and does not develop breast cancer. If we ask ourselves the question "what if she was treated for $x+1 = 5.5$ years?" then we see that the women enters the group of sensitive, thus modifying the stratum. On the other hand, if we consider a *decrement* of HRT duration we could observe women leaving the sensitive stratum to enter either the group of the "protected", i.e. those women that have a sufficiently high value of R not to develop HRT-induced cancers, or the group of "non-HRT-induced cancers". Assigning such individuals to the doomed group is not in line with the group definition "women with $R = 0$", thus creating some confusion.

If we could define the principal strata of interest without making use of the treatment variable X, and believed we have correctly placed every woman in its own stratum through our selection mechanism, then all the results of FR [2] about causality would hold, and β_1 would always be the causal direct effect of HRT duration on prognosis for sensitive women. With the current definition of principal strata that we have introduced in Section 4.2 following [10], β_1 would be a causal effect only "locally", i.e. for particular ranges of values of X such that HRT duration is "far enough" from the value of R for each woman not to induce a group belonging change with a variation of X. Such regions could not be derived from observed data using the procedure whe have proposed, since we are approximating stratum belonging through our logistic selection models, and not directly estimating R magnitude for each individual.

Chapter 6

Applications

6.1 The CAHRES Study

The Cancer and Hormones Replacement in Sweden (*CAHRES*) study is a nationwide, population-based case-control study of breast cancer occurrence among women, aged 50 to 74, without previously diagnosed breast cancer, born in Sweden and resident there between October 1, 1993, and March 31, 1995. Incident cases of invasive breast cancer were identified through the 6 Swedish regional cancer registries, and patients were asked through their physicians for written consent to accept a mailed questionnaire. Controls, frequency-matched to the expected age distribution of the cases, were randomly selected during the entire period of study from a continuously updated registry which provides national registration number, name, address and place of birth of all people residing in Sweden; women with a previous diagnosis of invasive cancer (other than non-melanoma skin cancer) were excluded from all analyses. For statistical reasons postmenopausal women only were included in the analyses, where age at menopause is defined as age at last menstrual period or age at bilateral oophorectomy, if one year or more prior to data collection (if later, women were considered premenopausal).

In Section 6.2 a subset of the original dataset consisting of 5929 individuals (2818 cases and 3111 controls) is considered, for which 6 variables are recorded:

- **case**, dichotomous, 0 if control, 1 if case
- **tumor grade**, polytomous $1 - 3$
- **tumor size** in cms, continuous
- **age** in years, continuous
- **bmi** in kg/m^2, continuous

- **hrt duration** in days, continuous.

We decided to use only complete data records, and this led to a total of 986 cases and 3009 controls, for which a continuous outcome variable, *NPI*, is constructed as explained in Section 6.2.

In Section 6.3 we present a similar analysis aiming, this time, to a discrete outcome, namely 5-years survival. After cleaning the dataset from missing values we have a total of 5092 individuals with complete data, 2351 of which are cases and 2741 controls, and we consider 7 variables:

- **death**, dichotomous, 0 if alive after 5 years, 1 if deceased
- **case**, dichotomous, 0 if control, 1 if case
- **age class**, continuous, five equally large classes from age 50 to 74
- **bmi** in kg/m^2, continuous
- **parity**, discrete
- **menarche**, continuous, age at first menstruation
- **hrt duration**, continuous.

6.2 Continuous Outcome: Nottingham Prognostic Index

We consider here a continuous outcome, the *Nottingham Prognostic Index* NPI, defined as $\ln(size + grade)$ of the cancer, where:

- **size** is the size of tumor in cms
- **grade** is a $1-3$ severity score based on the *modified Bloom-Richardson grading*,

the higher the index value, the worse the condition.

Our models, as presented in Section 5.2, are fitted to the observed data from the CAHRES study, employing both a plain logistic and a quadratic logistic selection mechanism (more on this can be found in Chapter 7). HRT duration, Age and BMI for each woman in the study are considered as covariates.

A comparison is made, with respect to a standard linear regression, aimed at assessing goodness of fit of the various models. A plain linear regression, or *naïve analysis*, fits the mean value of the NPI conditional on HRT, Age

and BMI, not considering possible differences due to the existence of a group distinction in the population. Such a model shows here a better fit, in terms of both correlation and residual variance, to the observed prognosis data. Our two models perform overall in the same way, with a slightly better fit for the one with the more complete selection model (quadratic logistic), albeit very small.

HRT effect estimates for the three models:

- PS Model, Plain Logistic Selection: $\beta_1 = -0.0430241$
- PS Model, Quadratic Logistic Selection: $\beta_1 = -0.0587184$
- Naïve Model: -0.0225993.

Summary of fit statistics for the three models:

Goodness of Fit	Correlation	Residual Variance
Plain Logistic	0.0820283	0.2639186
Quad.Logistic	0.0820442	0.2639178
Naïve	0.1066138	0.2626862

Table 6.1: Goodness of Fit - Continuos Outcome Application

The expression of the likelihood for the model with a quadratic logistic selection mechanism can be written down as follows:

$$\pi_{\underline{\alpha},X,C} := \frac{expit(\alpha_0+\alpha_3\text{Age}+\alpha_4\text{BMI})}{expit(\alpha_0+\alpha_1\text{HRT}+\alpha_2\text{HRT}^2+\alpha_3\text{Age}+\alpha_3\text{BMI})}$$

$$\ln L = \sum_{K_1} \ln \left[\pi_{\underline{\alpha},X,C} \frac{e^{-\frac{(NPI-\mu_0-\mu_1\text{AGE}-\mu_2 BMI)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} + (1-\pi_{\underline{\alpha},X,C}) \frac{e^{-\frac{(NPI-\beta_0-\beta_1\text{HRT}-\beta_2\text{AGE}-\beta_3\text{BMI})^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} \right] + \\ + \sum_{K} \ln expit(\alpha_0 + \alpha_1\text{HRT} + \alpha_2\text{HRT}^2 + \alpha_3\text{Age} + \alpha_4\text{BMI}) \tag{6.1}$$

We have allowed for AGE,BMI to have a different effect in the two groups, and did not employ ascertainment probabilities (which were not known). Estimates for such model are presented in Table 6.2.

Problems arise when trying to obtain the standard error estimates for these parameters from the hessian matrix computed by the optimization routine (using a Nelder-Mead method). Such matrix happens to be indefinite in sign, thus indicating the estimate to be a saddle point in the parameter space; different optimization methods did not give different results, and a deeper investigation of the outputs revealed that the selection model assigned all

the women to the first group (the doomed). This could either indicate that the covariates we have considered (*HRT* duration, *AGE,BMI*) do not bring relevant information for our selection mechanism to discriminate between the groups, or that there is no detectable effect of *HRT* on prognosis, which in turn leads to assignment of all women to the stratum where the therapy does not actually affect the outcome.

α_0	-0.6281073847	μ_0	2.6574134337	β_0	2.9479989324
α_1	0.4069481028	μ_1	-0.0008324058	β_1	-0.0587183965
α_2	-0.0992194920	μ_2	0.0104455145	β_2	0.0217099265
α_3	0.2749661877			β_3	0.0139408707
α_4	0.0544716150				

Table 6.2: Estimates for the model with quadratic logistic selection

The following output summarizes the results for the naïve analysis:

```
glm(formula = index ~ hrt1000 + age + bmi1)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
-1.796423  -0.346899  -0.008588   0.323511   2.127771

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.7141979  0.1861384  14.582   <2e-16 ***
hrt1000     -0.0225993  0.0105323  -2.146   0.0321 *
age         -0.0009448  0.0025683  -0.368   0.7130
bmi1         0.0091708  0.0040786   2.249   0.0248 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The standard analysis detects a significant effect of *HRT* duration on prognosis, showing a somewhat protective effect (the coefficient for therapy duration is negative); *AGE* does not appear to have a significant effect on *NPI*, whereas, albeit weakly, *BMI* seems to play a role.

Following the standard analysis interpretation one would conclude that there seems to be a protective effect of hormonal therapy duration on patients conditions after developing breast cancer, while a larger body mass index apparently pushes towards the opposite direction, i.e. a worsening of the prognosis.

It must be remarked, however, that *BMI* can be of difficult interpretation, since identical values of such measure could indicate very different situations from the medical point of view: a particular value of *BMI* could be attained either by being, say, not very tall and slightly overweight or by being very tall and well fit.

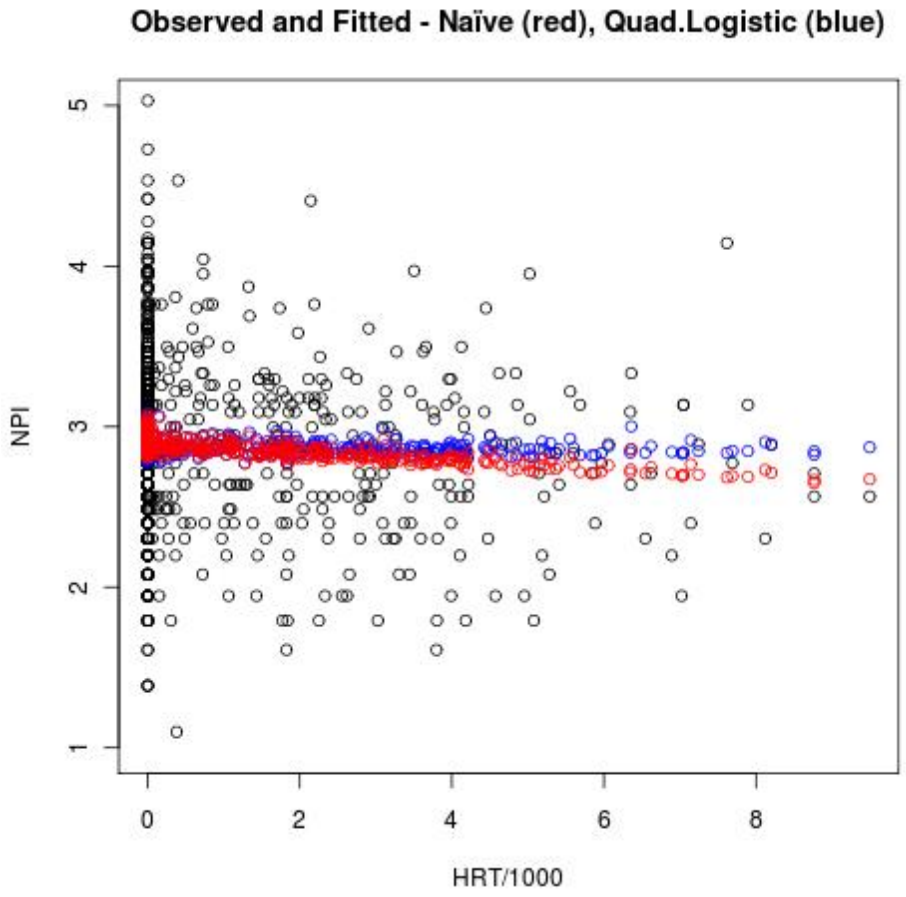

Figure 6.1: *NPI* versus *HRT* in thousands of days

Figure 6.1 represents a plot of *NPI* versus *HRT*, with fitted data from the naïve model and the quadratic logistic selection one overimposed. As aforementioned, our model assigns all the women to the first group, where no *HRT* effect on prognosis is present, we then see a fairly constant trend (not decreasing nor increasing) in *NPI* as duration grows (blue balls): variations are due to *AGE* and *BMI* effect. The red balls, representing fitted values from a standard analysis, show a decreasing trend in prognosis as *HRT* duration increases, as we expected from the previous output.

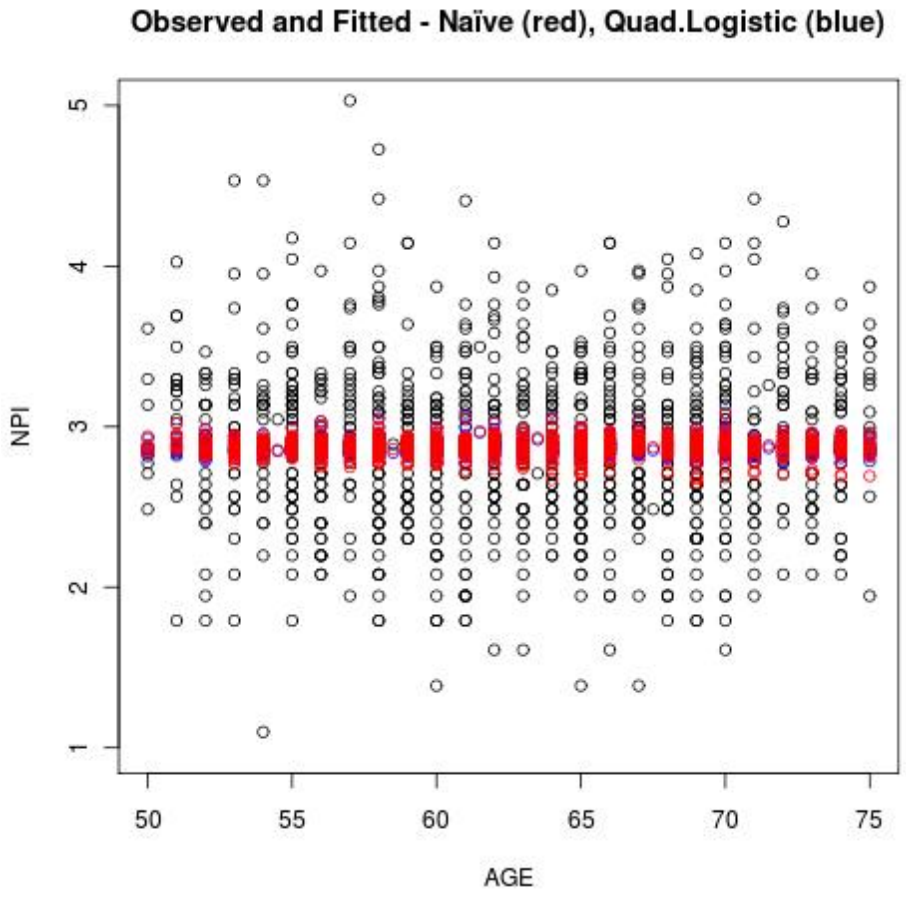

Figure 6.2: *NPI* versus *AGE*

Figure 6.2 shows that both the quadratic logistic selection based model and the standard model do not detect an appreciable effect of the age of the women with respect to their breast cancer prognosis, as measured with the *NPI*. This means that being older or younger appears not to affect the severity of the breast cancer for the women in this dataset.

It appears, moreover, that both models detect a small, but significant, effect of *BMI* on prognosis (lower plot), as can be seen in Figure 6.3. Still, as remarked before, *BMI* is not an unambiguous measure and should, therefore, be carefully evaluated when trying to draw conclusions (or possibly inference).

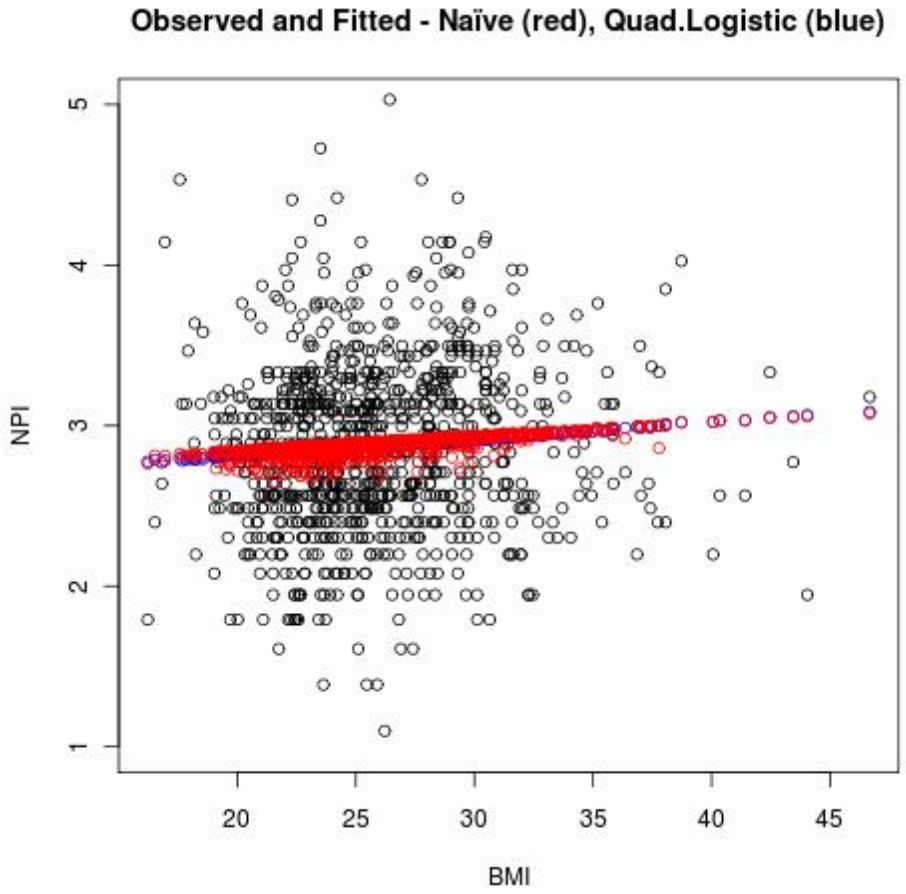

Figure 6.3: *NPI* versus *BMI*

6.3 Discrete Outcome: 5-Years Survival

A dichotomous (0/1) outcome is considered, i.e. 5-years survival (Y), for which models with plain and quadratic logistic selection mechanism are fitted and compared to a standard logistic regression. The dataset consists of 5092 individuals with complete data, 2351 cases and 2741 controls, for which the following covariates are measured (together with Y): HRT (X), BMI (C_1), Age at Menarche (C_2), Age Class (C_3), Parity (C_4).

In order to compare our model to a standard logistic regression in terms of performance we predict the outcome using both models, build two contingency tables of fitted against observed values, and calculate sensitivity and specificity values for each such table.

Such analysis has been carried out both with and without covariates, yelding slightly different results in terms of HRT effect estimate and goodness of fit. Exploratory investigations suggested that a quadratic term for HRT in the standard logistic model was not to be considered, while it has been included as a regressor in the sensitive component of the mixture in both settings and for the selection model in the covariates case.

Sensitivity turns out to be higher for the PS model prediction in both cases, whereas we observed a certain improvement in specificity for this model from the no covariates situation (where it's slightly lower) to the one involving covariates (slightly higher).

The likelihood for the PS (Principal Stratification) model with a quadratic logistic selection, letting $\pi_{\underline{\alpha},X,C} := \frac{expit(\alpha_0+\alpha_3 C_1+\alpha_4 C_2+\alpha_5 C_3+\alpha_6 C_4)}{expit(\alpha_0+\alpha_1 X+\alpha_2 X^2+\alpha_3 C_1+\alpha_4 C_2+\alpha_5 C_+\alpha_6 C_4)}$, is:

$$\begin{aligned}\ln L = &\sum_{K_1} \ln[\pi_{\underline{\alpha},X,C} \frac{e^{Y(\beta_0+\beta_1 C_1+\beta_2 C_2+\beta_3 C_3+\beta_4 C_4)}}{1+e^{\beta_0+\beta_1 C_1+\beta_2 C_2+\beta_3 C_3+\beta_4 C_4}} + \\ &+(1-\pi_{\underline{\alpha},X,C}) \frac{e^{Y(\gamma_0+\gamma_1 X+\gamma_2 X^2+\gamma_3 C_1+\gamma_4 C_2+\gamma_5 C_3+\gamma_6 C_4)}}{1+e^{\gamma_0+\gamma_1 X+\gamma_2 X^2+\gamma_3 C_1+\gamma_4 C_2+\gamma_5 C_3+\gamma_6 C_4}}] + \\ &+\sum_{K} \ln expit(\alpha_0+\alpha_1 X+\alpha_2 X^2+\alpha_3 C_1+\alpha_4 C_2+\alpha_5 C_+\alpha_6 C_4)\end{aligned} \tag{6.2}$$

A choice has been made, to keep the baseline log-odds-ratio of being a case (α_0) constant and equal to a fixed value, specifically $\alpha_0 = -6.0$. This choice should reflect possible previous knowledge about the study population, therefore we decided not to estimate it together with the other parameters. A small sensitivity analysis has been carried out, showing little difference in sensitivity and specificity values for values of α_0 in the range $[-7.0, -4.0]$.

The HRT-related parameters' estimates we obtained are:

HRT Effect	No Covariates	With Covariates
PS Model γ_1	0.1716152	0.29960684
PS Model γ_2	-0.3005484	-0.03214295
Naïve Model	-0.14793	-0.144746

Table 6.3: HRT effect on 5-years survival, PS vs Naïve, with and without covariates

Tables follow, which summarize predicted versus observed survival for a plain logistic model (Table 6.3) and for our model with quadratic logistic selection mechanism.

(a) No Covariates

	Pred. Alive	Pred. Dead
Obs. Alive	1677	308
Obs. Dead	308	58

(b) With Covariates

	Pred. Alive	Pred. Dead
Obs. Alive	1678	307
Obs. Dead	307	59

Table 6.4: Naïve model predicted vs fitted

(a) No Covariates

	Pred. Alive	Pred. Dead
Obs. Alive	1667	318
Obs. Dead	306	60

(b) With Covariates

	Pred. Alive	Pred. Dead
Obs. Alive	1682	303
Obs. Dead	305	61

Table 6.5: PS model predicted vs fitted

The following tables present the specificity and sensitivity results. These measures are widely used in statistics applied to the medical field, usually to evaluate the power of a new test (a new screening technique, for example) as opposed to the standard methods in use in terms of correct classification of, say, healthy and unhealty patients. In this case, the observed values are thought as being the standard method, and the model prediction as the new one. Specificity measures the proportion of correct predictions of survival over the actual survived, whereas sensitivity measures the proportion of correct predictions of death over the actual non-survivors; clearly, the higher these two values, the better the model's fit.

(a) No Covariates

	Sensitivity	Specificity
Naïve Model	0.1585	0.8448
PS Model	0.1639	0.8398

(b) With Covariates

	Sensitivity	Specificity
Naïve Model	0.1612	0.8453
PS Model	0.1667	0.8473

Table 6.6: Sensitivity and Specificity Comparison

Overall, our model shows a slight improvement in terms of fit based on these two indexes.

Chapter 7

Simulation Studies

7.1 Rationale

In order for our simulation studies to take place we first of all need to set up the mechanism that will generate the data we will be fitting our models to. Such structure is required to be to a certain extent flexible, since we will need to be able to simulate data under different "scenarios", and will therefore depend on parameters which we will tune. This notwithstanding, the mechanism will of course reflect our beliefs about what the relationships among the involved variables could be like.

The focus is on a continuous outcome, say Y, which can be regarded as a prognostic index and the effect of a covariate X, which mimics HRT duration; for clarity and simplicity all simulations will be carried out in the absence of covariates other than HRT. We also need to generate a value for R to be able to let people in different groups ("healthy","sensitive", "doomed", when we believe this distiction exists) have different characteristics, thus creating (or eliminating) a group induced heterogeneity in the outcome.

The first simulation study, which is presented in section 7.3, will address the issue of correct targeting of the true parameters' values through estimation with the model proposed in equation (5.6). The aim is to show the benefits of putting some effort in the development of the selection model (the $\pi_{\underline{\alpha},X}$) and to underline the intepretational difficulties of which [10]'s implicit model suffers.

In section 7.4 we will present the results from an intensive simulation study aimed at comparing our proposal to a standard approach modeling. The main issue we address is "when is one of the two approaches better than the other in terms of explanatory power?"; in order to give an answer to

such question we tune our generating mechanism so to be able to investigate various scenarios and evaluate goodness of fit measures (namely correlations between observed and predicted values and models' residual variances).

To depict in a clear way which situations we are going to simulate we make use of graphs, in order to better show the connections we believe may exist among variables. Figure 7.1 presents a setting in which the following effects exist:

- direct effect of X on Y
- direct effect of Z on Y
- joint effect of R and X, one independent of the other, on Z
- direct effect R on Y

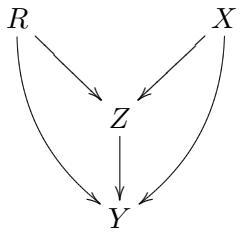

Figure 7.1: Underlying Truth: Strong Group Effect on Prognosis

In this situation our model, which assumes the presence of heterogeneity induced by sub-groups of people on the outcome, is therefore correct, and we would expect it to perform better than a standard analysis in predicting the outcome.

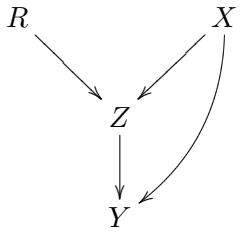

Figure 7.2: Underlying Truth: No Group Effect on Prognosis

Figure 7.2 shows the opposite situation, i.e. the total absence of a group effect on the outcome, thus making a standard model, which does not take

group belonging into account, the most correct tool for an analysis.

To be able to study with more accuracy the adequateness of our model we propose to analyze various intermediate situations between the two we just presented. Trying to gradually "shift" from 7.1 to 7.2, we make our comparisons on scenarios that can be represented with Figure 7.3 by tuning, time by time, the strength of the direct association of R and Y.

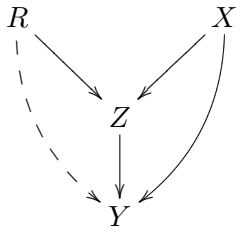

Figure 7.3: Underlying Truth: Intermediate Group Effect on Prognosis

The aim is to study the behaviour of our model, compared to the standard one, analyzing datasets which are gradually moving away from our assumption of different sub-groups, until the point in which such assumption is completely wrong (i.e. no groups). We need to do this since in reality R is unobservable, we are estimating its effect via our selection mechanism $\pi_{\underline{\alpha},X}$, and we would like to feel comfortable using our model also when we do not have any clue on what the underlying truth could be.

7.2 Simulation Structure

Code has been written in *R 2.10.1* language running under *UBUNTU 10.04 LTS - Lucid Lynx* for these simulations. The function that generates the datasets and estimate the models accept the following inputs:

- number of simulations
- size of each simulated dataset
- proportion of "doomed" p
- proportion of not-treated p'
- minimum and maximum value for R if "sensitive" (a, b)
- prognosis mean value for the "non HRT-induced" tumors μ_0

- prognosis mean value for the "HRT-induced" tumors β_0
- HRT gradient for the mean prognosis of the "HRT-induced" tumors β_1
- prognosis standard deviation for the "non HRT-induced" tumors σ_0
- prognosis standard deviation for the "HRT-induced" tumors σ_1
- a 3-elements vector defining the scenario, namely weights for μ_0, β_0 and β_1 called A.

The default values are tuned so to yeld 2% of Z=1 (cases) (1% doomed + 1% HRT induced), a balanced (number of cases=number of controls) sample is then drawn accordingly to fixed ascertainment probabilities and the estimation procedure applies. The parameters matrix is stored and possible convergence issues are recorded at each loop (see the code for details).

The stochastic mechanism generating values for R, X, Z and Y is structured as follows:

$$\varphi_R = \begin{cases} 0 & p \\ U(a,b) & 1-p. \end{cases} \tag{7.1}$$

A uniform distribution is chosen as no direct information is available on the unobservable variable R in the absence of covariates, leading to a simple way of splitting the women in the two subgroups without inducing any dependence with the X variable; R is in fact something we may call an "intrinsic" characteristic, and should not be affected by anything but random variation in this simplified setting. Applying our model to real data will require instead that we formulate some associational assumption linking R to the available covariates, so to allow the selection mechanism to assign each woman to one of the two strata.

$$\varphi_X = \begin{cases} 0 & p' \\ \lambda e^{-\lambda x} & 1-p'. \end{cases} \tag{7.2}$$

The choice for a zero-inflated exponential is motivated by exploratory analysis of real datasets, and mimics quite well the real HRT duration distribution in such data with an estimated λ of 0.422594, thus fixed in the code at such a level.

$$\varphi_{Z|R=r,X=x} = \begin{cases} 0 & r > x \\ 1 & r \leq x. \end{cases} \tag{7.3}$$

As in Section 4.2, assumption *II* and further considerations lead us to structure the probability of actually developing breast cancer contrasting the values for R and X for each woman. If $r > x$ this mean that the subject has an intrinsic resistance to developing HRT-induced cancer which protects her, thus keeping her safe; if this is not the case then a woman could either be a case because she is doomed ($R = 0$) or because her resistance is lower than what required not to develop the tumor. Generating data in this way ensures that a woman with $R = 0$ will *always* be a case (X can never be less than 0).

$$\varphi_{Y|Z=1,R=r,X=x} = \begin{cases} N(\mu_0, \sigma) & r = 0 \\ N(\beta_0 + \beta_1 x, \sigma) & r > 0. \end{cases} \tag{7.4}$$

As stated already, the choice for a normal distribution in the two subgroups is motivated by real data analysis; moreover, the distributional assumptions reflect the belief that if a woman develops a non HRT-induced cancer, then HRT shouldn't have any effect at all on the prognosis (a justification for this can be found in the medical literature). This is encoded by the absence of X in the mean prognosis level for the doomed group, assumed to be constant (in real data application we may want to let it vary accordingly to measured covariates, as already seen in Chapter 6, but still *not* to HRT duration).

7.3 Modeling $m(x, C)$: Parameters Targeting and Implicit Model Limitations

The simulation study we present in this section is aimed at comparing estimate properties for our model with respect to those for the implicit model mentioned in 4.3 and proposed by [10], in a case/control setting with no covariates. The estimand of interest is the quantity in equation (4.1), re-expressed, under [10]'s implicit model assumptions, using the identity link $g(\cdot) = \cdot$ and through equation (4.4) as:

$$m(x; \xi, \alpha) = \frac{\xi_1 X}{1 - \frac{expit(\alpha_0)}{expit(\alpha_0 + \alpha_1 X)}} \tag{7.5}$$

assuming simple linear and logistic models for $E(Y|Z=1,X)$ and $\pi(X)$ as follows

$$\begin{aligned} E(Y|Z=1,X,\xi) &= \xi_0 + \xi_1 X \\ logit\pi(X;\alpha) &= \alpha_0 + \alpha_1 X. \end{aligned} \tag{7.6}$$

As opposed to such model, our proposal estimates $m(x)$ as a difference in the mean levels (again, the identity link $g(\cdot) = \cdot$ has been chosen) of the two subgroups of doomed and sensitive women, defined as in (5.5); following (4.1) we then have:

$$\begin{aligned} m(x) &= E\{Y(x)|0 < R \leq x\} - E\{Y(x)|R=0\} = \\ &= \beta_0 + \beta_1 X - \mu_0 = \\ &= (\beta_0 - \mu_0) + \beta_1 X = \\ &= \psi_0 + \psi_1 X. \end{aligned} \tag{7.7}$$

The ψ parameters have a direct interpretation, ψ_0 being an intrinsic, non-HRT-dependent difference in prognosis between women with different tumor subtypes and ψ_1 the x-gradient on $m(x)$; analogous parameters in the case of the explicit model would be ξ_1/α_1 (resulting from $\lim_{x\to 0^+} m(x;\xi,\alpha)$) and ξ_1, which is also the coefficient for X in the classic linear regression $E(Y|Z=1,X)$.

500 to 1000 samples of 200000 units each have been generated, out of which all of the cases and an approximately equivalent number of controls for each sample were drawn. With the default settings this leads to about 4000 cases and a similar number of controls at each step. This has been done three times, once for each of the following possible scenarios:

1. strong group effect on difference in prognosis, as in Fig.7.1
2. weak group effect on difference in prognosis, as in Fig.7.3
3. no group effect on difference in prognosis, as in Fig.7.2

In scenario **1.** data are generated so that the underlying true mechanism is correctly described by our model, whereas in scenario **3.**, where no distinction exist between doomed and sensitives, the correct model would be

a simple linear regression $E(Y|Z=1,X)$. Scenario **2.** recreates a half-way situation, the strength of the group effect being still present, but weaker than in Scenario **1.**. What we want to show is the ability of the two models to correctly (unbiasedly) target the true parameter values (that we fix in the generating mechanism) indexing the true model for $m(x;\tilde{\psi}) = \tilde{\psi}_0 + \tilde{\psi}_1 X$. The generating mechanisms and the true models for the three scenarios are:

1.
$$Y|Z=1,X,R=0 \sim N(\mu_0,\sigma)$$
$$Y|Z=1,X,R>0 \sim N(\mu_0'+\mu_1'X,\sigma)$$

$m(x;\tilde{\psi}) = \tilde{\psi}_0 + \tilde{\psi}_1 X$, with $\tilde{\psi}_0 = \mu_0' - \mu_0, \tilde{\psi}_1 = \mu_1'$

2.
$$Y|Z=1,X,R=0 \sim N(\mu_0 + \frac{\mu_1'}{k}X,\sigma)$$
$$Y|Z=1,X,R>0 \sim N(\mu_0'+\mu_1'X,\sigma)$$

$m(x;\tilde{\psi}) = \tilde{\psi}_0 + \tilde{\psi}_1 X$, with $\tilde{\psi}_0 = \mu_0' - \mu_0, \tilde{\psi}_1 = \mu_1'(1-1/k), k>1$

3.
$$Y|Z=1,X,R=0 \sim N(\mu_0+\mu_1X,\sigma)$$
$$Y|Z=1,X,R>0 \sim N(\mu_0+\mu_1X,\sigma)$$

$m(x;\tilde{\psi}) = 0$, i.e. no direct group effect nor difference in prognosis imputable to HRT.

At each simulation step, parameters are estimated for our model and the implicit one, values are stored and eventually sample distributions are obtained. Histograms of such distributions are plotted, together with confidence intervals (at a 95% confidence level) for the estimates; a vertical red line on the histograms represents the true parameter value we have fixed for each scenario. In addition, R^2 statistics are computed at each step for our model and for a simple linear regression, sample distributions are plotted and the R^2 mean values considered, to assess which model has the best fit to the simulated data.

Different choices for the equations in (7.6) have been considered, but more complex models did not show better performance than these. As for our proposal, two different way of modelling the selection model $\pi_{\underline{\delta},X}$ have been investigated:

1. a plain logistic selection, $logit\pi_{\underline{\delta},X} = \delta_0 + \delta_1 X$
2. a quadratic logistic selection, $logit\pi_{\underline{\delta},X} = \delta_0 + \delta_1 X + \delta_2 X^2$.

This effort in the modelization of $\pi_{\underline{\delta},X}$ is aimed at showing whether, having (hopefully) improven the selection mechanism, such additional knowledge could result in a reduction of the bias in our model's estimates.
The actual generating value in each situation have been chosen as follows:

1.
$$Y|Z = 1, X, R = 0 \sim N(2.5, 0.5)$$
$$Y|Z = 1, X, R > 0 \sim N(1.9 + 0.053X, 0.5)$$

$\tilde{\psi_0} = 1.9 - 2.5 = -0.6, \tilde{\psi_1} = 0.053$

2.
$$Y|Z = 1, X, R = 0 \sim N(2.5 + \frac{0.053}{4}X, 0.5)$$
$$Y|Z = 1, X, R > 0 \sim N(1.9 + 0.053X, 0.5)$$

$\tilde{\psi_0} = 1.9 - 2.5 = -0.6, \tilde{\psi_1} = 0.053(1 - 1/4) = 0.03975$

3.
$$Y|Z = 1, X, R = 0 \sim N(1.9 + 0.053X, 0.5)$$
$$Y|Z = 1, X, R > 0 \sim N(1.9 + 0.053X, 0.5)$$

$m(x; \tilde{\psi}) = 0.$

Results for Scenario 1.
Fig.7.4 shows the sample distributions for the estimates of ψ_0, ψ_1 under the two proposed models (plain logistic and quadratic logistic selection) and of $\xi_1/\alpha_1, \xi_1$ under the implicit model; the vertical red lines represent the true target values, $\tilde{\psi_0} = -0.6, \tilde{\psi_1} = 0.053$ in this scenario. As stated before, the implicit model parameters appear not to have a straightforward interpretation, also being quite far away from the values of a simple linear formulation of the difference in prognosis $m(x; \tilde{\psi})$. The non linearity of the model brings uncertainty on how to read ξ_1/α_1 and ξ_1 in terms of $m(x)$; ξ_1 is, moreover, the X coefficient we obtain in the linear regression $E(Y|Z = 1, X)$, thus not being eligible as informative of an x-gradient on $m(x)$ if a group difference actually exists (as in this scenario).

As can be seen, neither of the implicit model estimates lies on a range of value that makes it comparable with the true $\tilde{\psi}$ parameters, indicating that

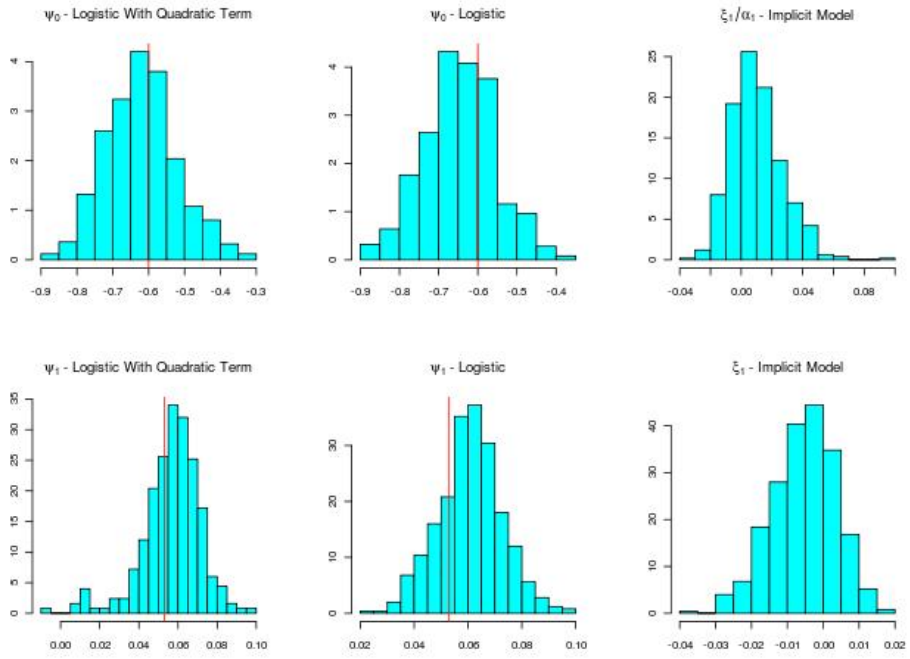

Figure 7.4: Sample Distributions of Parameters Estimates - Scenario 1.

they are estimating something different. The ψ estimates from our model appear to correctly target the estimand of interest, also showing a modest improvement in terms of bias if we use the quadratic logistic selection model; keep in mind that in this scenario a difference due to group belonging does actually exist, so we expected our model to perform better with respect to the implicit one or to a simple linear regression, as can also be seen from the R^2 statistics in Fig.7.5 (red lines are their mean values, also plotted over each histogram).

	Quadratic Logistic	Plain Logistic	Implicit Model
ψ_0	-0.6192 (-0.8127, -0.4257)	-0.6451 (-0.8253, -0.4649)	×
ψ_1	0.0567 (0.0272, 0.0862)	0.0607 (0.0375, 0.0838)	×
ξ_1/α_1	×	×	0.0101 (-0.0225, 0.0427)
ξ_1	×	×	-0.0054 (-0.0227, 0.0119)

Table 7.1: Confidence Intervals of Parameters Estimates - Scenario 1.

Table 7.1 contains the confidence intervals for the parameters estimates: employing the quadratic logistic model reduces a little bit (though not in a statistically significant way) the bias with respect to the plain logistic

one, and they both appear to be on target. The implicit model estimates, however, detect no difference in prognosis at all, the confidence intervals for both parameters contains the zero at a 95% confidence level, thus making the estimated $m(x;\xi,\alpha)$ not significantly different from zero.

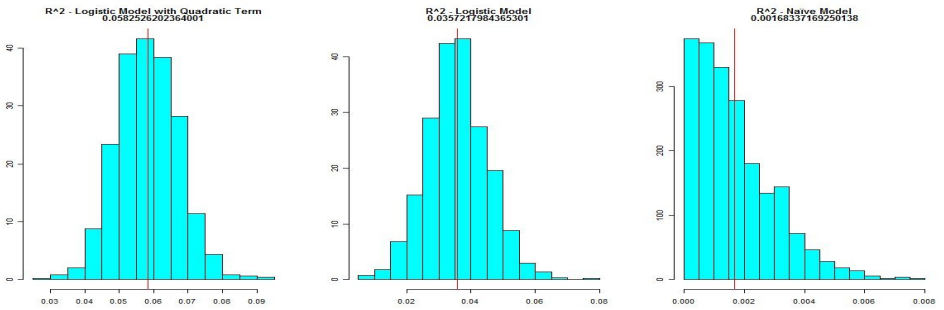

Figure 7.5: R^2 Sample Distributions and Their Mean Values - Scenario 1.

Results for Scenario 2.
Sample distributions are plotted for the parameters estimates and for the R^2 as in the previous scenario. A table with confidence intervals follows.

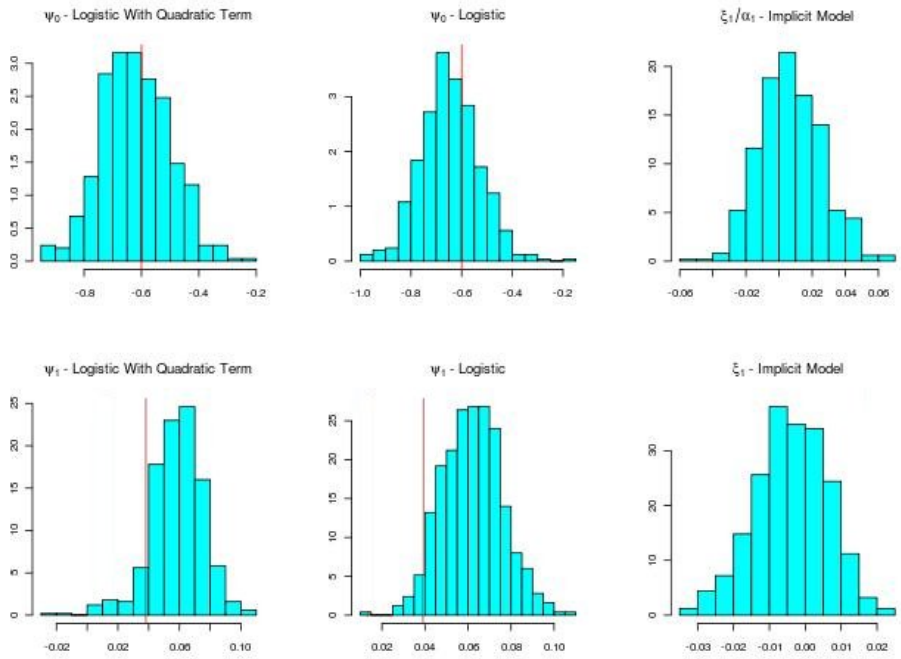

Figure 7.6: Sample Distributions of Parameters Estimates - Scenario 2.

The weakness of the link between group and prognosis (induced by introducing a small effect of HRT on prognosis in the doomed group), makes it more difficult for our model to correctly asses the group membership, resulting in a slightly increased bias in the estimates of $\tilde{\psi}_1$, while the baseline non-group-dependent difference $\tilde{\psi}_0$ is still well targeted.

	Quadratic Logistic	Plain Logistic	Implicit Model
ψ_0	-0.6195 (-0.8493, -0.3897)	-0.6438 (-0.8682, -0.4194)	×
ψ_1	0.05832 (0.0245, 0.0922)	0.0620 (0.0346, 0.0895)	×
ξ_1/α_1	×	×	0.0074 (-0.0292, 0.0440)
ξ_1	×	×	-0.0040 (-0.0239, 0.0160)

Table 7.2: Confidence Intervals of Parameters Estimates - Scenario 2.

Again, the implict model fails to detect any effect at all on difference in prognosis, and so does a simple linear regression, not detecting ($\xi_1 \simeq 0$) an existing direct HRT effect on prognosis.

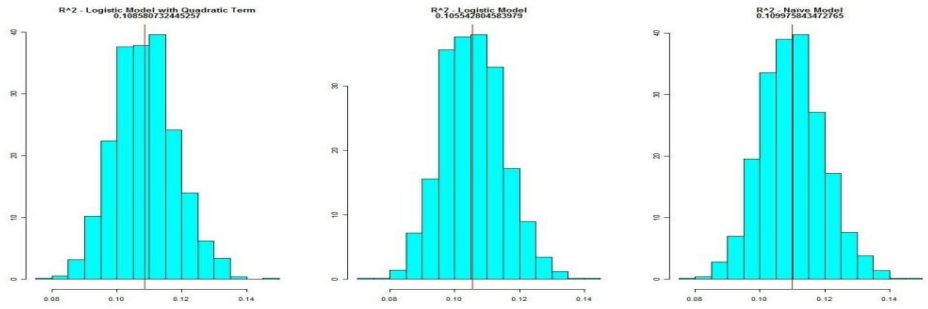

Figure 7.7: R^2 Sample Distributions and Their Mean Values - Scenario 2.

Figure 7.7 shows, nonetheless, that in such a mid-way situation between Scenario 1. and 3. the goodness of fit of our models, albeit slightly higher, is not statistically different from the naïve model's one.

Results for Scenario 3.
The mixture model fails most of the times to reach convergence, the reason for this being the fact that there actually is no group effect, thus X is not a predictor of belonging to a stratum or another (there exist no such distinction). Whenever this happens in a real analysis, the advice should be to either spend more efforts on the selection model parameterization and try different predictors, or to consider that there could actually be no group distinction as we are trying to model. When the model converges the results

appear to be in line with the underlying true scenario for what concern $\tilde{\psi}_0$, while the estimates for $p\tilde{si}_1$ exhibit a somewhat stronger than before bias, as can be seen from the sample distributions plots 7.8,7.9 and the confidence intervals table 7.3.

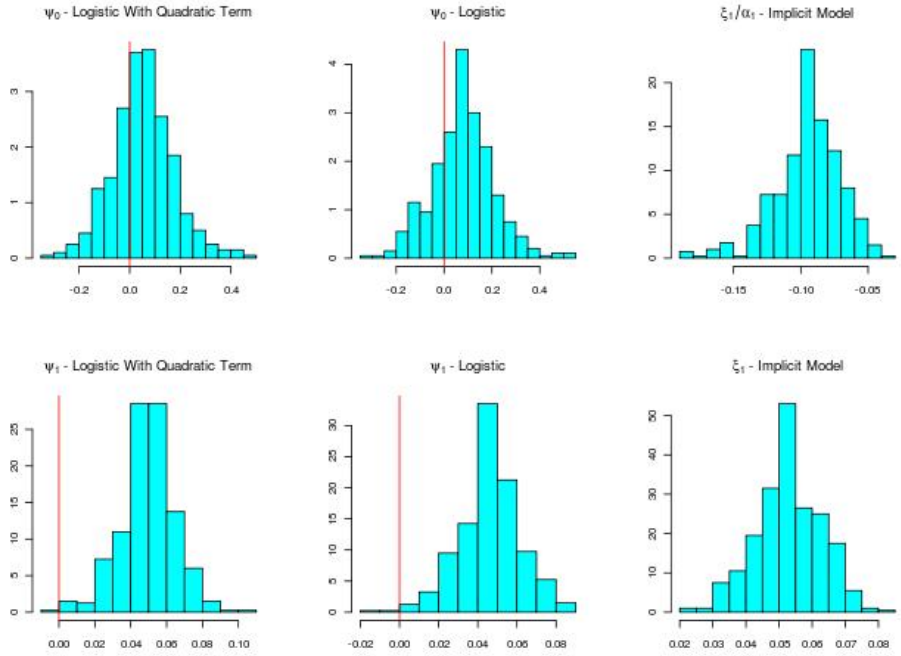

Figure 7.8: Sample Distributions of Parameters Estimates - Scenario 3.

	Quadratic Logistic	Plain Logistic	Implicit Model
ψ_0	0.0530 (-0.1844, 0.2904)	0.0781 (-0.1744, 0.3305)	×
ψ_1	0.0498 (0.0198, 0.0797)	0.0461 (0.0157, 0.0764)	×
ξ_1/α_1	×	×	-0.0948 (-0.1432, -0.0463)
ξ_1	×	×	0.0529 (0.0337, 0.0722)

Table 7.3: Confidence Intervals of Parameters Estimates - Scenario 3.

ψ_0 correctly estimates, under both the quadratic and the plain logistic selection model, the true magnitude of $\tilde{\psi}_0$, whereas the analogous baseline value for the implicit model detects a significant intrinsic difference in prognosis which is of difficult interpretation. ψ_1 shows now a larger bias than what obtained in the previous two scenarios, and it appears clear here that including the quadratic term in the selection model did not bring any improvement. Note that the naïve model parameter, ξ_1 correctly detects an

existing effect of HRT on prognosis (not on thmulticolumne difference), this being the scenario in which this model is actually the correct interpretation of such association, also yelding a good estimate of its magnitude.

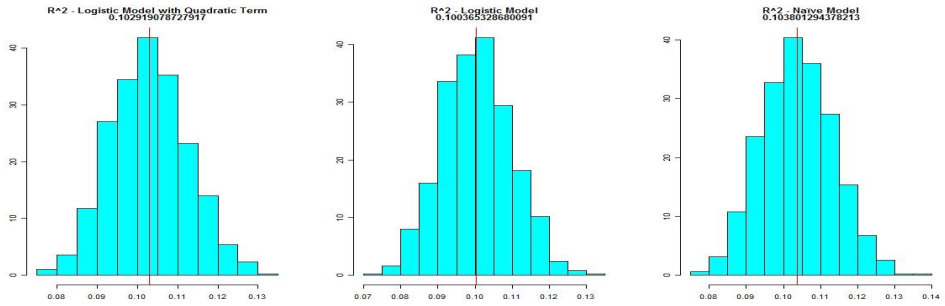

Figure 7.9: R^2 Sample Distributions and Their Mean Values - Scenario 3.

The R^2 sample distributions and their mean value show, moreover, that a plain linear regression performs slightly better (though the difference is not statistically significant) than both our models.

7.4 Goodness of Fit Comparison with the Standard Approach

The aim of this simulation study is to compare our models with a standard (naïve) analysis, i.e. a simple linear regression, in terms of goodness of fit to the simulated data. The comparison is made on the basis of statistics such as the correlation coefficient ρ (between observed and fitted outcome data), and the residual variance. Various scenarios are considered and indexed by the vector A as defined in Section 7.2:

Scenario	A
I.	(1,0,0)
II.	(1,0,1/4)
III.	(1,0,1/2)
IV.	(1,0,3/4)
V.	(1,0,1)
VI.	(0,1,1)

Table 7.4: Scenarios for Goodness of Fit Comparison Simulation

Scenario I. is exactly the **Scenario 1.** from Section 7.3, generating data with a strong group effect on prognosis, and Scenario VI. is the former **Scenario 3.** of no group effect at all. Scenarios II. to V. represent a range

of possible intermediate situation between the latter two; a gradual shift is induced introducing an effect of HRT on prognosis in the doomed group of increasing strength until the point (V.) where its magnitude is equivalent in both groups, and thus any discrepancy is imputable only to an intrinsic (and constant over levels of HRT) baseline difference.

Under scenarios I. to V. the simulating mechanism has been tuned in order to generate different proportions of sensitives in the population, in particular four situations have been considered: $1\%, 0.7\%, 0.01\%$ and 0.002% of sensitives generated. Our aim was to test whether, even with a population with a very low number of sensitive individuals, our models would be still able to assign people to their stratum and retain an overall good fit. The statistics indicating the best fitting model are, from time to time, highlighted in the tables, and plots of ρ and residual variance against the percentage of sensitives are shown. 2000 runs with populations of size 200000 have been generated for each combination of scenario and sensitive percentage, and samples of approximately 4000 cases and a similar number of controls have been drawn for analysis out of each generated dataset.

Eventually, a particular analysis is presented, which investigate more deeply Scenario I., both varying the sensitives proportion and the baseline mean prognosis level in the two groups, providing graphical evidence of the behaviour of our models as compared to a standard linear regression. The same number of runs and population dimensions as before apply.

Results for Scenario I. - $A = (1, 0, 0)$

	Correlation			Residual Variance		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.62743	0.63530	0.57662	0.40416	0.39680	0.43566
0.7%	0.67910	0.68110	0.62718	0.37283	0.36956	0.40454
0.01%	0.61782	0.61881	0.60447	0.29617	0.29489	0.30090
0.002%	0.34952	0.35395	0.29242	0.25561	0.25270	0.26544

Table 7.5: Correlation and Residual Variance Results for Scenario I.

Under this scenario data are generated with a strong group effect on prognosis and no effect at all of HRT on the doomed group outcome. This is the situation where our models are the correct tool for interpretation of the underlying truth. As can be seen from Table 7.5 and from Figure 7.10 our models (and in particular the quadratic logistic one) perform better than a naïve analysis in terms of goodness of fit, as measured through the correlation coefficient ρ (the higher, the better) and the residual variance (the lower, the better). The differences between the fit of the two models employ-

ing the selection mechanism appears to be very small, still always favouring the quadratic logistic selection procedure. The plots show, moreover, a sort of "fork", an increasing difference in goodness of fit between our models and the standard one as the sensitive percentage in the population increases: we could think of such trend as indicating a better functioning of the selection models due to a greater presence of information, which allows for a more precise assignment. It is not clear, however, why the correlation coefficient shows a peak at 0.7% and then decreases at 1% of sensitives for all three models; more simulations (considering a broader range of values for such percentage), could help gaining more insight on this issue.

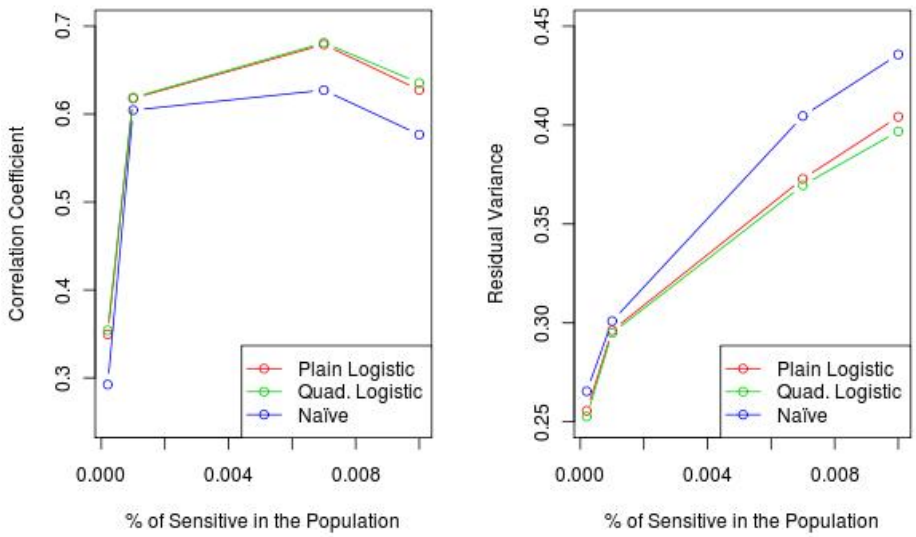

Figure 7.10: Correlation and Residual Plots for Scenario I.

Results for Scenario II. - $A = (1, 0, 1/4)$

	Correlation			Residual Variance		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.63151	0.64516	0.56965	0.40401	0.39015	0.43893
0.7%	0.67709	0.67910	0.63738	0.34947	0.34761	0.37474
0.01%	0.59586	0.61056	0.59517	0.30769	0.29852	0.29976
0.002%	0.30084	0.31038	0.28585	0.26543	0.26495	0.26736

Table 7.6: Correlation and Residual Variance Results for Scenario II.

Scenario II. introduces a small effect of HRT on prognosis in the doomed

group (1/4 of the effect in the sensitive group). This induces some difficulties for the selection models, that assign women to groups using the assumption that HRT affects the prognosis for the sensitive group only. This notwithstanding the very same comments for Scenario I. apply: the two models using the selection mechanism still perform better than the standard one (apart at 0.01%, where the naïve model appears to be slightly better than the plain logistic one, but still not than the quadratic logistic) and the fork is still evident from the plots in Figure 7.11.

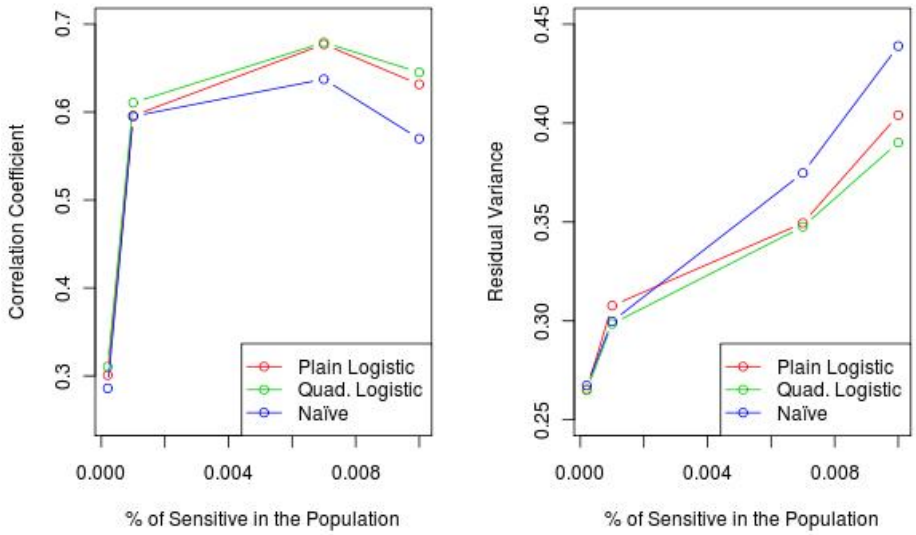

Figure 7.11: Correlation and Residual Plots for Scenario II.

Results for Scenario III. - $A = (1, 0, 1/2)$

	Correlation			Residual Variance		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.62190	0.63267	0.57237	0.40364	0.39323	0.42874
0.7%	0.64901	0.65388	0.61775	0.38607	0.38086	0.40003
0.01%	0.62510	0.63099	0.61984	0.30102	0.29691	0.30133
0.002%	0.33393	0.34767	0.32124	0.27523	0.27353	0.27604

Table 7.7: Correlation and Residual Variance Results for Scenario III.

The strength of the effect of HRT on the doomed group prognosis, introduced in Scenario II., is increased, reaching half of the magnitude in the sensitive group. No evident differences from the results in the previous sce-

nario is found, with the model employing the quadratic logistic selection still performing better than the other two.

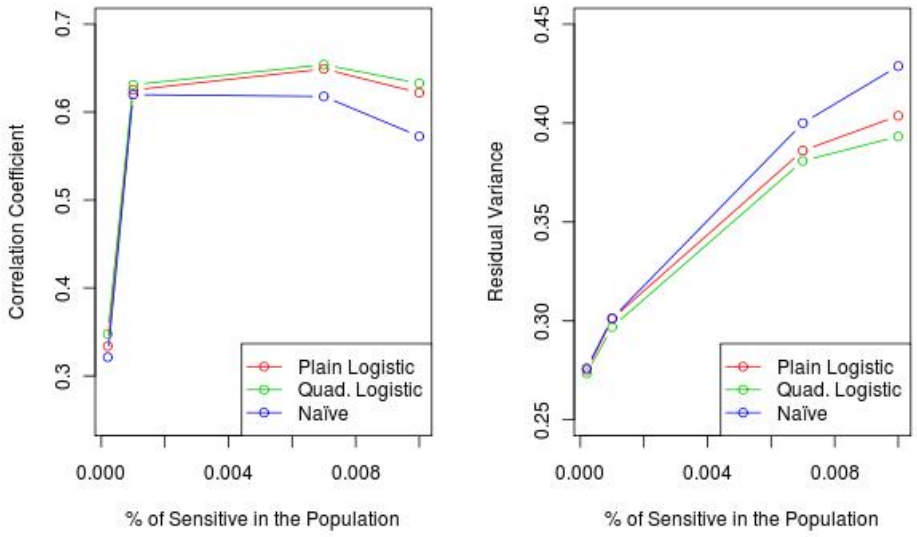

Figure 7.12: Correlation and Residual Plots for Scenario III.

Results for Scenario IV. - $A = (1, 0, 3/4)$

	Correlation			Residual Variance		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.63353	0.63871	0.58053	0.34829	0.34247	0.37952
0.7%	0.69669	0.70192	0.65219	0.38607	0.38086	0.40003
0.01%	0.61126	0.61451	0.60811	0.27998	0.27818	0.27894
0.002%	0.43310	0.43700	0.41858	0.25387	0.25227	0.25509

Table 7.8: Correlation and Residual Variance Results for Scenario IV.

HRT effect on doomed group prognosis is further increased to 3/4 times the effect among the sensitives. We are approaching a situation in which no difference in prognosis for the group will be imputable to HRT effect, thus depriving the selection mechanism of a valuable source of information to work well. The differences in goodness of fit at low levels of sensitive percentage is now less clear, the values being very close, but the fork still remains at higher levels. Again, the quadratic logistic selection allows for the best overall fit.

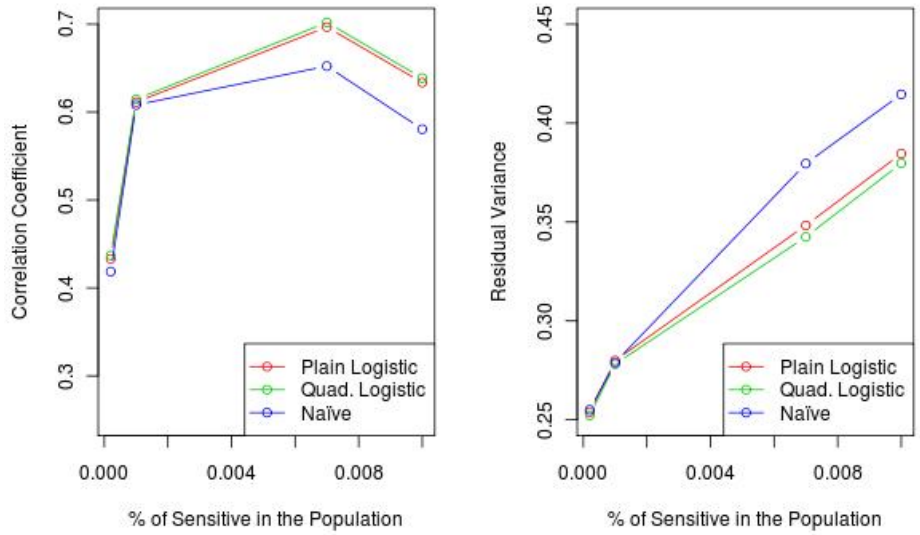

Figure 7.13: Correlation and Residual Plots for Scenario IV.

Results for Scenario V. - $A = (1, 0, 1)$

	Correlation			Residual Variance		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.66352	0.67029	0.64641	0.35420	0.34786	0.37205
0.7%	0.68397	0.68778	0.66678	0.34993	0.34614	0.35307
0.01%	0.63055	0.63329	0.63280	0.28442	0.28311	0.28059
0.002%	0.41097	0.41164	0.41294	0.24809	0.24796	0.24719

Table 7.9: Correlation and Residual Variance Results for Scenario V.

This is a limit situation where the only group difference in prognosis is imputable to baseline non-HRT-dependent characteristics; the strength of HRT effect is, indeed, the same in both groups (the proportionality coefficient being equal to 1 in this case). At low percentage of sensitives levels a naïve analysis appears to perform slightly better than our models, and the fork we have observed in the previous scenarios is now less evident. The selection mechanism is clearly having some difficulties assigning women to the correct stratum because of the lack of HRT-induced prognosis difference in the two groups; group distinction, though, does still exist thanks to a different baseline effect, which our models manage to detect, for the two groups. A better performance of our models in presence of more sensitive individuals is then still present.

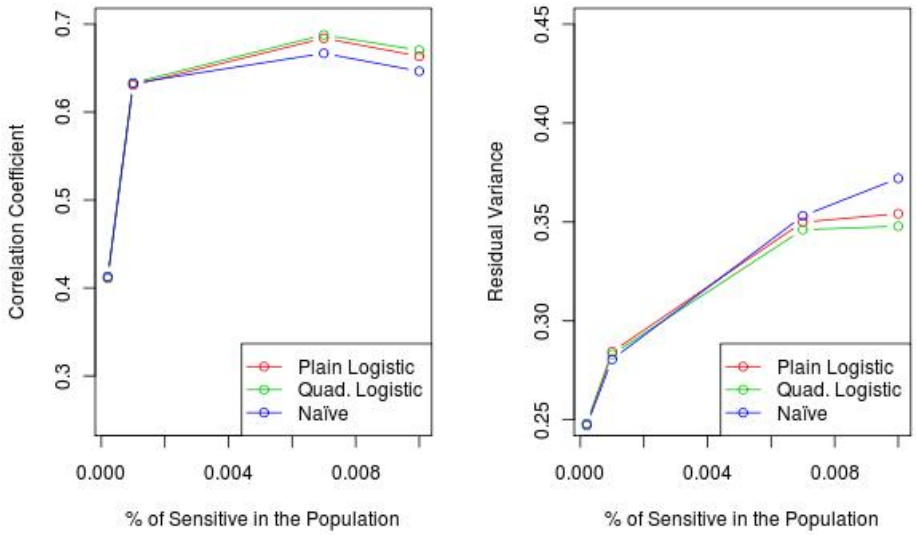

Figure 7.14: Correlation and Residual Plots for Scenario V.

Results for Scenario VI. - $A = (0, 1, 1)$

Correlation			Residual Variance		
Logistic	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
0.32054	0.32120	0.32239	0.25487	0.25476	0.25449

Table 7.10: Correlation and Residual Variance Results for Scenario VI.

In this scenario the selection mechanism is superfluous, since the dataset are generated with no group-induced difference in prognosis, i.e. we have the same baseline mean value and the same magnitude of HRT effect in the two groups, the only differences we can observe in terms of doomed and sensitive women are due to randomness. The standard analysis is here the correct tool, as Table 7.10 shows, yelding better fitting values as compared to the other two models. This is exactly what we expected, nonetheless the difference is very small, and we may regard at this as an indication of some sort of robustness our models show with respect to wrong initial assumptions.
Next, we propose a more insightful analysis for Scenario I. for a more complete evaluation of our models' performances, as opposed to the standard analysis; this is done analysing various sub-scenarios of the situation $A = (1, 0, 0)$.

Deeper investigation of Scenario I.
In this paragraph we present the result of a simulation study, on the line of those discussed until now, based entirely on Scenario I. and some of its sub-scenarios. Looking for a deeper understanding of our models' ability to describe the data, we propose goodness of fit measures along with plots showing, from time to time, the generated sample units together with the fitted models; in particular, the best performing model (the one employing the quadratic logistic mechanism) will be plotted over the data, together with the regression line from a standard analysis.

This study is structured as follows:

- datasets are generated following the scenario $A = (1, 0, 0)$
- three sensitive percentage values are considered, $1\%, 0.1\%, 0.01\%$
- three "distances" between mean non-HRT-dependent baseline values in the two groups are considered, i.e. (following the notation of equation (7.4)) $\beta_0 - \mu_0 = 1, 2, 3$.

Hence, a total of nine sub-scenarios is considered and 2000 datasets is generated for of each of such situations. Actual values used for the generating mechanism follow, where D stands for doomed and S for sensitive group:

1. $\beta_0 - \mu_0 = 1$

$$Y_D \sim N(3, 0.5)$$
$$Y_S \sim N(2 + 0.053X, 0.5)$$

2. $\beta_0 - \mu_0 = 2$

$$Y_D \sim N(4, 0.5)$$
$$Y_S \sim N(2 + 0.053X, 0.5)$$

3. $\beta_0 - \mu_0 = 3$

$$Y_D \sim N(5, 0.5)$$
$$Y_S \sim N(2 + 0.053X, 0.5)$$

Table 7.11 reports the mean correlation coefficient and the mean residual variance over the 2000 simulation runs for each of the models under the nine different sub-scenarios. Once again, the model employing the quadratic logistic selection mechanism appears to be the best fitting one, and the same

considerations as for Scenario I. analysis apply; it is interesting to note that, as the distance between the groups increases, our models yeld improving performances (i.e. the "fork" broadens), confirming a better functioning of the selection model in presence of a more marked distinction between doomed and sensitives.

Distance= 1	**Correlation**			**Residual Variance**		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.60869	0.61594	0.57205	0.40869	0.39594	0.43205
0.1%	0.67168	0.67847	0.64094	0.37166	0.35896	0.38683
0.01%	0.33338	0.33499	0.28825	0.26395	0.26355	0.27101
Distance= 2	**Correlation**			**Residual Variance**		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.67984	0.68451	0.59088	0.91994	0.91282	0.98500
0.1%	0.74160	0.75571	0.68038	0.75311	0.68222	0.79262
0.01%	0.44527	0.44578	0.36795	0.29358	0.29228	0.30969
Distance= 3	**Correlation**			**Residual Variance**		
% of Sensitive	**Logistic**	**Q.Logistic**	**Naïve**	**Logistic**	**Q.Logistic**	**Naïve**
1%	0.70944	0.71905	0.61348	1.62333	1.58392	1.78042
0.1%	0.78264	0.78310	0.71633	1.29746	1.19909	1.39495
0.01%	0.54228	0.56238	0.45431	0.32333	0.30804	0.34459

Table 7.11: Mean Correlations and Mean Residual Variances

Nine plots follow, every of them showing one randomly chosen simulated sample among those for each sub-scenario they represent, together with the related fitted model with quadratic logistic selection mechanism (in green) and the regression line from a naïve analysis (in red); on the x-axis we consider HRT duration in thousands of days (about 3 years time) while the y-axis represents the prognostic index.

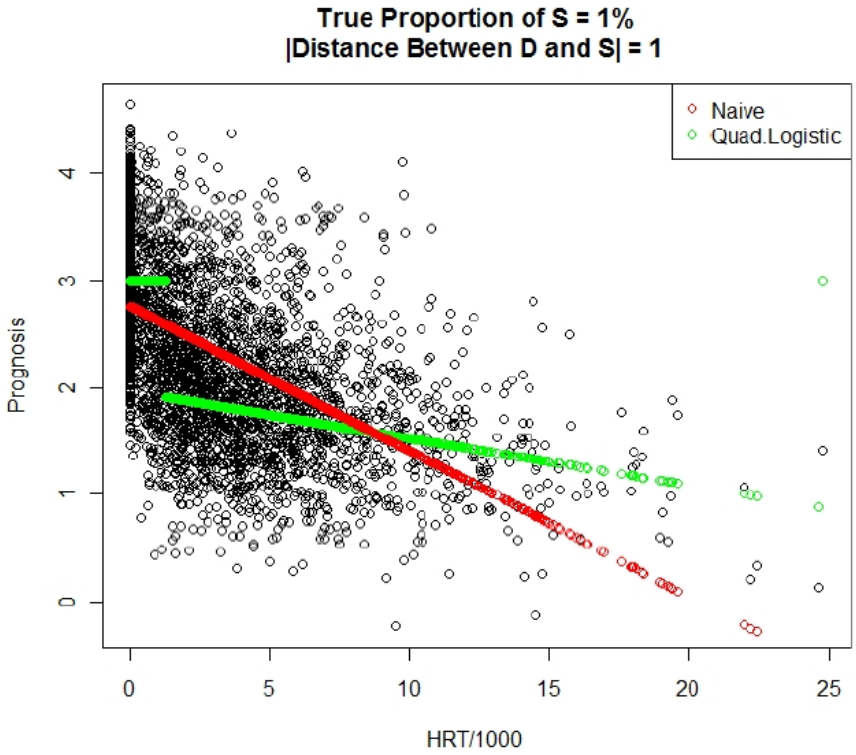

Figure 7.15: Plot for Sub-Scenario with Distance= 1, $S = 1\%$

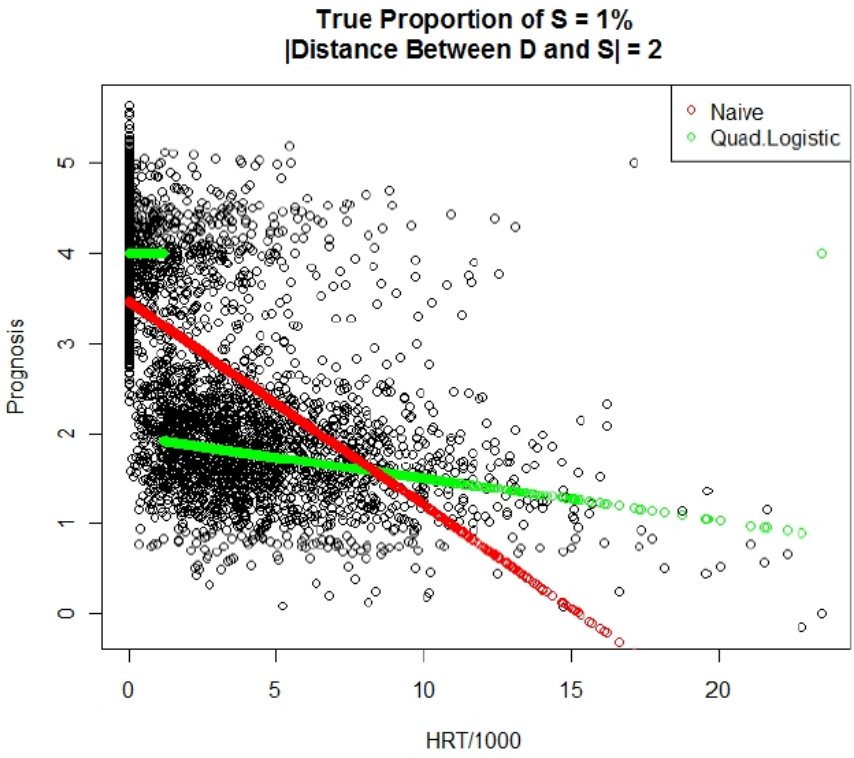

Figure 7.16: Plot for Sub-Scenario with Distance= 2, $S = 1\%$

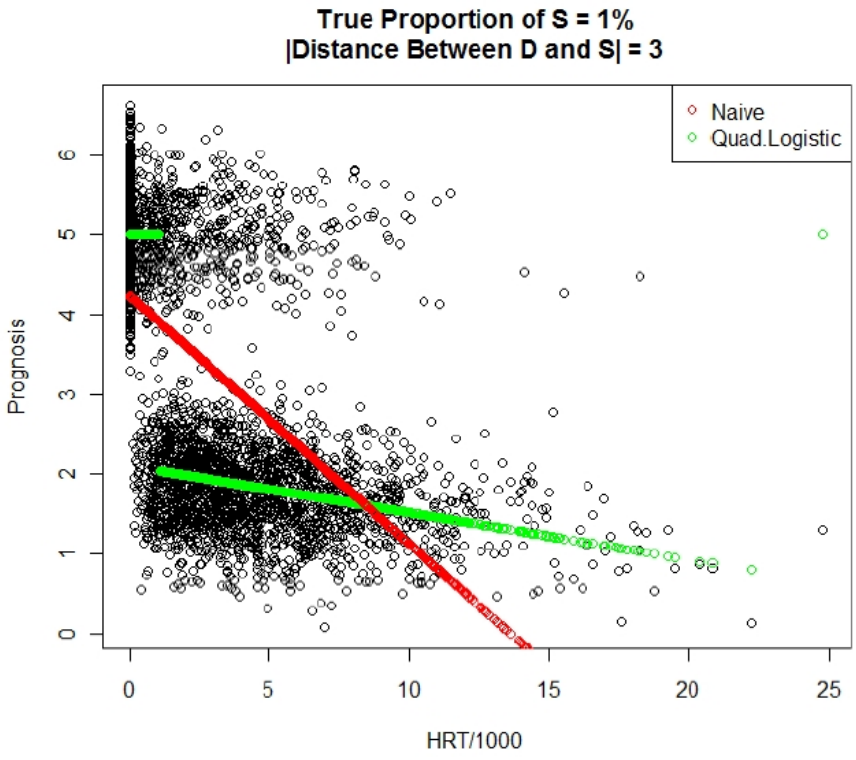

Figure 7.17: Plot for Sub-Scenario with Distance= 3, $S = 1\%$

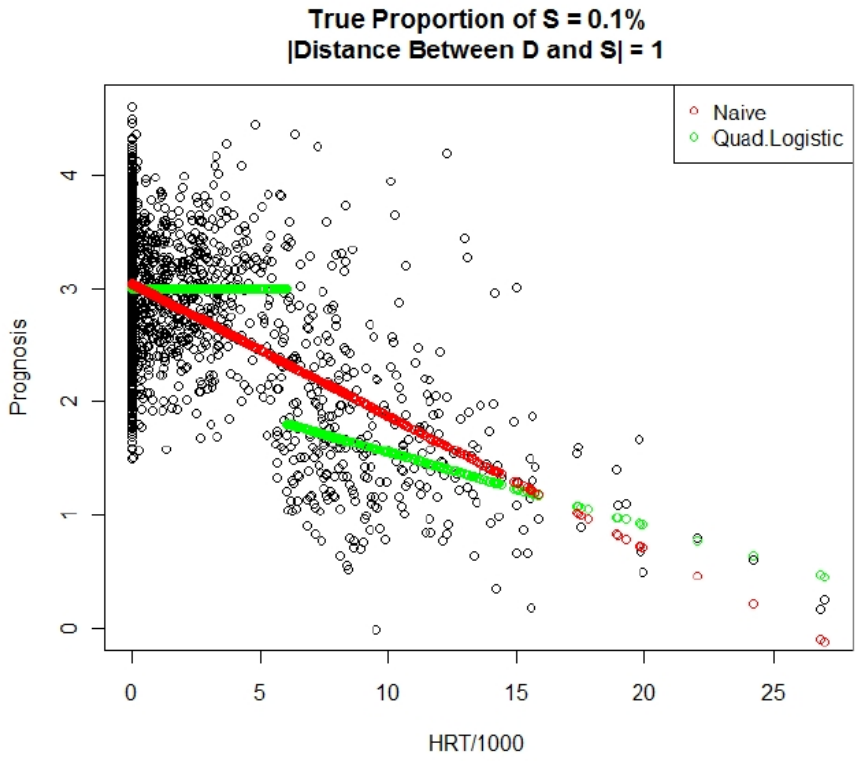

Figure 7.18: Plot for Sub-Scenario with Distance= 1, $S = 0.1\%$

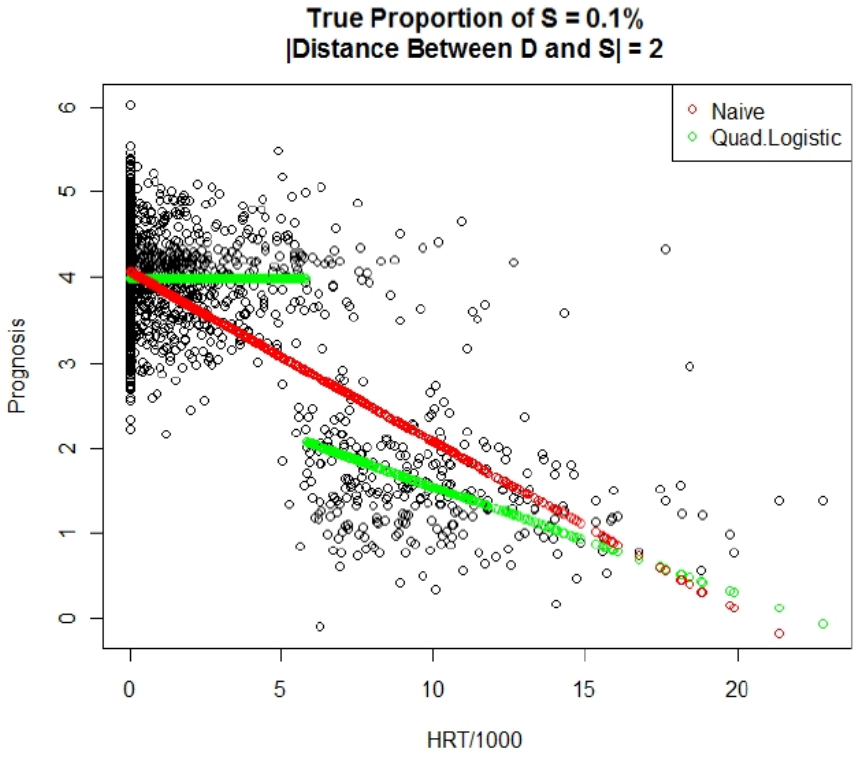

Figure 7.19: Plot for Sub-Scenario with Distance= 2, $S = 0.1\%$

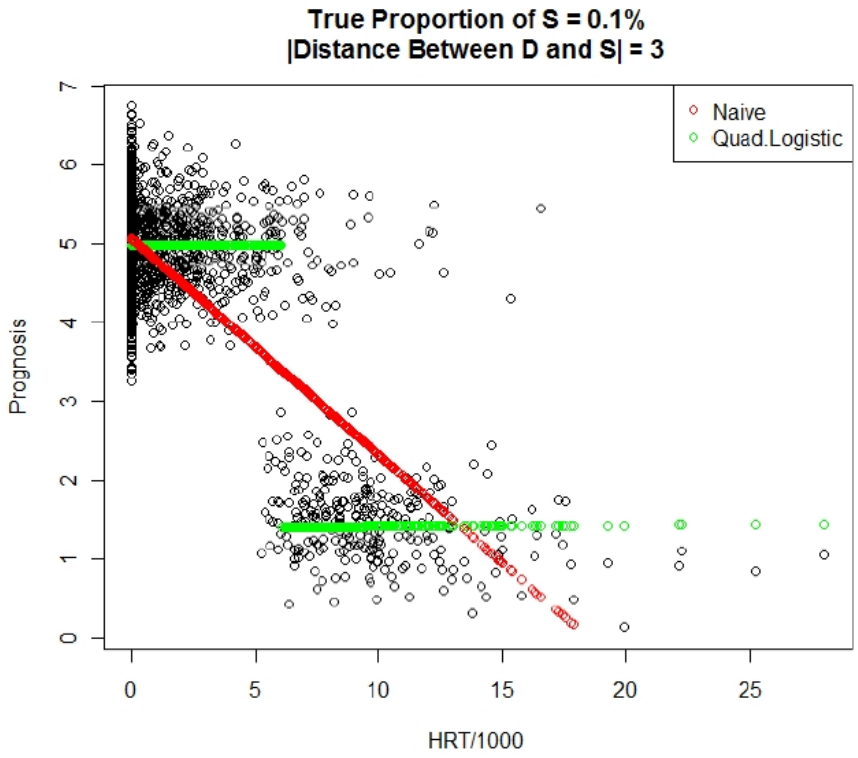

Figure 7.20: Plot for Sub-Scenario with Distance= 3, $S = 0.1\%$

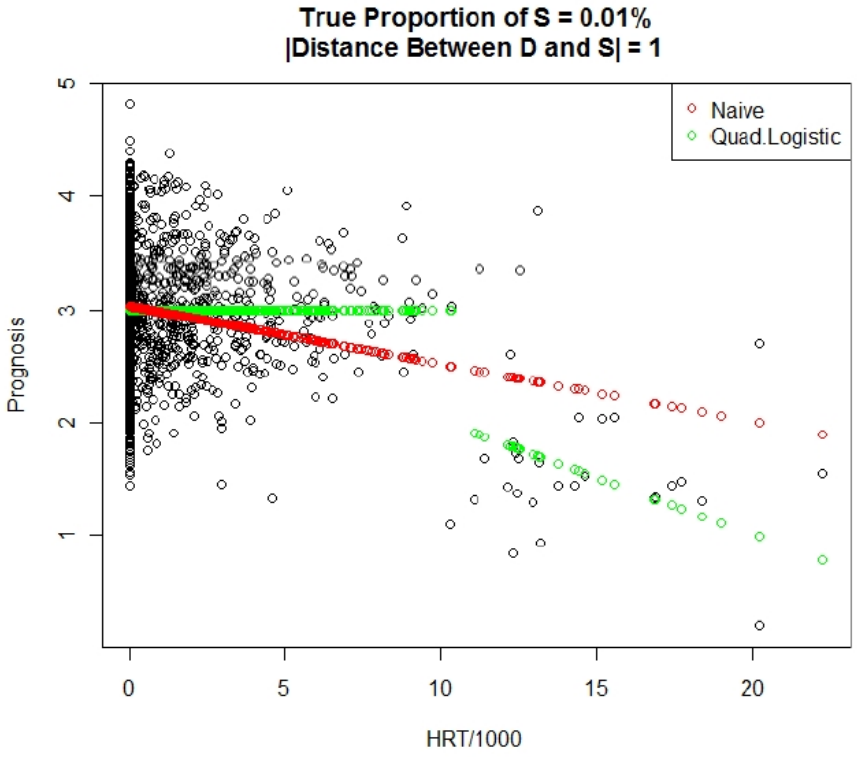

Figure 7.21: Plot for Sub-Scenario with Distance= 1, $S = 0.01\%$

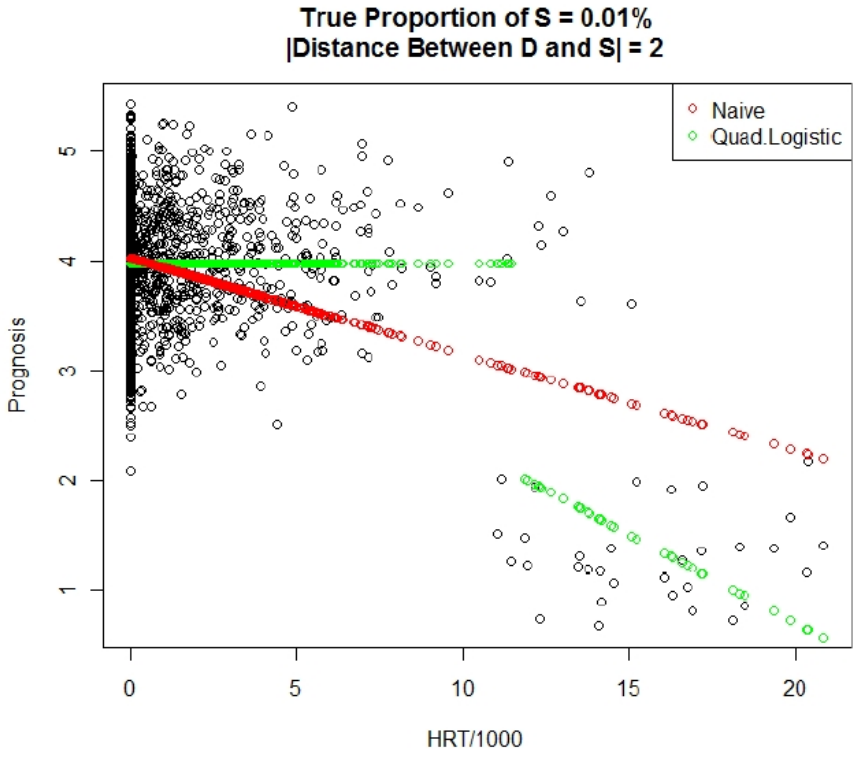

Figure 7.22: Plot for Sub-Scenario with Distance= 2, $S = 0.01\%$

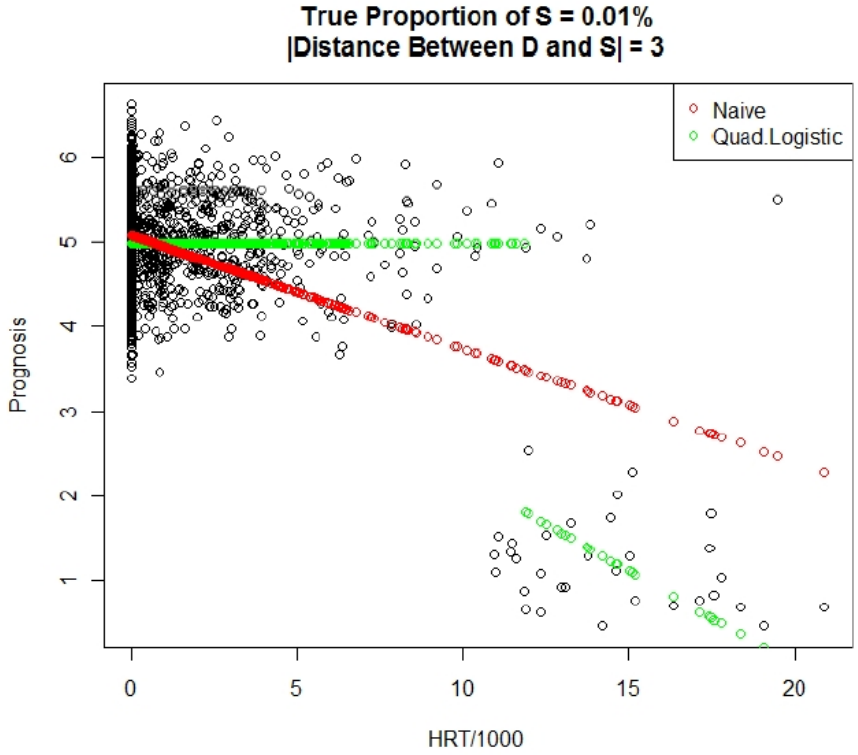

Figure 7.23: Plot for Sub-Scenario with Distance= 3, $S = 0.01\%$

Chapter 8

Conclusions

In my thesis I have used mixture modelling and clustering techniques within the framework of principal stratification in order to assess the possible causal link between Hormone Replacement Therapy (HRT) and Breast Cancer. HRT has been a widely prescribed medical treatment to relieve post-menopausal symptoms and protect women against heart diseases and osteoporosis during the last decades; by 2001, 15 million women were under annual treatment. Starting from 1998, with a clinical trial called HERS (Heart and Estrogen-progestin Replacement Study), faith in the protective power of HRT began to fade, since the study concluded that estrogen therapy increased, rather then decreased, the likelihood that women who had heart disease already would suffer a heart attack; in 2002, with the W.H.I. study (Women's Health Initiative) evidence was found that HRT constituted a strong risk factor for breast cancer (see [1]). While HRT might protect women against post-menopausal symptoms (osteoporosis and colorectal cancer among the worsts), these benefits are definitely outweighted by increased risks of heart disease, stroke, blood clots, breast cancer and perhaps even dementia. The New England Journal of Medicine, though, assessed that HRT may indeed protect women against heart disease if they begin taking early after menopause, but it is still decidedly deleterious for those women who begin later in life. The question of how many women have died prematurely or suffered strokes or breast cancer while they were taking a pill that their physicians had prescribed to protect them against heart disease remains still unanswered. There is therefore a need to detect causal mechanisms.

My work aims at providing an extension to the methods proposed by Sjölander [10], and is closely related to the works by Frangakis and Rubin [2] and Gilbert, Bosch and Hudgens [3]. As in [10], we focus on a continuous exposure (HRT duration) rather then on a binary treatment variable (as usually is in literature on this topic) and develop a modelling setting aimed at estimating the causal effect of the treatment on an outcome of interest (breast

cancer prognosis or survival), in observational studies, such as a case-control design. The main difference between our approach and Sjölander's lies on the fact that we do not use semi-parametric estimating methods, but rather directly parametrize potential outcomes in the two groups of women defined as "doomed" and "sensitive", i.e. those women who experience a cancer which is not HRT-induced and those who wouldn't have developed the tumor if not treated; this allows to use maximum likelihood estimation techniques and grants more modelling flexibility than the semi-parametric approach. Moreover, as opposed to [10], it could be possible with our approach to find actual causal effects (at least locally, see Section 5.3), even if at the expense of a somewhat heavier estimating procedure; greater flexibility may imply more parameters and a possibly more complex structure for the model (which could, however, reflect a deeper knowledge of the phaenomenon of interest). Maximum likelihood estimation also allows for a straightforward parametric assumptions testing.

Applications to real data from the CAHRES study (Chapter 6) deal with both a continuous (the Nottingham Prognostic Index, NPI) and a discrete (5-years survival) outcome. Comparison of the proposed model against standard regression techniques is made through goodness of fit measures and graphic representations for the NPI outcome, and lead to a very small improvement in fit using our method, the main problem being that the discriminatory power of the logistic selection model is not sufficient to assign women to different groups (assuming such distinction actually exists). Estimated covariates (age, BMI) effect on prognosis are comparable to those obtained through a standard linear regression. As for the discrete outcome, better results for our model are obtained, possibly thank to a larger complete-data set and a greater number of additional covariates (BMI, age at menarche, age class, parity). In order to compare our model with an ordinary logistic regression in terms of performance we predict the outcome (5-years survival) using both models, build the corresponding contingency tables of fitted against observed values, and calculate sensitivity and specificity values for each such table. Our method results in a better predictive power in terms of both indexes.

An important part of my work is devoted to simulation study (Chapter 7): various analyses have been made to compare our proposal to Sjölander's and to the classic analysis approach (mainly against standard regression models), in order to gain a deep insight of when and how results could be wrong and misleading when the model in use is not correctly specified. In the first part (Section 7.3) models for difference in prognosis in the two groups are compared in terms of bias in parameters' estimates through analysis of such estimates' sample distributions. Evidence is found that particular attention on the logistic selection model helps to improve correct targeting

of the true parameters' values, and that Sjölander's implicit model yelds estimates which are of difficult interpretation, biologically speaking. The second part (Section 7.4) dealt with assessing the goodness of fit of our model as compared to a standard linear regression for a continuous outcome under several possible scenarios; there is evidence of an overall slightly better fit in all the situations we considered, except for the one where our model is completely wrong with respect to the underlying truth (and even in this case, the performance is not so far from that of the right modelling structure).
Possible future extensions of the work may include the re-definition of the principal strata not in dependence of HRT duration, so not to have to deal with only-locally-causal effects, and the implementation of different estimation methods, such as the EM algorithm. Also considering the possibility of letting covariates to affect in the same way the two groups of women may lead to a somewhat simpler modelling structure, allowing for cleansing the outcome from their effect, focusing, for example, on residuals only.

In any case, to better address the causality issue, there is need for more data, possibly genetic-specific data of the patients, given the nature of the clustering we have considered, which we think could be almost entirely related to gene configuration. If this were the case, then refining our model, keeping into account such genetic information, could lead to something more sensible than what we found at the present stage of the work, i.e. quite a small difference in terms of model fit as compared with the standard approach (see Chapter 6).

Bibliography

[1] Caroline Antoine, Fabienne Liebens, Birgit Carly, Anne Pastijn, and Serge Rozenberg. Influence of hrt on prognostic factors for breast cancer: a systematic review after the women's health initiative trial. *Human Reproduction*, 19(3):741–756, January 2004.

[2] Constantine E. Frangakis and Donald B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, March 2002.

[3] Peter B. Gilbert, Ronald J. Bosch, and Michael G. Hudgens. Sensitivity analysis for the assessment of causal vaccine effects on viral load in hiv vaccine trials. *Biometrics*, 59(3):531–541, September 2003.

[4] Michael G. Hudgens, Antje Hoering, and Steven G. Self. On the analysis of viral load endpoints in hiv vaccine trials. *Statistics in Medicine*, 2003.

[5] Jerzy Neyman. On the application of probability theory to agricultural experiments: Essay on principles, section 9. *Translated in Statistical Science*, (5):465–480, 1990.

[6] Judea Pearl. *Causality: models, reasoning and inference.* Cambridge University Press, 2000.

[7] Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, (66):688–701, 1974.

[8] Donald B. Rubin. Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics*, (2):1–26, 1977.

[9] Donald B. Rubin. Bayesian inference for causal effects. *Annals of Statistics*, (6):34–58, 1978.

[10] Arvid Sjölander et al. *Causal Inference in Epidemiological Research.* PhD thesis, Karolinska Institutet, 2009.