University of Oxford

Some aspects of complex statistical dependencies

Christiana Kartsonaki

Nuffield College

Thesis submitted for the degree of Doctor of Philosophy in Statistics.

Department of Statistics, 1 South Parks Road,
Oxford OX1 3TG

June 2014

Some aspects of complex statistical dependencies

Christiana Kartsonaki

Nuffield College

Thesis submitted for the degree of Doctor of Philosophy in Statistics.

Hilary Term 2014

Abstract

In the first part parametric models for which the likelihood is intractable are discussed. A method for fitting such models when simulation from the model is possible is presented, which gives estimates that are linear functions of a possibly large set of candidate features. A combination of simulations based on a fractional design and sets of discriminant analyses is used to find an optimal estimate of the parameter vector and its covariance matrix. The procedure is an alternative to Approximate Bayesian Computation and Indirect Inference methods. A way of assessing goodness of fit is briefly described.

In the second part the aim is to give a relationship between the effect of one or more explanatory variables on the response when adjusting for an intermediate variable and when not. This relationship is examined mainly for the cases in which the response depends on the two variables via a logistic regression or a proportional hazards model. Some of the theoretical results are illustrated using a set of data on prostate cancer. Then matched pairs with binary outcomes are discussed, for which two methods of analysis are

described and compared.

Acknowledgements

I am mostly grateful to David Cox, who is probably the best supervisor in the world. He has offered to me invaluable help and support, beyond what one could expect. I am grateful for his continuous support, guidance and encouragement. He has made the process very pleasant and has shown to me how enjoyable doing interesting work can be.

I would like to thank Lorenzo Richiardi (University of Turin, Italy) for providing the data of which the analysis is described in Chapter 5.

I would like to thank Tom Snijders for helpful comments. I would also like to thank Lucy Carpenter for kindly offering her help and support. I am also grateful to Antonis Antoniou because the experience I gained during the year before starting to work on this thesis has undoubtedly been very helpful. I would also like to thank Michelle Jackson who encouraged me to do this.

I am grateful to my parents for making me the type of person who wants and is able to do this and for their continuous support. I am also grateful to Savvas for being there for me. Finally I would like to thank my friends who encouraged me to do this and who made this time even more pleasant.

Contents

Part I

The fitting of complex parametric models

Chapter 1

Background

1.1 Introduction

A first step in the formal analysis of a parametric statistical model typically involves the likelihood function. Some models, however, are too complicated for the likelihood to be available in useful form and then resort may be had first to the Generalized Method of Moments. That is, with, say, p unknown parameters, q features of the data, with $q \geq p$, all judged to depend sensitively on the parameter θ, are chosen, for all of which analytical formulae can be obtained for the expectations of the chosen features. If $q = p$ these are equated to the corresponding observed values, and the resulting nonlinear equations solved for the parameter estimates, hopefully finding a unique solution in the range of sensible values. If $q > p$, there are a number of possibilities, one being use of generalized least squares fitting with a formal covariance matrix, if necessary found by simulation. Another possibility is to choose p of the features for fitting, with the remaining $q - p$ used to assess model adequacy, quite possibly trying several choices to reach a judgemental compromise.

For example, to fit a five-parameter model of hourly rainfall, Rodriguez-Iturbe et al. (1987) used data at hourly, and aggregated at six-hourly and twelve-hourly periods, considering the mean, variance, skewness and lag one correlation of rainfall at each level and the proportions of zero values. Parameters were estimated in various ways, aiming to produce good fit for all 13 features considered. In this, and perhaps in other examples, the features used were not only routes to parameter estimation but also were of intrinsic interest, so that reasonable agreement between all observed and model values was desirable.

Jiang and Turnbull (2004) reviewed what they termed methods of indirect inference, hinging on the idea of a bridge or binding relation linking the distribution of empirically chosen features to the parameters in the underlying assumed model. Chapter 2 of the thesis is essentially concerned with how to establish that bridge. Their review included important work with a strong econometric emphasis, including various generalizations of the method of moments and the contributions of Gouriéroux and Monfort (1993) and Gouriéroux et al. (1993). In particular, the latter authors exploited possible connections with simpler models to guide the choice of relevant features.

A different approach, again for very complicated models where analytical evaluation of features is not possible, is the Approximate Bayesian Computation (ABC) scheme (Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002), extensively used both in genetical applications (Marjoram et al., 2003), as reviewed by Beaumont (2010), and in systems biology (Toni et al., 2009). In the simplest form of this, values of θ are chosen from a prior distribution, simulated data obtained from the model, and a suitable distance function used to examine consistency between the data and the simulated values. If there is reasonable consistency, the corresponding value of θ is included in the posterior region. The procedure is then repeated many times. In these methods the best way of choosing the statistics on which to base the procedure and the function for examining consistency is not clear.

In Chapter 2 a new method for fitting parametric models for which the likelihood is intractable is described (Cox and Kartsonaki, 2012). In this method a combination of a fractionally replicated design and sets of discriminant analyses is used to find an estimate of the parameter vector, together with its covariance matrix, the estimate being optimal within a specified family of estimates.

Chapters 1 and 2 of the thesis are based partly on a published paper (Cox and Kartsonaki, 2012) and partly on an extension of that work. The remainder of this chapter describes existing likelihood-free methods for complex parametric models. In Section 1.2, the ABC algorithm is discussed and in Section 1.3 Indirect Inference methods are described.

1.2 Approximate Bayesian Computation (ABC)

1.2.1 Introduction

Approximate Bayesian Computation (ABC) (Tavaré et al., 1997; Pritchard et al., 1999; Beaumont et al., 2002) is a method aiming to infer the posterior distribution when the likelihood function is intractable. It uses a simulation-based procedure and replaces the calculation of the likelihood by a comparison between the observed and simulated data.

Let θ be a parameter vector to be estimated. If $\pi(\theta)$ is the prior distribution and $f(\mathcal{D} \mid \theta)$ the likelihood of the data $\mathcal{D}$ given θ, which cannot be calculated, the aim is to approximate the posterior distribution $\pi(\theta \mid \mathcal{D}) \propto f(\mathcal{D} \mid \theta)\pi(\theta)$. The output is a sample of θ values from an approximation to the posterior.

A simple version of the ABC algorithm is as follows:

- Draw θ from $\pi(\theta)$.
- Simulate $\mathcal{D}' \sim f(\cdot \mid \theta)$.
- Accept θ if $\rho(\mathcal{D}, \mathcal{D}') \leq \epsilon$, otherwise reject.

Here ρ is a distance function used to compare the simulated with the observed data and $\epsilon > 0$ the tolerance level, that is the desired level of agreement between the simulated and observed data.

This generates observations from $\pi(\theta \mid \rho(\mathcal{D}, \mathcal{D}') < \epsilon)$. If ρ is the Euclidean distance, as ϵ tends to infinity, we get observations from the prior, $\pi(\theta)$, and in the limit as ϵ tends to zero, we generate observations from the posterior, $\pi(\theta \mid \mathcal{D})$. So the tolerance ϵ reflects the trade-off between computability and accuracy.

Another version of the ABC algorithm uses summary statistics $\mathcal{S}(\mathcal{D})$, to reduce the dimension, but also introducing a further approximation. The algorithm for the version with summary statistics is as follows:

- Draw θ from $\pi(\theta)$.
- Simulate $\mathcal{D}' \sim f(\cdot \mid \theta)$.
- Accept θ if $\rho(\mathcal{S}(\mathcal{D}), \mathcal{S}(\mathcal{D}')) \leq \epsilon$.

If $\mathcal{S}$ is sufficient, this is equivalent to the previous algorithm, since sufficiency implies that all information in $\mathcal{D}$ about θ is captured by $\mathcal{S}(\mathcal{D})$. In practice it is not usually possible to find a set of sufficient statistics; hence a set of informative non-sufficient statistics is used instead.

Diggle and Gratton (1984) and Rubin (1984) were the first to introduce ABC-related ideas, although the use of simulation for inference was introduced much earlier (Student, 1908a, 1908b; Ross, 1972). Diggle and Gratton (1984) developed a method of inference for models whose distribution theory is intractable, not identical to what is now known as ABC. They used kernel density estimation for estimating a log-likelihood from simulations of such models and developed an algorithm for maximizing this estimated log-likelihood function. More specifically, they suggested using a systematic simulation scheme to approximate the likelihood function in situations where its analytic form is intractable. Their method was based on defining a grid in the parameter space and using it to approximate the likelihood by running several simulations for each grid point. The approximation was then improved by applying smoothing techniques to the outcomes of the simulations.

Rubin (1984), when discussing the interpretation of Bayesian statements, pointed out that datasets simulated from a model under a range of parameter values can be used to assess likelihoods without explicit evaluation and described a hypothetical sampling mechanism that yields a sample from the posterior distribution. This sampling mechanism coincides with that of the ABC rejection scheme.

Another algorithm which produces a set of samples from a posterior distribution when it is not analytically tractable was proposed by Gordon et al. (1993).

Tavaré et al. (1997) were the first to propose an ABC algorithm. This was followed by Pritchard et al. (1999), and the term ABC was established by Beaumont et al. (2002), who extended the ABC methodology. See Beaumont (2010) and Sunnåker et al. (2013) for reviews.

Various modifications of ABC have been proposed, such as ABC with MCMC (Marjoram et al., 2003) and ABC with a sequential Monte Carlo sampler (Sisson et al., 2007). The ABC framework can be used for model selection, by estimating the posterior probabilities of different candidate models (Grelaud et al., 2009; Toni et al., 2010), but this can be problematic (Robert et al., 2011).

ABC introduces two approximations to the posterior distribution. The first is that the posterior distribution of the full dataset, $\pi(\theta \mid \mathcal{D})$, is approximated by $\pi(\theta \mid s_{\mathcal{D}}) \propto f(s_{\mathcal{D}} \mid \theta)\pi(\theta)$, where $s_{\mathcal{D}} = \mathcal{S}(\mathcal{D})$ is a vector of summary statistics of lower dimension than the data $\mathcal{D}$. Thus for $\pi(\theta \mid s_{\mathcal{D}})$ to be a good approximation to $\pi(\theta \mid \mathcal{D})$, the chosen statistics must be highly informative for the model parameters. The second approximation is that as $f(s_{\mathcal{D}} \mid \theta)$ is also likely to be computationally intractable if $f(\mathcal{D} \mid \theta)$ is, the ABC posterior is obtained by a finite number of simulations and the agreement between the values of the simulated statistics and $s_{\mathcal{D}}$ is determined by the tolerance ϵ or equivalently by a smoothing kernel which depends on ϵ.

There is a trade-off between the two approximations: if the dimension of $\mathcal{S}$ is large so that the first approximation $\pi(\theta \mid s_{\mathcal{D}}) \simeq \pi(\theta \mid \mathcal{D})$ is good, the second approximation may then be poor due to the inefficiency of kernel smoothing in large dimensions. Conversely, if the dimension of $\mathcal{S}$ is small and the second approximation is good (with small ϵ), any loss of information in using the summary statistics instead of the data means that the first approximation is inadequate. Thus ideally a low-dimensional and near-sufficient statistic $\mathcal{S}$ should be chosen. There are several methods which attempt to achieve this, most of which are based on dimension reduction ideas.

1.2.2 Choice of distance function and tolerance

The choice of the distance function ρ and the tolerance ϵ cannot be made based on a general rule. There are studies of the difference between the ABC posterior and the true posterior as a function of the tolerance ϵ (Sisson et al., 2007) and theoretical results for an upper bound for the error in the parameter estimates which is a function of ϵ (Dean et al., 2011). The convergence of the distributions when ϵ approaches zero, and how it depends on the distance measure used, is an important topic that has yet to be investigated in greater detail. In particular, it is difficult to distinguish errors due to the approximation from those due to poor model choice (Beaumont, 2010).

The choice of ϵ affects the accuracy of the ABC posterior $\pi_{\text{ABC}}(\theta \mid s_{\mathcal{D}})$ in approximating $\pi(\theta \mid s_{\mathcal{D}})$, as well as the average acceptance probability and hence the simulation error (Fearnhead and Prangle, 2012).

Beaumont et al. (2002) suggest correcting the output of ABC by using local linear weighted regression to weaken the effect of the discrepancy of the simulated statistics and $s_{\mathcal{D}}$, in order to reduce the error between $\pi_{\text{ABC}}(\theta \mid s_{\mathcal{D}})$ and $\pi(\theta \mid s_{\mathcal{D}})$.

Fearnhead and Prangle (2012) investigated the accuracy of the posterior, measured by the expected quadratic loss, as a function of ϵ. They also proposed 'noisy ABC', a version of ABC in which a specific form of noise is introduced to the summary statistics to characterize and compensate for the bias caused by a non-zero tolerance. They also showed that asymptotically as $\epsilon \to 0$, ABC gives estimates that are at least as accurate as or more accurate than any other estimator based on the same summary statistics.

1.2.3 Choice of summary statistics

The choice of summary statistics in ABC is very important, as a poor choice can have a big effect on both approximations involved.

One approach aiming to capture most of the information present in the data would be to use many statistics, but the accuracy and stability of ABC appears to decrease rapidly as the number of summary statistics increases (Beaumont, 2010; Csilléry et al., 2010). The more summary statistics, the more difficult it is to match the observations closely ('curse of dimensionality').

There are methods to guide the selection of a subset of statistics from a large set of candidate statistics (Joyce and Marjoram, 2008; Nunes and Balding, 2010). Joyce and Marjoram (2008) propose a method to choose approximately sufficient statistics. However the definition of such statistics is not clear and it is difficult to construct a general method for finding such statistics. Another approach is to weight the statistics appropriately, using a partial least squares regression (Wegmann et al., 2009).

Semi-automatic ABC (Fearnhead and Prangle, 2012) is an algorithm which aims to guide the choice of summary statistics to be used in ABC. They show that the optimal summary statistics, with respect to minimizing the quadratic loss of the parameter point estimates, are the posterior means of the parameters, which are unknown. The idea is that π_{ABC} should be aimed to be a good approximation only in terms of the accuracy of certain estimates of the parameters. The algorithm is as follows:

1. Use a pilot run of ABC to determine a region of non-negligible posterior mass.
2. Simulate sets of parameter values and data.
3. Use the simulated values and data to estimate the summary statistics, by linear regression with appropriate functions of data as explanatory variables.
4. Run ABC with this choice of summary statistics.

For step (3), let $f(\mathbf{y})$ be a vector of (possibly nonlinear) transformations of the data, the simplest choice being $f(\mathbf{y}) = \mathbf{y}$. For the i^{th} summary statistic, the responses are the simulated values of the i^{th} parameter, $\theta_i^{(1)}, \ldots, \theta_i^{(M)}$, where M is the number of datasets, and explanatory variables $f(\mathbf{y}^{(1)})$, $\ldots$, $f(\mathbf{y}^{(M)})$. The model $\theta_i = E(\theta_i \mid \mathbf{y}) + \epsilon_i = \beta_0^{(i)} + \beta_1^{(i)} f(\mathbf{y}) + \epsilon_i$ is fitted by least squares. Then the i^{th} summary statistic is $\hat{\beta}_1^{(i)} f(\mathbf{y})$. Then standard model comparison procedures can be used to choose between summary statistics that are obtained from linear regressions using different explanatory variables.

See Blum et al. (2013) for a review and comparison of the performance of the methods of dimension reduction for summary statistics in ABC. Blum et al. (2013) also proposed two additional techniques for dimension reduction.

1.2.4 Using ABC with large datasets

When the number of observations in the data is large but the number of parameters is small, the only issue is the cost of simulation, otherwise ABC is expected to work well. However if the number of parameters is also large, it might be a problem as the number of summary statistics will also have to be large.

If simulating data or calculating summaries for large n is slow, one possibility is to resort to subsampling, the simplest approach being to use a fixed subset of the data, which is likely to be unsatisfactory, or alternatively to use different random sub-samples at each iteration, which is preferable.

Buzbas and Rosenberg (2013) proposed Approximate Approximate Bayesian Computation (AABC), a method which extends ABC to models in which simulating data is expensive. They first simulate a computationally feasible number of datasets and then use these datasets as fixed background information to inform a non-mechanistic statistical model, based on the empirical distributions of the data simulated from the correct model, which approximates the correct parametric model and enables efficient simulation of a large number of datasets by Bayesian resampling methods. This involves further approximations in addition to the other approximations involved in ABC.

High-dimensional datasets and high-dimensional parameter spaces can require an extremely large number of parameter points to be simulated to obtain a reasonable level of accuracy for the posterior inferences. Analytical formulae have been derived for the error of the ABC estimators as functions of the dimension of the summary statistics (Blum, 2010; Fearnhead and Prangle, 2012).

The scheme proposed by Fearnhead and Prangle (2012) projects the data into estimates of the parameter posterior means, which are used as summary statistics and have the same dimension as the parameters.

How to make ABC practically feasible for problems involving high-dimensional target parameter spaces is currently an open issue.

1.2.5 Discussion

The choice of tolerance, distance function, summary statistics, number of simulations or competing models cannot currently be based on general rules.

Also, no general way of testing model adequacy exists. However, Ratmann et al. (2009) proposed an algorithm which explicitly accounts for discrepancies between the model and the data, termed ABC under model uncertainty. They augment the likelihood of the data with unknown error terms that correspond to chosen checking functions and provide Monte Carlo methods for sampling from the associated joint posterior distribution.

1.3 Indirect Inference

Indirect inference methods (Gouriéroux and Monfort, 1993; Jiang and Turnbull, 2004) estimate 'auxiliary' parameters β under a tractable model and map these to an estimate of θ via simulation, where θ denotes the parameter of the model from which the data were assumed to be generated. That is, an 'auxiliary model' is introduced, which is misspecified and typically not even generative, but can easily be fitted to the data. This has parameter vector β, with estimator $\hat{\beta}$. These auxiliary parameters aim to capture aspects of the empirical distribution of the data. The parameters θ of the data-generating model are then estimated by trying to match the auxiliary parameters.

Suppose that we have data x and auxiliary parameter estimates $\hat{\beta} \equiv \hat{\beta}(x)$. For each θ we can generate a simulated $\tilde{x}(\theta)$ of the same size as the data, leading to auxiliary estimates $\tilde{\beta}(\theta) \equiv \hat{\beta}(\tilde{x}(\theta))$. The indirect inference estimate $\hat{\theta}$ is the value of θ for which $\tilde{\beta}(\theta)$ is as close to $\hat{\beta}$ as possible. More generally, a symmetric, positive-definite matrix W is used and the quadratic form $\left\{\hat{\beta} - \tilde{\beta}(\theta)\right\}^{\mathrm{T}} W \left\{\hat{\beta} - \tilde{\beta}(\theta)\right\}$ is minimized, with the entries of the matrix chosen to give more or less relative weight to the different auxiliary parameters. That is, the idea is to minimize (in θ) a distance between estimators $\hat{\beta}$ of parameters β of a pseudo-model for the data and for observations simulated under the true model and the parameter θ.

Drovandi et al. (2011) propose combining ABC with indirect inference, as a way to deal with the issue of choosing summary statistics in ABC. In particular, they propose choosing the summary statistics in ABC to be the estimators of the auxiliary parameters of indirect inference.

Alternative approaches include the Method of Simulated Moments (MSM) (McFadden, 1989), which is a special case of indirect inference, and the synthetic likelihood (Wood, 2010).

1.4 Discussion

In Chapter 2 a new method is described in which a combination of a fractionally replicated design and sets of discriminant analyses is used to find an estimate of the parameter vector, together with its covariance matrix, the estimate being optimal within a specified family of estimates.

The distinctive features of the proposed method are the use of a fractionally replicated design based on Hadamard matrices to guide the simulations, combined with sets of discriminant analyses, one for each component parameter. These lead by linear interpolation between values at the known design points to explicit estimates of the components of the parameter vector, together with the covariance matrix of the estimates. From this, approximate confidence intervals or regions for components or sets of components follow. The justification of these estimates broadly parallels the first-order asymptotic theory of maximum likelihood estimates, with optimal linear combinations of the assumed features playing the role of components of the score function. That is, the proposed procedures are optimal among all those for which dependence on the underlying parameter value is locally linear.

Chapter 2

Fitting parametric models with intractable likelihoods using summary statistics

2.1 Introduction

Suppose we have a set of data $\mathcal{D}$ which are assumed to have been generated from a model that can be simulated and we want to estimate its parameters. Starting from some initial parameter values and suitable spacing for each, design points are specified, at which we simulate from the model. The design is specified by columns of a Hadamard matrix of suitable dimensions.

A set of suitable features, i.e. summary statistics, of the distribution is chosen, the values of which are averaged over all simulation runs. Then a linear combination of the differences of those statistics evaluated at different design points is formed. The method of Lagrange multipliers is used to find the coefficients that maximize this linear contrast, subject to a given variance, or equivalently to minimize its variance subject to a given separation. Then the corresponding linear combination of the features of the data is formed and the parameters are estimated by linear interpolation between the values of this linear combination at different design points. The covariance matrix of the estimates can be calculated.

As the starting parameter value might be far from the true one, the procedure can be run iteratively, by re-centering the parameter values at the estimates of the previous step, and possibly by adjusting the spacings according to the standard errors of the parameter estimates. The procedure has been implemented in the R package 'ssfit' (Kartsonaki, 2013).

The present chapter is an extended version of a published paper (Cox and Kartsonaki, 2012). In Section 2.2 a simple example with a one-dimensional parameter is presented for illustration. Then a general formulation of the method is given in Section 2.3 and more complicated examples are presented in Sections 2.4, 2.5 and 2.6. In Section 2.7 various open issues are outlined.

2.2 A simple example

2.2.1 One-dimensional case

We first consider a very simple example to illustrate the arguments involved. Suppose we have a distribution with an unknown scalar parameter θ, and that we can simulate observations from this distribution. We suppose that a set of data is available for analysis and we select q features as the basis of the estimation procedure. We choose an initial value θ_0 of the parameter and a suitable displacement h. This specifies two design points, namely $\theta_0 \pm h$. Then a large number r of simulations are run from the distribution at each of these two parameter values, typically but not necessarily with sample size the same as that of the data to be analyzed. For each simulation run the q features are calculated.

We form a $q \times r$ matrix Z_i for each design point i with the simulated values of q features of the distribution. Let $\bar{Z}$ be the $q \times 2$ matrix of the values of the q features averaged over all simulation runs at each design point. Postmultiplication of $\bar{Z}$ by a 2×1 contrast matrix gives a $q \times 1$ column vector $\bar{V}$ determining the differences between the means of the q features at the upper and lower levels of θ. The covariance matrices of the individual simulated values Z at each of the two points are calculated from the simulated values; we use their average Σ. Let $Z_{\mathcal{D}}$ be a $q \times 1$ vector containing the values of the features calculated from the data.

We want the linear combination of features which gives an unbiased estimate of θ and has minimum variance among all such estimates. This is achieved by first finding the $q \times 1$ column vector ℓ such that for a given separation, $\ell^{\mathrm{T}}\bar{V}$ of $E(\ell^{\mathrm{T}}Z_{\mathcal{D}})$ between the two base levels, $\mathrm{var}(\ell^{\mathrm{T}}Z_{\mathcal{D}}) = \ell^{\mathrm{T}}\Sigma\ell$ is minimized. That is, we want the vector ℓ that will discriminate most sharply between the distributions at the two base levels of θ. Then the resulting combination is standardized to produce the correct expected value at the base levels of θ. Linear interpolation of the observed value between the base levels then produces the required estimate.

The first step leads to the vector of coefficients $\ell = \lambda^{-1}\Sigma^{-1}\bar{V}$, where the standardizing coefficient λ is arbitrary, but may, for example, be chosen to be $(\bar{V}^{\mathrm{T}}\Sigma\bar{V})^{1/2}$ to produce unit variance for $\ell^{\mathrm{T}}Z_{\mathcal{D}}$. The vector $Z_{\mathcal{D}}$ determines the value of the linear discriminant $y_{\mathcal{D}} = \ell^{\mathrm{T}}Z_{\mathcal{D}}$. The expected values at the two base points are then $\bar{y}_i = \ell^{\mathrm{T}}\bar{Z}_i$, where $\bar{Z}_i$ is the i^{th} column of $\bar{Z}$. This leads in the final interpolatory step to the point estimate

$$\tilde{\theta} = \theta_0 + d\left(y_{\mathcal{D}} - \frac{\bar{y}_1 + \bar{y}_2}{2}\right), \tag{2.2.1}$$

with variance

$$\mathrm{var}(\tilde{\theta}) = d^2\ell^{\mathrm{T}}\Sigma\ell, \tag{2.2.2}$$

where $d = 2h/(\bar{y}_2 - \bar{y}_1)$.

2.2.2 Normal mean

Consider the estimation of the mean μ of a normal distribution from a sample of size 100. Assuming that the variance σ^2 is known and equal to one, we generate 1000 datasets $\mathcal{D}_1, \ldots, \mathcal{D}_{1000}$ of size $n = 100$ each, with $\mu = 2.3$. We generate several datasets with the same parameter value in this example, to obtain an average of the parameter estimates and to be able to calculate the variances empirically, in order to compare them to the variances obtained theoretically using the method. We choose $\mu_0 = 2$ as the initial value and spacing $h = 0.25$. We then perform $r = 10000$ simulations, for each simulating a sample of size $n = 100$ from the points $\mu_0 \pm h$. We use $q = 3$ statistics: the average of the first and last order statistic, $(x_{(1)}+x_{(n)})/2$, the median and the sum of the order statistics from 2 to $n-1$, $x_{(2)}+x_{(3)}+\ldots+x_{(n-1)}$. We then obtain $\text{ave}(\tilde{\mu}) - \mu = -0.005$ and $\text{var}(\tilde{\mu}) = 0.00988$, where $\tilde{\mu}$ is the point estimate of μ obtained from the method, the average taken over the 1000 estimates obtained from the simulated datasets $\mathcal{D}_1, \ldots, \mathcal{D}_{1000}$.

The estimate $\tilde{\mu}$ obtained is a linear combination of the features used in the procedure. In this example, the estimate should be very close to the maximum likelihood estimate, $\hat{\mu} = \bar{X}$, the sample mean, and its variance should be equal to the variance of $\bar{X}$, σ^2/n. The coefficients obtained from the method of Lagrange multipliers should be such that $y_{\mathcal{D}}$ is proportional to $(x_{(1)} + \cdots + x_{(n)})/n$. So we expect to obtain ℓ_1 proportional to $2/n$, ℓ_2 close to zero and $\ell_3 \propto 1/n$. Here we got $\ell_1 = 0.1956$, $\ell_2 = -0.0305$ and $\ell_3 = 0.1010$.

Results for the estimation of μ where the underlying model is a normal distribution of mean 5 and unit variance are shown in Table 2.2.1. To investigate the effect of the starting point and displacement h on the outcome, various values of μ_0 and h were used. The sample size was $n = 100$, $r = 10000$ simulations were performed and 1000 datasets $\mathcal{D}_i$ were generated. If the initial estimate $\tilde{\mu}_1$ fell outside the baseline levels $\mu_0 \pm h$, the simulations were repeated with μ_0 reset to $\tilde{\mu}_1$ and the procedure was repeated until the estimate was within the interval defined by the base-points. As a precaution one additional iteration was then used. The theoretical variance was calculated using Equation (2.2.2) and the empirical variance is the sample variance among the 1000 estimates produced, one for each dataset $\mathcal{D}_i$, $i = 1, \ldots, 1000$.

$\boldsymbol{h} = \mathbf{0.05}$

$\boldsymbol{\mu_0}$	**4**	**5**	**6**	**8**
mean($\tilde{\mu}$)	4.9985	5.0023	5.0021	5.0068
theoretical var($\tilde{\mu}$)	0.00988	0.00985	0.00980	0.01016
empirical var($\tilde{\mu}$)	0.00974	0.00987	0.01090	0.00994
iterations	3	2	3	3

$\boldsymbol{h} = \mathbf{0.5}$

$\boldsymbol{\mu_0}$	**4**	**5**	**6**	**8**
mean($\tilde{\mu}$)	4.9931	5.0056	4.9973	4.9982
theoretical var($\tilde{\mu}$)	0.00999	0.00996	0.00997	0.00989
empirical var($\tilde{\mu}$)	0.01000	0.01043	0.01000	0.00994
iterations	3	2	3	3

Table 2.2.1: *Estimated values of μ for a linear estimation problem. $n = 100$. μ_0 is the starting value and h the spacing.*

A key aspect of the method is it is restricted to linear combinations of the chosen features. In this particular example the defining features are linear and the defining parameter is a location parameter. This implies that the outcome does not depend on the difference between the base point μ_0 and the unknown population mean μ or at all critically on the displacement h.

To investigate in this simple case the effect of nonlinearity one might use nonlinear features; instead we have taken estimation of a nonlocation parameter, $\phi = \log \mu$. That is, we take as base points for the simulation $\phi_0 \pm h$, that is for μ the values $\mu_0 e^{\pm h}$, where $\mu_0 = e^{\phi_0}$. To test the method we took as the underlying model a normal distribution of mean 5 and variance one and starting values corresponding to $h = 0.2, \mu_0 = 4, 5, 6, 8$. The results are shown in Table 2.2.2.

The first two values of h are approximately one-half and twice the standard error of the resulting estimate and give essentially equivalent and satisfactory results. But at $h = 0.2$, and even more strikingly at $h = 0.5$, biases in the estimation of ϕ and μ are encountered which, while numerically small, are substantial compared with the standard error. Note that, for example, $h = 0.5$ defines base-points at $\mu = \mu_0 e^{\pm 0.5}$, that is $0.61\mu_0$ and $1.65\mu_0$. This is a wide enough range for the curvature of the log function to be not entirely negligible, so that the presumed linearity of the estimate over the range in question is not a good approximation.

$\boldsymbol{h} = \mathbf{0.01}$

$\boldsymbol{\mu_0}$	**4**	**5**	**6**	**8**
mean($\tilde{\phi}$)	1.6096	1.6099	1.6105	1.6090
mean($\tilde{\mu}$)	5.0007	5.0024	5.0053	4.9976
theoretical var($\tilde{\phi}$)	4.12×10^{-4}	4.05×10^{-4}	3.82×10^{-4}	3.95×10^{-4}
empirical var($\tilde{\phi}$)	4.22×10^{-4}	4.14×10^{-4}	4.05×10^{-4}	3.72×10^{-4}
iterations	3	2	3	4

$\boldsymbol{h} = \mathbf{0.05}$

$\boldsymbol{\mu_0}$	**4**	**5**	**6**	**8**
mean($\tilde{\phi}$)	1.6081	1.6075	1.6087	1.6090
mean($\tilde{\mu}$)	4.9935	4.9903	4.9965	4.9977
theoretical var($\tilde{\phi}$)	4.01×10^{-4}	4.07×10^{-4}	4.06×10^{-4}	3.94×10^{-4}
empirical var($\tilde{\phi}$)	4.14×10^{-4}	3.94×10^{-4}	4.07×10^{-4}	3.82×10^{-4}
iterations	3	2	3	4

$\boldsymbol{h} = \mathbf{0.2}$

$\boldsymbol{\mu_0}$	**4**	**5**	**6**	**8**
mean($\tilde{\phi}$)	1.5889	1.5899	1.5893	1.5896
mean($\tilde{\mu}$)	4.8980	4.9032	4.9003	4.9016
theoretical var($\tilde{\phi}$)	4.12×10^{-4}	4.09×10^{-4}	3.96×10^{-4}	4.09×10^{-4}
empirical var($\tilde{\phi}$)	4.31×10^{-4}	3.99×10^{-4}	3.60×10^{-4}	4.22×10^{-4}
iterations	3	2	2	3

$\boldsymbol{h} = \mathbf{0.5}$

$\boldsymbol{\mu_0}$	**4**	**5**	**6**	**8**
mean($\tilde{\phi}$)	1.4892	1.4893	1.4867	1.4866
mean($\tilde{\mu}$)	4.4336	4.4339	4.4227	4.4220
theoretical var($\tilde{\phi}$)	4.50×10^{-4}	4.65×10^{-4}	4.42×10^{-4}	3.77×10^{-4}
empirical var($\tilde{\phi}$)	4.64×10^{-4}	4.30×10^{-4}	4.66×10^{-4}	3.99×10^{-4}
iterations	2	2	2	2

Table 2.2.2: *Estimated values of $\phi = \log \mu$. Nonlinear estimation problem. $n = 100$. μ_0 is the starting value for μ and h the spacing for ϕ.*

2.3 A general formulation

We now give a general formulation for a model with a $p \times 1$ parameter vector θ, using a set of features Z defining a $1 \times q$ row vector, where $q > p$. On the basis of prior knowledge we specify a simulation base level θ_0 and appropriate displacements h, so that simulations will be run at parameter values $\theta_0 \pm h$. That is, the i^{th} component of θ will be set at values $\theta_{i0} \pm h_i$. This specifies 2^p possible parameter points, but it is sufficient to use a fraction of these allowing estimation of main effects. For this we choose a $k \times p$ design matrix D consisting of orthogonal columns, each containing an equal number of entries 1 and -1. Following Plackett and Burman (1946) this can be achieved by taking p columns of a $k \times k$ Hadamard matrix ($k \geq p$) to form a matrix D, specifying a design. Hadamard matrices consist of a first column of elements 1, the remaining columns being formed from mutually orthogonal columns of elements 1 and -1. They are known to exist for all values of k that are a multiple of four, up to $k = 428$ (Kharaghani and Tayfeh-Rezaie, 2005) and for $k = 764, 1004, 2524, 23068, 28324, 32996$ (Đoković, 2010; Đoković et al., 2013). See also Cox and Reid (2000, p. 116–, 261).

Thus for $k = 4$ a Hadamard matrix is

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix}.$$

Then for $p = 3$ we take D to be the last three columns of this 4×4 matrix. Each row defines a set of parameter values corresponding to three component parameters. For example, the second row defines the parameter point $(\theta_{01} - h_1,\ \theta_{02} + h_2,\ \theta_{03} - h_3)$.

The simulation procedure is that at each of the k design points corresponding to D a suitably large number of simulation runs is made, each run determining a value of the feature vector Z. From these are calculated a $q \times k$ matrix of means $\bar{Z}$ and a pooled covariance matrix Σ.

The $q \times p$ matrix $\bar{Z}D$ has rows defined by features and columns defined by the component parameters, and the entries are proportional to corresponding contrasts of the means between the two levels of the parameter component. For each component we apply the argument of Section 2.2, that is we consider the linear combination of feature contrasts that is most sensitive, that is has minimum variance subject to a given expected value of the associated contrast of levels. That is, we aim to find among estimates that are linear in the features estimates which locally are minimum variance unbiased estimates. Also we find the covariance matrix of the resulting estimates.

If $Z_{\mathcal{D}}$ is a single observed value of the $q \times 1$ feature vector derived from the data under analysis and if ℓ is a column vector of coefficients, the variance of $\ell^{\mathrm{T}} Z_{\mathcal{D}}$ is $\ell^{\mathrm{T}} \Sigma \ell$ and the difference between its expected values at the two levels of the first component is $\ell^{\mathrm{T}}(\bar{Z}D)_1$, where the suffix 1 denotes the first column of the matrix in question. Use of the method of Lagrange multipliers leads to the linear discriminant $(D_1)^{\mathrm{T}} \bar{Z}^{\mathrm{T}} \Sigma^{-1} Z_{\mathcal{D}}$ for the first component, and in general to the vector of linear discriminants

$$Y_{\mathcal{D}} = L^{\mathrm{T}} Z_{\mathcal{D}}, \tag{2.3.1}$$

where $L = \Sigma^{-1} \bar{Z} D$. It is assumed that $\mathbb{E}(Y_{\mathcal{D}})$ is a linear function of θ. This, combined with the orthogonality of the columns of D, ensures that each element of $Y_{\mathcal{D}}$ yields an estimate of the relevant component of θ.

Thus the covariance matrix of $Y_{\mathcal{D}}$ derived from the single observation $Z_{\mathcal{D}}$ for the data under analysis is

$$\operatorname{cov}(Y_{\mathcal{D}}) = L^{\mathrm{T}} \Sigma L. \tag{2.3.2}$$

From the overall simulation means we find

$$\bar{Y} = L^{\mathrm{T}} \bar{Z}. \tag{2.3.3}$$

It remains to convert (2.3.1) into a point estimate of θ and its covariance matrix. This is done by linear interpolation between the average values of (2.3.1) at the two simulation levels for each component.

The differences between the means of the new variables at the end points of the range are

$$A = 2k^{-1} \mathrm{diag}\left\{ D^{\mathrm{T}} \bar{Z}^{\mathrm{T}} L \right\} = \mathrm{diag}\left\{ \alpha_1, \ldots, \alpha_p \right\}, \tag{2.3.4}$$

where the diagonal matrix is formed from the diagonal elements of the relevant matrix.

It follows by linear interpolation that the estimate of θ is

$$\begin{aligned}\tilde{\theta} &= \theta_0 + 2\mathrm{diag}\,\{h_1, \ldots, h_p\}\, A^{-1}(Y_{\mathcal{D}} - L^{\mathrm{T}}\bar{Z}\mathbb{1}k^{-1}) \\ &= \theta_0 + 2C(Y_{\mathcal{D}} - \bar{Y}\mathbb{1}k^{-1}),\end{aligned} \tag{2.3.5}$$

where $\mathbb{1}$ is a $k \times 1$ vector of elements one and $C = \mathrm{diag}\,\{h_1/\alpha_1, \ldots, h_p/\alpha_p\}$. It follows then that, neglecting errors in $\bar{Z}$,

$$\mathrm{cov}(\tilde{\theta}) = 4CL^{\mathrm{T}}\Sigma LC. \tag{2.3.6}$$

Interpretation of the form of the resulting estimates may often be best done by standardizing the features to have unit standard deviation. Equivalently, the matrix of defining coefficients L is transformed to L^*, where $L^* = (\mathrm{diag}\{\Sigma\})^{1/2}L$.

When the estimate lies outside the interval $(\theta_0 - h,\ \theta_0 + h)$, we iterate until it lies inside the design region in the parameter space, so that the final estimate is obtained by interpolation, not extrapolation.

If the sample size of the data is too large so that simulated samples of that size cannot be easily generated, a smaller sample size could be used in the simulations and then the variances can be multiplied by a factor to account for the ratio of the data and simulation sample sizes.

2.4 A two-dimensional example: Weibull distribution

Suppose we have a sample from the Weibull distribution with density

$$f(x) = \frac{\gamma}{\rho}\left(\frac{x}{\rho}\right)^{\gamma-1} e^{-\left(\frac{x}{\rho}\right)^{\gamma}}, \quad \rho, \gamma > 0, \quad x > 0$$

and want to estimate the parameters ρ and γ. Simulated datasets were used, with ρ equal to 1 and γ taking the values 1, 1.5, 2 and 3. The sample size used was $n = 100$. Choosing initial values ρ_0, γ_0 and spacings (h_1, h_2), four design points are specified by two columns of a 4×4 Hadamard matrix. $r = 10000$ simulations of size $n = 100$ each were run from the distribution at each of these points. It is plausible in view of the form of the density and implied broad nature of the likelihood to base the estimation on sums of powers of the observations, although of course the appropriate powers are unknown, hence the following statistics were used:

$$z_1 = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$z_2 = \frac{1}{n}\sum_{i=1}^{n} \sqrt{X_i},$$

and

$$z_3 = \frac{1}{n}\sum_{i=1}^{n} (X_i)^{1.5}.$$

We form an $r \times q$ matrix Z_i for each design point i with entries the simulated values of the $q = 3$ chosen features. Let $\bar{Z}$ be a $q \times 4$ matrix of the values of the q features averaged over all simuation runs at each design point. The covariance matrix $\Sigma_Z^{(i)}$ of Z at each point i is then calculated from the simulated values. We assume that $\Sigma_Z^{(i)}$ is at least approximately constant, hence we use an average $\Sigma_Z \equiv \Sigma$ over the four matrices. Multiplying $\bar{Z}$ by the appropriate columns of the Hadamard matrix gives a $q \times 2$ matrix of contrasts $\bar{V}$.

We then seek a $q \times 2$ matrix L of coefficients that maximizes the linear combinations that are given by the rows of $L^{\mathrm{T}}\bar{V}$, subject to a given constant variance $v = L^{\mathrm{T}}\Sigma L$, set equal to 1 here. Using the method of Lagrange multipliers, we obtain the matrix of coefficients $L = \frac{1}{\lambda_i}\ \Sigma^{-1}\ \bar{V}_i, \quad i = 1, 2,$ where $\lambda_i = (\bar{V}_i^{\mathrm{T}}\ \Sigma^{-1}\ \bar{V}_i)^{1/2}, \quad i = 1, 2,$ and $\bar{V}_i$ is the i^{th} column of the matrix $\bar{V}$.

We obtain the linear combinations $\bar{Y} = L^{\mathrm{T}}\bar{Z}$ from the simulations and $Y_{\mathcal{D}} = L^{\mathrm{T}}\bar{Z}_{\mathcal{D}}$ from the data. The estimates of the parameters are given by:

$$\tilde{\rho} = \rho_0 + h_1 \frac{4y_{\mathcal{D}_1} - (\bar{y}_{11} + \bar{y}_{12} + \bar{y}_{13} + \bar{y}_{14})}{\bar{y}_{11} - \bar{y}_{12} + \bar{y}_{13} - \bar{y}_{14}}$$

and

$$\tilde{\gamma} = \gamma_0 + h_2 \frac{4y_{\mathcal{D}_2} - (\bar{y}_{21} + \bar{y}_{22} + \bar{y}_{23} + \bar{y}_{24})}{\bar{y}_{21} + \bar{y}_{22} - \bar{y}_{23} - \bar{y}_{24}} \ ,$$

where y_{ij} represents the (i, j) element of the matrix Y $(i = 1, 2; j = 1, \ldots, 4)$, and $y_{\mathcal{D}_i}$ is the i^{th} element of $Y_{\mathcal{D}}$ $(i = 1, 2)$. The covariance matrix of $(\tilde{\rho}, \tilde{\gamma})$ is given by

$$4CL^{\mathrm{T}}\Sigma LC,$$

where

$$C = \operatorname{diag}\left\{ \frac{2h_1}{\bar{y}_{11} - \bar{y}_{12} + \bar{y}_{13} - \bar{y}_{14}}, \frac{2h_2}{\bar{y}_{21} + \bar{y}_{22} - \bar{y}_{23} - \bar{y}_{24}} \right\} .$$

For each set of true parameter values, 1000 data sets were generated to give a set of estimates $\{(\tilde{\rho}_i, \tilde{\gamma}_i); i = 1, \ldots, 1000\}$. Then the mean and sample variance of ρ and γ were calculated (denoted as empirical variance), as well as the mean of the variances calculated from equation (2.4) (theoretical variance). In addition, the asymptotic variances of the maximum likelihood estimates $(\hat{\rho}, \hat{\gamma})$ were calculated for comparison. The results obtained are shown in Table 2.4.1.

Except for the estimation of γ at the highest spacing the estimates are essentially unbiased. There are some discrepancies between the empirical and theoretical variances although the reason for this is unclear. However the theoretical variances here are based on a single set of simulations, while the empirical variances are obtained from a set of several estimates based on that single set of simulations but different generated datasets.

$(\boldsymbol{\rho},\ \boldsymbol{\gamma})$	**(1, 1)**			**(1, 1.5)**		
(ρ_0, γ_0)	(1, 1)			(1, 1.3)		
$h_1 = h_2 = h$	0.05	0.1	0.2	0.05	0.1	0.2
mean$(\tilde{\rho})$	0.998	0.991	0.993	0.962	0.961	0.970
mean$(\tilde{\gamma})$	1.005	1.019	1.063	1.464	1.482	1.544
theoretical var$(\tilde{\rho})$	0.01031	0.01074	0.01155	0.00605	0.00620	0.00654
empirical var$(\tilde{\rho})$	0.01002	0.01029	0.00971	0.00464	0.00427	0.00459
theoretical var$(\tilde{\gamma})$	0.00569	0.00610	0.00692	0.00929	0.01008	0.01193
empirical var$(\tilde{\gamma})$	0.00559	0.00546	0.00574	0.00588	0.00542	0.00595
asymptotic (var$(\hat{\rho})$, var$(\hat{\gamma})$)	$(0.01109, 0.00608)$			$(0.00493, 0.01368)$		
$(\boldsymbol{\rho},\ \boldsymbol{\gamma})$	**(1, 1.5)**			**(1, 2)**		
(ρ_0, γ_0)	(1, 1.5)			(1, 2.3)		
$h_1 = h_2 = h$	0.05	0.1	0.2	0.05	0.1	0.2
mean$(\tilde{\rho})$	1.003	0.994	0.999	1.025	1.020	1.008
mean$(\tilde{\gamma})$	1.505	1.528	1.595	1.972	2.016	2.215
theoretical var$(\tilde{\rho})$	0.00455	0.00470	0.00494	0.00191	0.00204	0.00217
empirical var$(\tilde{\rho})$	0.00433	0.00433	0.00465	0.00262	0.00266	0.00271
theoretical var$(\tilde{\gamma})$	0.01292	0.01331	0.01620	0.02896	0.03370	0.04393
empirical var$(\tilde{\gamma})$	0.01203	0.01271	0.01296	0.04688	0.04668	0.04687
asymptotic (var$(\hat{\rho})$, var$(\hat{\gamma})$)	$(0.00493, 0.01368)$			$(0.00277, 0.02432)$		
$(\boldsymbol{\rho},\ \boldsymbol{\gamma})$	**(1, 3)**			**(1, 3)**		
(ρ_0, γ_0)	(1, 2.5)			(1, 3)		
$h_1 = h_2 = h$	0.05	0.1	0.2	0.05	0.1	0.2
mean$(\tilde{\rho})$	0.977	0.980	0.995	0.998	0.993	1.003
mean$(\tilde{\gamma})$	2.909	2.986	3.222	3.039	3.136	3.516
theoretical var$(\tilde{\rho})$	0.00165	0.00173	0.00184	0.00116	0.00122	0.00130
empirical var$(\tilde{\rho})$	0.00099	0.00115	0.00154	0.00110	0.00115	0.00135
theoretical var$(\tilde{\gamma})$	0.03691	0.04186	0.05494	0.05498	0.06122	0.08688
empirical var$(\tilde{\gamma})$	0.02030	0.02187	0.02337	0.04986	0.04917	0.06186
asymptotic (var$(\hat{\rho})$, var$(\hat{\gamma})$)	$(0.00123, 0.05471)$			$(0.00123, 0.05471)$		

Table 2.4.1: *Estimated values and variances for the parameters of the Weibull distribution.* ρ_0 *and* γ_0 *are the starting values of the parameters and* h *is the spacing.* $\tilde{\rho}$, $\tilde{\gamma}$ *are the estimates obtained from the method and* $\hat{\rho}$, $\hat{\gamma}$ *the maximum likelihood estimates.*

2.5 An example: Superposition of renewal processes

As a relatively simple but realistic example where calculation of a likelihood function is not feasible we consider estimation connected with a superimposition of point processes. This was considered by Cox and Smith (1954) motivated by an issue in neurophysiology. The data and some further discussion are given by Cox and Lewis (1966). The data, provided by Professor B. Katz and Dr P. Fatt, University College London, give 799 time intervals in units of 1/50 sec. between successive signals along a nerve fibre. The values have mean 10.95 and variance 109.40. The data appear superficially like a realization of a stationary Poisson process but there are some departures. To study these it was assumed on subject-matter grounds that the generating process consisted of an unknown number of statistically identical but independent renewal processes. It is known that even if the distribution defining the renewal processes is not exponential some of the local properties of a Poisson process are mimicked after superposition (Cox and Smith, 1954).

For the present discussion we assume a process which is the superposition of an unknown number N of renewal processes, illustrated in Figure 2.5.1, with waiting times that follow a Gamma(α, β) distribution with density

$$f(x) = \frac{\alpha(\alpha x)^{\beta-1}e^{-\alpha x}}{\Gamma(\beta)}, \quad x > 0, \quad \alpha, \beta > 0.$$

Because the mean interval between successive points in the process is rela-

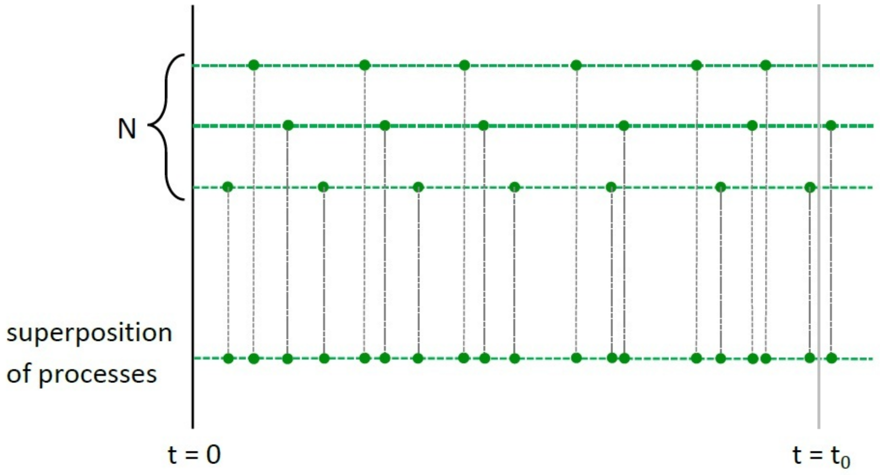

Figure 2.5.1: *Illustration of N superimposed processes.*

tively well estimated, we take as parameters the mean interval, $\theta = \beta/(N\alpha)$, β and N, the last being of primary interest.

Initial values θ_0, β_0, N_0 and spacings (h_1, h_2, h_3) are chosen. Then four design points are specified by three columns of a 4×4 Hadamard matrix. Then r simulations of size n are run from the distribution at each of these points.

To choose features for inclusion we may consider the data as specified either as here by intervals between successive points or by counts of numbers of points falling in intervals of various lengths. Simple statistics based on the former are likely to be useful if N is small, whereas the latter become more relevant if N is larger. We therefore take a mixture of the two types. In either case the mean rate of occurrence, $z_1 = R_{(0,t_0)}/t_0$, where $R_{(0,t_0)}$ is the number of points observed in the interval $(0, t_0)$ is certain to be required. We also use z_2, the coefficient of variation of intervals, x_i, between successive points,

$\sqrt{\{\sum_{i=1}^{n}(x_i - \bar{x})^2/n\}}/\bar{x}$, and z_3 the sum of the first five autocorrelations of the intervals. In addition we include z_4, the mean of the intervals, and z_5, z_6, z_7, z_8, the Poisson index of dispersion, variance divided by mean, for intervals of length $1, 5, 10$ and 20.

We form an $r \times q$ matrix Z_i for each design point i with the simulated values of the q selected features. Let $\bar{Z}$ be a $q \times 4$ matrix of the values of the q features averaged over all simuation runs at each design point. The covariance matrix $\Sigma_Z^{(i)}$ of Z at each point i can be calculated from the simulated values. We use an average Σ of the four matrices. Multiplying $\bar{Z}$ by the appropriate columns of the Hadamard matrix gives a $q \times 3$ matrix of contrasts $\bar{V}$.

We then find a $q \times 3$ matrix L of coefficients to obtain the linear combinations $\bar{Y} = L^{\mathrm{T}}\bar{Z}$ from the simulations and $Y_{\mathcal{D}} = L^{\mathrm{T}}\bar{Z}_{\mathcal{D}}$ from the data. The estimates of the parameters are given by:

$$\tilde{\theta} = \theta_0 + h_1 \frac{4y_{\mathcal{D}_1} - (\bar{y}_{11} + \bar{y}_{12} + \bar{y}_{13} + \bar{y}_{14})}{\bar{y}_{11} - \bar{y}_{12} + \bar{y}_{13} - \bar{y}_{14}} \ ,$$

$$\tilde{\beta} = \beta_0 + h_2 \frac{4y_{\mathcal{D}_2} - (\bar{y}_{21} + \bar{y}_{22} + \bar{y}_{23} + \bar{y}_{24})}{\bar{y}_{21} + \bar{y}_{22} - \bar{y}_{23} - \bar{y}_{24}}$$

and

$$\tilde{N} = N_0 + h_3 \frac{4y_{\mathcal{D}_3} - (\bar{y}_{31} + \bar{y}_{32} + \bar{y}_{33} + \bar{y}_{34})}{\bar{y}_{31} - \bar{y}_{32} - \bar{y}_{33} + \bar{y}_{34}} \ ,$$

where y_{ij} is the (i, j) element of the matrix Y $(i = 1, 2, 3;\ j = 1, \ldots, 4)$ and

$y_{\mathcal{D}_i}$ is the i^{th} element of $Y_{\mathcal{D}}$. The covariance matrix of $(\tilde{\theta}, \tilde{\beta}, \tilde{N})$ is given by $4CL^{\mathrm{T}}\Sigma LC$, where

$$C = \operatorname{diag}\left\{\frac{2h_1}{\bar{y}_{11} - \bar{y}_{12} + \bar{y}_{13} - \bar{y}_{14}}, \frac{2h_2}{\bar{y}_{21} + \bar{y}_{22} - \bar{y}_{23} - \bar{y}_{24}}, \frac{2h_3}{\bar{y}_{31} - \bar{y}_{32} - \bar{y}_{33} + \bar{y}_{34}}\right\}.$$

Using an iterative procedure, re-centering the parameters at the estimates of the previous step, 2000 simulations were run with $h_1 = h_2 = 0.5$ and $h_3 = 1$. The starting values were chosen to be $\theta_0 = 10$, $\beta_0 = 40$ and $N_0 = 20$. After approximately 38 iterations, the estimates obtained are shown in Table 2.5.1(a).

(a)		**(b)**	
$\tilde{\theta} = 12.3$	s.e.$(\tilde{\theta}) = 4.64$	$\tilde{\theta} = 4.88$	s.e.$(\tilde{\theta}) = 0.25$
$\tilde{\beta} = 55.7$	s.e.$(\tilde{\beta}) = 6.79$	$\tilde{\beta} = 7.07$	s.e.$(\tilde{\beta}) = 1.82$
$\tilde{N} = 100.2$	s.e.$(\tilde{N}) = 6.00$	$\tilde{N} = 2.20$	s.e.$(\tilde{N}) = 0.30$

Table 2.5.1: *Estimates for (a) the data of Fatt and Katz, (b) artificial example with $N = 2$, $\theta = 4.5$, $\beta = 9$.*

The estimate of N is only broadly comparable to the estimate of about 170 obtained by Cox and Lewis (1966) by a quite informal approach. Even though the notional standard error of $\tilde{N}$ reported in Table 2.5.1(a) is relatively small the estimate itself implies that only a few events have been recorded from each of the component processes and this must throw doubt on the suitability and interpretability of the model.

Because of this we also analyzed data from an artificial series of 87 events obtained by superimposing two independent renewal processes, each with a

Gamma distribution of intervals of mean 9 and coefficient of variation 33% (Cox and Lewis, 1966). The period of observation is 420 units. The mean interval of the observed data is 4.82.

To estimate $\theta = \beta/(N\alpha)$, the mean interval of the pooled process, β and N, 50 iterations were run, with 2000 simulations each. Spacings were chosen to be $h_1 = h_2 = 0.5$ and $h_3 = 1$. Starting from the values $\theta_0 = 5$, $\beta_0 = 10$ and $N_0 = 5$, the results obtained are shown in Table 2.5.1(b). The estimates of the three components were almost uncorrelated. The predominant contributions to the estimate of N were relatively high values of the coefficient of variation of intervals between successive points combined with relatively low values for the overdispersion of counts.

2.6 Hidden Markov chain

A Hidden Markov model (Zucchini and MacDonald, 2009) was considered as another example. Let $X_1, \ldots, X_n$ be binary $(0, 1)$ variables which form a two-state Markov chain. The transition probabilities are given by

$$\begin{aligned} \mathbb{P}(X_t = 1 \mid X_{t-1} = 0) &= \alpha \\ \mathbb{P}(X_t = 0 \mid X_{t-1} = 1) &= \beta. \end{aligned} \tag{2.6.1}$$

The initial condition is

$$\mathbb{P}(X_1 = 1) = \frac{\alpha}{\alpha + \beta}.$$

Suppose that X_t is not observed, but instead we observe Y_t which has independent error of obervation given by

$$\begin{aligned} \mathbb{P}(Y_t = 1 \mid X_t = 0) &= \epsilon \\ \mathbb{P}(Y_t = 0 \mid X_t = 1) &= \eta. \end{aligned} \tag{2.6.2}$$

We simplify the problem by assuming that η is equal to 0. We observe $Y_1, \ldots, Y_n$, independent given the Xs.

We use the following features:

- proportion of 1's
- proportion of 1's following a 1
- proportion of 1's following a 0
- proportion of 1's following 11
- proportion of 1's following 10
- proportion of 1's following 01
- proportion of 1's following 00,

since for a hidden Markov chain the Markov property does not hold and information lies in the dependence between future and past states conditionally on the present state.

We generate a simulated dataset for analysis with parameter values $\alpha = \beta = 0.25$ and $\epsilon = 0.3$. We run $r = 10000$ simulations, generating samples of size $n = 500$ at each design point. We take $h = 0.05$ and starting parameter values $\alpha_0 = \beta_0 = \epsilon_0 = 0.4$.

By running 50 iterations, we obtain $\tilde{\alpha} = 0.280$, $\tilde{\beta} = 0.188$ and $\tilde{\epsilon} = 0.306$. The covariance matrix of $\tilde{\theta} = (\tilde{\alpha}, \tilde{\beta}, \tilde{\epsilon})$ is

$$\mathrm{cov}(\tilde{\theta}) = \begin{pmatrix} 2.90 \times 10^{-4} & -9.25 \times 10^{-5} & 2.81 \times 10^{-4} \\ -9.25 \times 10^{-5} & 3.49 \times 10^{-4} & -1.11 \times 10^{-4} \\ 2.81 \times 10^{-4} & -1.11 \times 10^{-4} & 3.75 \times 10^{-4} \end{pmatrix},$$

therefore the standard errors of $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\epsilon}$ are 0.017, 0.019 and 0.019, respectively. Also $\tilde{\alpha}$ and $\tilde{\epsilon}$ are highly correlated, while $\tilde{\alpha}$ with $\tilde{\beta}$ and $\tilde{\beta}$ with $\tilde{\epsilon}$ have small negative correlation.

The coefficients obtained at the final iteration are

$$L = \begin{pmatrix} 75.62 & -28.02 & 66.33 \\ -10.09 & -4.46 & -53.65 \\ -24.67 & -30.42 & 44.77 \\ -52.31 & 17.91 & 3.01 \\ -8.68 & 43.60 & -3.81 \\ 24.78 & 125.41 & 3.95 \\ -0.29 & -3.48 & 6.31 \end{pmatrix}.$$

The values of the seven features calculated from the data are roughy equal to those calculated from a dataset with parameters $\tilde{\alpha}$, $\tilde{\beta}$ and $\tilde{\epsilon}$ (Table 2.6.1).

Values of features calculated from data generated by:	
the true parameter values	the estimated parameter values
0.644	0.730
0.469	0.589
0.176	0.140
0.353	0.488
0.080	0.064
0.116	0.102
0.096	0.076

Table 2.6.1: *Values of features calculated using true and estimated parameter values for the hidden Markov chain example.*

2.7 Discussion

There are a number of issues that arise in applying the procedure of this chapter. Even in the relatively simple examples discussed here some choices are involved that are difficult to formalize.

First, as with other iterative procedures, a starting point has to be chosen. In more complicated examples, and especially when there is little prior information about plausible parameter values, use of multiple starting points will be desirable. To avoid errors from extrapolation we continued until the estimated parameter values lie within the region spanned by the base points used in simulation, so that the final estimate is obtained by interpolation, not extrapolation. Then at least one precautionary iteration is made, mainly to check on stability. It might also be reasonable in some cases to perform a few more iterations after the procedure seems to have converged and then take $\tilde{\theta}$ to be the average of the estimates obtained from these extra iterations.

The choice of numerical values for the components of h is more difficult. As with other forms of numerical differentiation, a compromise is needed between the instability induced by unsuitably small h and the bias arising from nonlinearity in the objective function. There is in some cases a need to avoid h so large that inadmissible parameter values might become involved. Formal discussion of optimality does not seem possible. We have followed the informal rule of ending with values of h rather smaller than the standard error of the estimate. It would also be possible to employ a stochastic approximation in which the spacing h decreases as iteration proceeds.

Calculation of the condition number of the matrix Σ gives an indication either of redundant features or of nearly linear relations between the features, calling either for omission or redefinition. A large condition number of the covariance matrix of the estimates suggests reparameterization.

The method hinges on local linearity of the dependence of the optimal estimate on the chosen features. This could in principle be checked from the start by using a design more complicated than that used here, for example a central composite design allowing estimation of quadratic dependence. If, however, the procedure is used iteratively information about nonlinearity is built up by comparing distinct sections with differing simulation bases.

As with other applications of discriminant analysis and similar techniques, it will in applications often be helpful to use simplified sets of coefficients in which, for example, small coefficients are replaced by zero and others by simple integers. For this it may be helpful to standardize the features by their standard deviation and the matrix, L^*, of coefficients with such standardized variables calculated. In particular, sometimes the estimates can be approximated by statistics with a direct interpretation. In some cases also dimensional considerations can guide the choice of features as when one or more of the component parameters are location or regression parameters or when some are dimensionless. We have not incorporated this formally into the procedure.

There is the possibility at some intermediate stage of the analysis of simplifying the procedure by removing from the analysis features that either appear to make no clear contribution to the estimation either to any of the components of the parameter θ or to specific components of interest as assessed component by component. While we shall not suggest a formal procedure for doing this, some comparison of the estimated coefficients with their formal standard errors could be used. Especially if the number of features is initially large there may be computational advantages including improved convergence resulting from the simplification.

To test whether some $\ell_i = 0$, one could run several iterations of the procedure after the results are stable enough to suggest convergence and calculate an empirical standard error of each coefficient ℓ_i. Then each ℓ_i divided by its empirical standard error could be compared to a standard normal distribution.

In calculating the covariance matrix of the final estimates errors of estimation arising in the simulation, that is errors in $\bar{Z}$ and in Σ, are ignored. They can indeed be made small unless simulation is expensive. Moreover to use the method, the defining combinations may be fixed by a single initial calculation in which case calculations of precision would be for those fixed values.

One limitation of the method is the restriction to estimates that are linear combinations of the defining features of the data. Although in a sense most reasonable estimates are asymptotically locally linear and *ad hoc* modifications to the procedure could be made by changing the defining features in the light of preliminary results, a more systematic approach would be to use central composite designs to explore second degree forms of the estimates, possibly leading to modified definitions of the defining features.

Finally we make some brief comparisons with alternative approaches, especially the various forms of ABC. Both the proposed method and most versions of ABC and Indirect Inference are based on simulations and chosen summary statistics based on which the observed and simulated data are compared. Distinctive aspects of the present approach are the use of Hadamard matrices to focus the simulations and the calculation of individual linear discriminants to study component parameters. Thus the present procedure gives an explicit form for the estimate of each component. In some cases this may lead to fruitful interpretation and simplification. Study of the covariance matrix of the estimates could in principle be used to find a simplifying linear reparameterization.

An advantage of the method over ABC is that an explicit prior does not have to be introduced. Informal prior knowledge guides the choice of starting point, but in principle a poor choice merely slows convergence. ABC gives a sample from the posterior distribution of the parameters, but it does not give an explicit form of point estimates or of the posterior. The contribution of each summary statistic in ABC is unclear, while in the proposed method we only specify which statistics potentially contribute to the estimation of the parameters, but the amount by which they contribute is determined automatically by the method. The performance of the method cannot be compared directly with that of ABC, as the output of the two methods is different. However, it would be possible to compare the mean of an ABC posterior to the point estimate obtained by the proposed method. In very simple examples the results obtained from the two methods agree.

Indirect Inference methods require the derivation of a 'bridge relation' or 'binding function' which is used to compare the auxiliary statistics calculated from the data to those calculated from simulations, whereas the proposed method gives the optimal linear combination of the chosen statistics.

Finally, fitting of a model should in principle be accompanied by a check, informal or formal, of model adequacy. This aspect is discussed briefly in Chapter 3.

Chapter 3

Goodness of fit

3.1 Introduction

When $q > p$ there is the possibility of obtaining information about the adequacy of the assumed model, in effect by finding combinations that should, if the model is correct, have zero mean. This leads to a statistic broadly analogous to a residual sum of squares in a least-squares analysis.

When $q > p$, only p linear functions of Z are used for estimation and so it should be possible to find $q - p$ orthogonal linear combinations that locally at least are independent of the parameter and available to test goodness of fit. We essentially do this by finding a $q - p$ dimensional vector of zero mean, evaluating its covariance matrix and hence forming a chi-squared like statistic. In effect this is done by an analogue of analysis of variance in which the variability in the data is partitioned into a part associated with the fitted model and a residual compared with its expectation. We illustrate this first by an extremely simple example and then give one rather more realistic one.

3.2 A very simple example

Suppose that the data consist of a pair of observations (x_1, x_2), independently normally distributed with mean μ and unit variance. The optimal estimate of μ is their average $(x_1 + x_2)/2$, with variance $1/2$. The corresponding squared norm is $\left\{(x_1 + x_2)/2\right\}^2/(1/2)$. The squared norm of the full data is $x_1^2 + x_2^2$ and the difference $\left\{(x_1 - x_2)/2\right\}^2/(1/2)$ has under normal theory a chi-squared distribution with one degree of freedom. If the observations are independent but not normally distributed, the statistic has expectation one but not a chi-squared distribution and therefore retains some useful properties.

The value of the test statistic $\left\{(x_1 - x_2)/2\right\}^2/(1/2)$ was calculated for 10000 generated pairs of $N(0,1)$ observations. The QQ plot comparing those values to the quantiles of a χ^2_1 distribution is shown in Figure 3.2.1. The plot shows that the empirical distribution of the test statistic in this simple example agrees with its theoretical distribution.

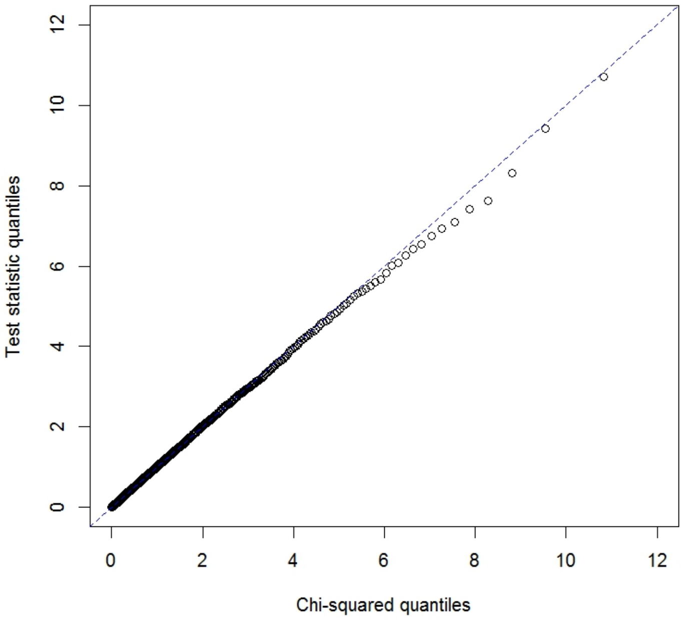

Figure 3.2.1: *QQ plot of test statistic* $\left\{(x_1 - x_2)/2\right\}^2/(1/2)$.

3.3 General case

Suppose that Z_T is the difference of the features from the overall simulation mean, $Z_T = Z_\mathcal{D} - \bar{Z}_0$. We have that $Y_\mathcal{D} = L^\mathrm{T} Z_\mathcal{D}$ and $\bar{Y} = L^\mathrm{T}\bar{Z}$, where $L = \Sigma^{-1}\bar{Z}D$. Let $Z_\mathcal{D} - \bar{Z}_0 = B(Y_\mathcal{D} - \bar{Y}_0) + R$, where B is a $q \times p$ matrix and R is a $q \times 1$ matrix.

A central result is that if a random vector Z has covariance matrix Σ, then $\|Z\|^2 = Z^\mathrm{T}\Sigma^{-1}Z$ is unchanged by the linear transformation of Z to AZ, where A is non-singular. This is because the covariance matrix of AZ is $A\Sigma A^\mathrm{T}$.

We have that $\Sigma_Y = \mathrm{cov}(Y_\mathcal{D}) = L^\mathrm{T}\Sigma L$ and is $p \times p$ and $\mathrm{cov}(Z_\mathcal{D}, Y_\mathcal{D}) = \Sigma L$

and is $q \times p$. The $q \times 1$ residual vector R is uncorrelated with $Y_{\mathcal{D}}$ under the model, i.e. $\mathrm{cov}(Z_{\mathcal{D}}, Y_{\mathcal{D}}) = B\mathrm{cov}(Y_{\mathcal{D}})$. Then $\Sigma L = BL^{\mathrm{T}}\Sigma L$ and

$$B = \Sigma L(L^{\mathrm{T}}\Sigma L)^{-1}. \tag{3.3.1}$$

The squared norm of the difference of the observed $Z_{\mathcal{D}}$ and the overall simulation mean $\bar{Z}_0$ is

$$\left\|Z_{\mathcal{D}} - \bar{Z}_0\right\|^2 = (Z_{\mathcal{D}} - \bar{Z}_0)^{\mathrm{T}}\Sigma^{-1}(Z_{\mathcal{D}} - \bar{Z}_0)$$

and

$$\begin{aligned} \left\|B(Y_{\mathcal{D}} - \bar{Y}_0)\right\|^2 &= \{B(Y_{\mathcal{D}} - \bar{Y}_0)\}^{\mathrm{T}}\Sigma^{-1}\{B(Y_{\mathcal{D}} - \bar{Y}_0)\} \\ &= (Y_{\mathcal{D}} - \bar{Y}_0)^{\mathrm{T}}B^{\mathrm{T}}\Sigma^{-1}B(Y_{\mathcal{D}} - \bar{Y}_0) \end{aligned}$$

and the matrix $B^{\mathrm{T}}\Sigma^{-1}B$ is equal to

$$(L^{\mathrm{T}}\Sigma L)^{-1}L^{\mathrm{T}}\Sigma\Sigma^{-1}\Sigma L(L^{\mathrm{T}}\Sigma L)^{-1} = (L^{\mathrm{T}}\Sigma L)^{-1} = {\Sigma_Y}^{-1},$$

so that

$$\left\|B(Y_{\mathcal{D}} - \bar{Y}_0)\right\|_{\Sigma}^{\ 2} = \left\|Y_{\mathcal{D}} - \bar{Y}_0\right\|_{\Sigma_Y}^{\ 2}.$$

Now $\mathbb{E}(R)$ cannot depend on θ for else $Y_{\mathcal{D}}$ would not be efficient. We assume that locally near $\theta = \theta_0$, the estimate of θ is based on $L^{\mathrm{T}}Z$, where L is $q \times p$, with covariance matrix $L^{\mathrm{T}}\Sigma L$ and hence $\|L^{\mathrm{T}}Z\|^2 = Z^{\mathrm{T}}L(L^{\mathrm{T}}\Sigma L)^{-1}L^{\mathrm{T}}Z$.

If $\theta = \theta_0$, then $\mathbb{E}(R) = 0$ and hence $\mathbb{E}(R) = 0$ locally and if the features are normally distributed $\|R\|^2$ has a χ^2 distribution with $q - p$ degrees of freedom. In general $\mathbb{E}(\|R\|^2) = q - p$, although if normality of the features does not hold the χ^2 distribution will not apply. Finally the test statistic is computed as

$$T = \|R\|^2 = \|Z_{\mathcal{D}}\|^2 - \|Y_{\mathcal{D}}\|^2 = Z_{\mathcal{D}}{}^{\mathrm{T}}\Sigma^{-1}Z_{\mathcal{D}} - Y_{\mathcal{D}}{}^{\mathrm{T}}\Sigma_Y{}^{-1}Y_{\mathcal{D}}. \tag{3.3.2}$$

To see the properties of T, first transform Z to Z^*, so that the covariance matrix is I. Then by orthogonal transformations of Z^*, make the first p components a linear combination of $L^{\mathrm{T}}Z_{\mathcal{D}}$ with expectations depending on θ and the last $q - p$ combinations have zero mean under the assumed local model. That is, choose $q - p$ further variables W to complete the transformation

$$\begin{pmatrix} Y_{\mathcal{D}} \\ W_{\mathcal{D}} \end{pmatrix} = \begin{pmatrix} L_{\mathrm{T}} & \mathbb{O} \\ \mathbb{O} & M \end{pmatrix} Z_{\mathcal{D}}.$$

We can choose M, for example by Gram–Schmidt orthogonalization, such that each W_j is uncorrelated with $Y_{\mathcal{D}}, W_1, \ldots, W_{j-1}$ and has unit variance. That is

$$\mathrm{cov}\begin{pmatrix} Y_{\mathcal{D}} \\ W_{\mathcal{D}} \end{pmatrix} = \begin{pmatrix} \Sigma_Y & \mathbb{O} \\ \mathbb{O} & \mathbb{I} \end{pmatrix}.$$

Thus, because squared norms are invariant under non-singular transformations,

$$\|Z_{\mathcal{D}}\|^2 = \left\| \begin{matrix} Y_{\mathcal{D}} \\ W_{\mathcal{D}} \end{matrix} \right\|^2 = Y_{\mathcal{D}}{}^{\mathrm{T}}\Sigma_Y{}^{-1}Y_{\mathcal{D}} + W_{\mathcal{D}}{}^{\mathrm{T}}W_{\mathcal{D}} = \|Y_{\mathcal{D}}\|^2 + \|W_{\mathcal{D}}\|^2.$$

If the model is correct and the linear approximation holds, $W_1, \ldots, W_{q-p}$ are uncorrelated random variables of unit variance.

The availability of such a check on model adequacy is, in principle at least, an advantage of the current method over ABC.

3.4 Example

Let $x_1, \ldots, x_n$ be a sample from a $N(\mu, 1)$ distribution. The method was applied to 10000 simulated datasets of size $n = 100$ with mean 5. Two iterations were run for each, with $\mu_0 = 5$, $h = 1$, $r = 1000$ and the features specified in Section 2.2.2. Thus 10000 values of the test statistic were generated. Figure 3.4.1 is a histogram of the values of the test statistic and figure (3.4.2) is a QQ plot of the quantiles of the test statistic against those of a chi-squared distribution with 2 degrees of freedom.

The plots show that the empirical distribution of the test statistic agrees with

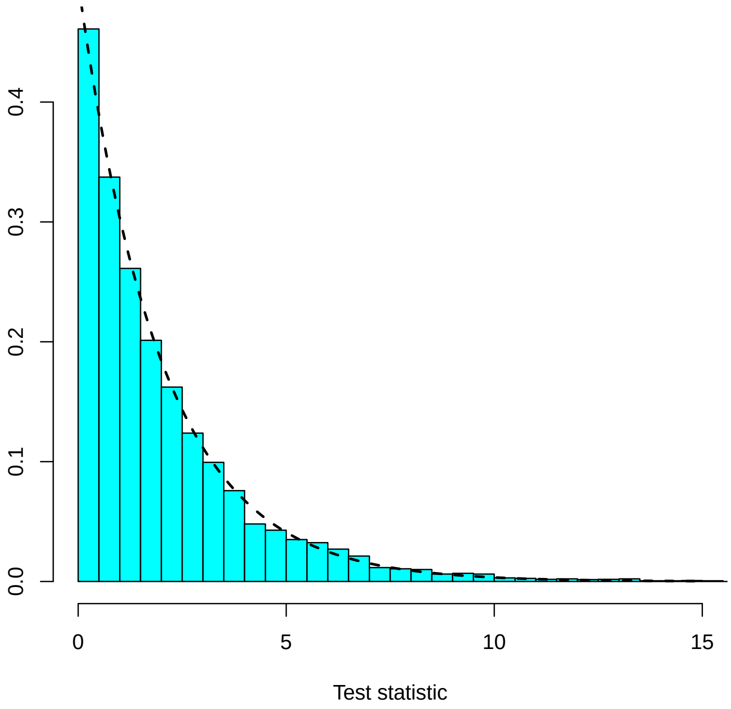

Figure 3.4.1: *Histogram of the values of the goodness of fit test statistic for a normal distribution with $n = 100$. The dashed line is the density function of the χ^2 distribution with 2 degrees of freedom.*

its theoretical distribution. The average of the values of the test statistic was 2.05 and the empirical variance 4.76.

For the first simple example the test statistic if calculated directly as $\frac{1}{2}(x_1 - x_2)^2$ has exactly a χ^2 distribution. The more complicated process implied by the use of the general method induces some distortion in the distribution, confined in this instance to the upper tail of values above the 5% point; see Figure 3.4.2.

For more complicated examples there might be some departures from the

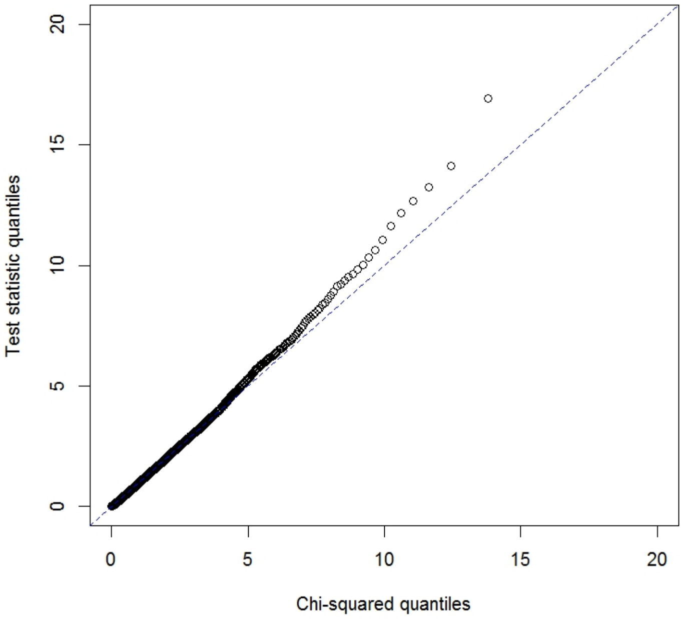

Figure 3.4.2: *QQ plot of the test statistic for a normal distribution with* $n = 100$*. The* 1% *point of the* χ^2_2 *distribution is* 9.21 *and the* 5% *point is* 5.99*.*

expected distribution of the test statistic.

Part II

Some issues connected with studies of dependence

Chapter 4

Direct and indirect effects in logistic regression and survival models

4.1 Introduction

In linear least squares regression there exists a simple decomposition of the effect of an exposure on an outcome into two parts in the presence of an intermediate variable. This decomposition is described and then analogous decompositions for other models are examined, in particular for logistic regression and proportional hazards models.

Let (Y, X, W) be random variables with finite variance. Let β_{YX} denote the coefficient of X in the linear least squares regression of Y on X and let

$\beta_{YX.W}$ denote the corresponding coefficient in the regression of Y on X and W, thus representing the dependence of Y on X given W. Then Cochran (1938) showed that

$$\beta_{YX} = \beta_{YX.W} + \beta_{YW.X}\beta_{WX}. \tag{4.1.1}$$

The *total effect* on Y of unit increase in X, β_{YX}, can thus be split into two parts, which can be interpreted as two pathways of dependence. The first is a contribution from X to Y with W fixed, sometimes called the *direct effect*, and the second is an *indirect effect* from X to W combined with the contribution of W to Y with X fixed. In the path diagram shown (Figure 4.1.1) these correspond respectively to a direct edge from X to Y and to a path from X to W and from W to Y. Equation (4.1.1) is sometimes called the *path formula*.

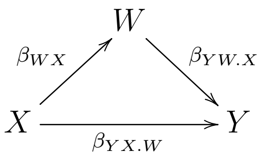

Figure 4.1.1: *Path diagram.*

In the epidemiological literature, for a situation where the associations shown in Figure 4.1.1 are assumed to hold, X is commonly referred to as the *exposure*, W as the *mediator* or *intermediate variable* and Y as the *outcome.*

The coefficient $\beta_{YX.W}$ is the direct (relative to the intermediate variable W)

or unmediated (by W) effect and $\beta_{YW.X}\beta_{WX}$ is the indirect or mediated effect. That is, $\beta_{YW.X}\beta_{WX}$ describes the part of the effect of X on Y which is explained by the path through W, while $\beta_{YX.W}$ is the effect of X on Y which is unexplained by W. The total effect β_{YX} includes both the path through W and the remaining unexplained effect (Figure 4.1.2). In a randomized experiment with X as the treatment and Y the outcome the standard analyses concentrate on β_{YX}.

$$X \xrightarrow[\beta_{YX}]{} Y$$

Figure 4.1.2: *Path diagram – total effect.*

For models other than linear regression, the above decomposition does not apply directly in such a simple form. Cox (2007) gave a generalization of (4.1.1) to quantile regression and pointed out the connection with notions of partial, directed and total differentiation.

The primary interest in most analyses is often in the total effect; however the decomposition may have explanatory power. In many applications the decomposition of an effect into an indirect path through a mediator and through a direct path with respect to the mediator, or more generally the decomposition of an effect into several path-specific effects, is of interest.

For example if X is a treatment and Y is a health outcome such as a cardiovascular event, there might be an intermediate outcome such as blood pressure and it might be of interest to estimate the proportion of the effect of the treatment on the health outcome which is mediated by a reduction in blood pressure and the proportion that affects directly the outcome in addition to the effect via the reduction of blood pressure.

Path analysis dates back to Wright (1921), who introduced path diagrams and considered linear relationships between variables.

This issue has been examined in the causal inference literature (Robins and Greenland, 1992; Pearl, 2001), which is based on counterfactuals and leads to general definitions of direct and indirect effects which can be estimated using special techniques such as the G-computation. Vansteelandt (2012) gives a review of approaches for estimating direct and indirect effects. Pearl (2012) described the mediation formula, which gives a decomposition of the total effect into a direct and indirect effect and is not model-specific.

Another approach is using structural equation models (Goldberger, 1972; Baron and Kenny, 1986) which provide methodology of defining and estimating such effects but become difficult for non linear models.

MacKinnon and Dwyer (1993) discuss measures of mediation using structural equation modelling approaches. Tein and MacKinnon (2003) examined estimates of mediated effects for survival data using a simulation study. MacKinnon et al. (2007) give a review of the mediation analysis literature and discuss concepts related to mediation.

There exist many applications in epidemiology (Lager et al., 2012; Ploubidis et al., 2013) and sociology (Kuha and Goldthorpe, 2010; Breen et al., 2013).

VanderWeele and Vansteelandt (2010) give a technique to estimate direct and indirect effect odds ratios that applies when the outcome is rare and the mediator continuous. Lange and Hansen (2011) and VanderWeele (2011) described a decomposition into direct and indirect effects for survival models. Aalen et al. (2012) discuss how the concepts of causal modelling and mediation can be viewed from a process point of view.

Muthén (2011) gives a review of mediation analysis in the causal inference framework using counterfactuals and its relation to the structural equation modelling approaches.

Gail et al. (1984) considered the problem of omitting one covariate from a generalized linear model and a Cox proportional hazards model. They compared estimating β from the model with linear predictor $\alpha + \beta x_1$ to estimating the effect of x_1 from the model with linear predictor $\alpha^* + \beta^* x_1 + \gamma x_2$ when x_1 is the treatment in a randomized experiment and found that for many models the estimator $\hat{\beta}$ based on the first model will be a biased estimator of the treatment effect β^* in the second model. They showed that for linear or exponential regression $\hat{\beta}$ is an asymptotically unbiased estimator of β^*, while for logistic regression models and Cox models with censored data $\hat{\beta}$ is a biased estimator of β^*; the bias for Cox regression depends on the amount of censoring. The estimate $\hat{\beta}$ from the model with the covariate omitted is closer to zero compared to the estimate $\hat{\beta}^*$ from the model with the covariates included. Struthers and Kalbfleisch (1986), Schumacher et al. (1987) and Bretagnolle and Huber-Carol (1988) also examined bias in Cox regression models and found results similar to those of Gail et al. (1984).

The motivation for the general discussion is to understand the relations between different analyses of the same set of data, to aid the comparison of analyses of different sets of data, possibly reported on a different model basis, as well as to understand when the path formulae of least squares regression are an adequate approximation.

In this Chapter various model-specific decompositions into direct and indirect effects are presented, for some commonly used models. The aim of the present Chapter is to study such models when marginalizing over a mediator and to give simple extensions of the path formula. First, in Section 4.2, Cochran's formula is presented for the cases in which the exposure and/or the mediator is a vector. In Section 4.3 a relationship analogous to Cochran's formula for logistic regression models is given. In Section 4.4 the corresponding formula for the case in which the mediator is binary but the outcome is continuous is given. Then a specific case in which the mediator consists of two components which are assumed to be associated in a specific way is presented in Section 4.5. In Section 4.6 an approximate version of the path formula for generalized linear models is given and in Section 4.7 the corresponding relationships for proportional hazards models are examined. Some of the results given in this Chapter will be illustrated using a particular dataset in Chapter 5.

4.2 Cochran's formula

4.2.1 Cochran's formula with multivariate exposure

Cochran's formula, given by equation (4.1.1), can easily be generalized to a multidimensional exposure $\mathbf{X}$. Suppose that

$$\begin{aligned} Y &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \epsilon \\ X_p &= \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x_{-p}} + \eta, \end{aligned} \tag{4.2.1}$$

where $\mathbf{x}$ and $\boldsymbol{\beta}$ are the $(p+1) \times 1$ vectors $\mathbf{x} = (1, x_1, \ldots, x_p)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\boldsymbol{\gamma}$ is the $p \times 1$ vector $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{p-1})^{\mathrm{T}}$, $\mathbf{x}_{-p}$ and $\boldsymbol{\beta}_{-p}$ are $p \times 1$ vectors consisting of the first p elements of $\mathbf{x}$ (a constant term and $p-1$ variables) and $\boldsymbol{\beta}$, respectively, and ϵ, η are normally distributed error terms with zero mean and variance σ^2_ϵ and σ^2_η, respectively.

Here we assume that Y depends on $X_1, \ldots, X_{p-1}$ and in particular X_p, while X_p depends on the other elements of $\mathbf{X}$, as in Figure 4.2.1, but with no stochastic dependence between the elements of $\mathbf{X}_{-p}$ and $X_1, \ldots, X_{p-1}$ are treated as fixed. We assume that there are no interaction terms in the models.

$$\begin{tikzcd}
 & X_p \arrow[dr, "\beta_p"] & \\
\mathbf{X}_{-p} \arrow[ur, "\gamma"] \arrow[rr, "\boldsymbol{\beta}_{-p}"'] & & Y
\end{tikzcd}$$

Figure 4.2.1: *Path diagram – multivariate exposure.*

Then from (4.2.1), we have that Y depends on $\mathbf{x}_{-p}$ via the linear regression model

$$\begin{aligned}
Y \mid \mathbf{x}_{-p} &= \boldsymbol{\beta}_{-p}^{\mathrm{T}}\mathbf{x}_{-p} + \beta_p x_p + \epsilon \\
&= \boldsymbol{\beta}_{-p}^{\mathrm{T}}\mathbf{x}_{-p} + \beta_p(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_{-p} + \eta) + \epsilon \\
&= (\boldsymbol{\beta}_{-p}^{\mathrm{T}} + \beta_p\boldsymbol{\gamma}^{\mathrm{T}})\mathbf{x}_{-p} + \beta_p\eta + \epsilon \\
&= (\boldsymbol{\beta}_{-p}^{\mathrm{T}} + \beta_p\boldsymbol{\gamma}^{\mathrm{T}})\mathbf{x}_{-p} + \epsilon^*, \qquad (4.2.2)
\end{aligned}$$

say, where $\epsilon^* \sim N(0,\ \sigma_\epsilon^2 + \beta_p\sigma_\eta^2)$. Therefore the generalization of (4.1.1) for a multivariate exposure is

$$\boldsymbol{\beta}^* = \boldsymbol{\beta}_{-p} + \beta_p\boldsymbol{\gamma}. \qquad (4.2.3)$$

4.2.2 Cochran's formula with multivariate mediator

Suppose now that the mediator $\mathbf{X}_{-p}$ is multidimensional, but there is a single exposure variable x_p. Let $\mathbf{x}$ and $\boldsymbol{\beta}$ be the $(p+1)\times 1$ vectors $\mathbf{x} = (1, x_1, \dots, x_p)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^{\mathrm{T}}$, let $\boldsymbol{\eta}$ be the $(p-1)\times 1$ vector $\boldsymbol{\eta} = (\eta_1, \dots, \eta_{p-1})$, $x_p^* = (1\ x_p)^{\mathrm{T}}$, $A = [\boldsymbol{\alpha}_0 \vdots \boldsymbol{\alpha}_1]$, where $\boldsymbol{\alpha}_0$ and $\boldsymbol{\alpha}_1$ are the $(p-1)\times 1$ vectors $\boldsymbol{\alpha}_0 = (\alpha_{01}, \alpha_{02}, \dots, \alpha_{0p-1})^{\mathrm{T}}$ and $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{12}, \dots, \alpha_{1p-1})^{\mathrm{T}}$. Let

$$\begin{aligned}
Y &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \epsilon \\
\mathbf{X}_{-p} &= Ax_p^* + \boldsymbol{\eta}, \qquad (4.2.4)
\end{aligned}$$

where $\boldsymbol{\eta} \sim N(0, \Sigma)$ and $\Sigma = \text{diag}\{\sigma^2_\eta\}$.

Here we assume that Y depends on $\mathbf{X}$, while $\mathbf{X}_{-p}$ depends on X_p, as illustrated by Figure 4.2.2, but with no stochastic dependence between the components of $\mathbf{X}_{-p}$ conditionally on X_p.

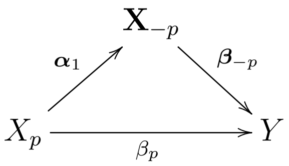

Figure 4.2.2: *Path diagram – multivariate mediator.*

Then from (4.2.4) we have that

$$\begin{aligned} Y \mid x_p &= \boldsymbol{\beta}^{\mathrm{T}}_{-p}\mathbf{x}_{-p} + \beta_p x_p + \epsilon \\ &= \boldsymbol{\beta}^{\mathrm{T}}_{-p}(\boldsymbol{\alpha}_0 + \boldsymbol{\alpha}_1 x_p + \boldsymbol{\eta}) + \beta_p x_p + \epsilon \\ &= \boldsymbol{\beta}^{\mathrm{T}}_{-p}\boldsymbol{\alpha}_0 + (\boldsymbol{\beta}^{\mathrm{T}}_{-p}\boldsymbol{\alpha}_1 + \beta_p)x_p + \boldsymbol{\beta}^{\mathrm{T}}_{-p}\boldsymbol{\eta} + \epsilon, \qquad (4.2.5) \end{aligned}$$

therefore

$$\boldsymbol{\beta}^*_0 = \boldsymbol{\beta}^{\mathrm{T}}_{-p}\boldsymbol{\alpha}_0$$

and the generalization of (4.1.1) for a multivariate mediator is

$$\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_{-p}^{\mathrm{T}}\boldsymbol{\alpha}_1 + \beta_p.$$

4.2.3 Cochran's formula with multivariate exposure and multivariate mediator

Suppose now that both the exposure and the mediator are vectors. We now denote the exposure by $\mathbf{X}$ ($p \times 1$) and the mediator by $\mathbf{Z}$ ($q \times 1$). Let $\boldsymbol{\beta}$ be a $p \times 1$ vector of coefficients, $\boldsymbol{\gamma}$ a $q \times 1$ vector of coefficients and A a $q \times p$ matrix of coefficients. Let $\boldsymbol{\eta}$ be a $q \times 1$ vector. Suppose that

$$\begin{aligned} Y &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z} + \epsilon \\ \mathbf{Z} &= \mathbf{A}\mathbf{x} + \boldsymbol{\eta}, \end{aligned} \tag{4.2.6}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $\boldsymbol{\eta} \sim N(0, \Sigma)$ with $\Sigma = \mathrm{diag}\{\sigma_\eta^2\}$.

Here we assume that Y depends on both $\mathbf{Z}$ and $\mathbf{X}$, and that $\mathbf{Z}$ depends on $\mathbf{X}$, as shown in Figure 4.2.3, but with no stochastic dependence between the elements of $\mathbf{X}$ or $\mathbf{Z}$.

Then from (4.2.6),

$$\begin{tikzcd}
& \mathbf{Z} \arrow[dr, "\gamma"] & \\
\mathbf{X} \arrow[ur, "A"] \arrow[rr, "\boldsymbol{\beta}"'] & & Y
\end{tikzcd}$$

Figure 4.2.3: *Path diagram – multivariate exposure and mediator.*

$$\begin{aligned}
Y \mid \mathbf{x} &= \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}(A\mathbf{x} + \boldsymbol{\eta}) + \epsilon \\
&= (\boldsymbol{\beta}^{\mathrm{T}} + \boldsymbol{\gamma}^{\mathrm{T}}A)\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\eta} + \epsilon. \qquad (4.2.7)
\end{aligned}$$

Therefore we have that the vector of coefficients of the regression of Y on $\mathbf{x}$ is

$$\boldsymbol{\beta}^{*} = \boldsymbol{\beta} + A\boldsymbol{\gamma}. \qquad (4.2.8)$$

A multivariate version of Cochran's formula has been presented, showing how the vector of coefficients in a linear regression model in which a multidimensional intermediate variable is omitted depends on the coefficients obtained from fitting the linear model on a multidimensional exposure and the intermediate variable and those from the linear model with the intermediate variable as the outcome. The vector of coefficients $\boldsymbol{\beta}$ in (4.2.8) is the direct effect of the exposures on the outcome, while $A\boldsymbol{\gamma}$ is the indirect effect that is mediated, or explained, by $\mathbf{Z}$.

4.3 Logistic regression

4.3.1 Logistic regression model with continuous mediator

An analogous decomposition is derived for logistic regression, when the outcome variable is binary and the intermediate variable is continuous. A different approximation for the same case is then derived and then the two approximations are compared. Then the first approach is generalized for the case in which the outcome and/or the mediator is multivariate and finally the case in which the mediator also is binary is considered.

We assume that there are not interactions between the explanatory variables in the models considered and for a more detailed interpretation there is the assumption that there is no unmeasured confounding.

Let Y be a binary and X_2 a continuous random variable. Let X_1 be either a continuous or discrete variable, treated as fixed. Let Y depend on X_1 and X_2 via the logistic regression model

$$\mathbb{P}(Y = 1 \mid X_1 = x_1, X_2 = x_2) = L(\beta_0 + \beta_1 x_1 + \beta_2 x_2), \tag{4.3.1}$$

where $L(x) = e^x/(1 + e^x)$ is the logistic function and let X_2 depend on X_1 via a linear regression model

$$X_2 = \gamma_0 + \gamma_1 x_1 + U, \tag{4.3.2}$$

where $U \sim N(0, \sigma^2)$ and U is uncorrelated with Y given X_1 and X_2. By substituting equation (4.3.2) into (4.3.1) and averaging over U we have

$$\mathbb{P}(Y = 1 \mid X_1 = x_1) = \mathbb{E}_U\big\{L\big((\beta_0 + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)x_1 + \beta_2 U\big)\big\}. \quad (4.3.3)$$

The expected value in (4.3.3) cannot in general be calculated analytically, but it can be evaluated numerically in any particular case. However, $L(x)$ can be approximated by $\Phi(kx)$, where $\Phi(\cdot)$ is the standard normal cumulative distribution function and k is a tuning constant. With $k = 0.607$ the approximation has a small relative error except when the probability of success is either very close to 0 or very close to 1. In the extremes the normal curve approaches its limit more rapidly than the logistic (Cox and Snell, 1989, p. 21–22). Therefore

$$\begin{aligned}\mathbb{P}(Y = 1 \mid X_1 = x_1) &\simeq \mathbb{E}_U\big\{\Phi\big(k(\beta_0 + \beta_2\gamma_0) + k(\beta_1 + \beta_2\gamma_1)x_1 + k\beta_2 U\big)\big\} \\ &= \int_{-\infty}^{\infty} f_U(u)\Phi(\alpha + k\beta_2 u)du, \qquad (4.3.4)\end{aligned}$$

where $k = 0.607$ and $\alpha = k(\beta_0 + \beta_2\gamma_0) + k(\beta_1 + \beta_2\gamma_1)x_1$.

To write the probability density function of U in terms of the standard normal probability density function $\phi(\cdot)$ we make the change of variable $V = U/\sigma$ and thus the integral in (4.3.4) becomes

$$
\begin{aligned}
\mathbb{E}_U\big\{\Phi(\alpha + k\beta_2 U)\big\} &= \int_{-\infty}^{\infty} \frac{1}{\sigma}\phi\left(\frac{u}{\sigma}\right)\Phi(\alpha + k\beta_2 u)du \\
&= \int_{-\infty}^{\infty} \frac{1}{\sigma}\phi\left(\frac{u}{\sigma}\right)\mathbb{P}\left(Z \leq \alpha + \beta\sigma\frac{u}{\sigma}\right)du,
\end{aligned}
$$

where $\beta = k\beta_2$ and $Z \sim N(0,1)$. Making the substitution $y = u/\sigma$ the previous expression becomes

$$
\begin{aligned}
& \int_{-\infty}^{\infty} \phi\left(y\right)\mathbb{P}\left(Z \leq \alpha + \beta\sigma y\right)dy \\
= & \int_{-\infty}^{\infty} \phi\left(y\right)\mathbb{P}\left(Z - \beta\sigma y \leq \alpha\right)dy \\
= & \ \mathbb{P}\left(Z - \beta\sigma Y \leq \alpha\right), \qquad (4.3.5)
\end{aligned}
$$

where $Y \sim N(0,1)$. Let $W' = Z - \beta\sigma Y$. Since Y has a standard normal distribution, W' follows a $N(0,\ 1 + \beta^2\sigma^2)$ distribution. Then

$$
W = \frac{W'}{\sqrt{1 + \beta^2\sigma^2}}
$$

has a standard normal distribution and the probability in (4.3.5) is equal to

$$\begin{aligned}\mathbb{P}\left(\sqrt{1+\beta^2\sigma^2}W \leq \alpha\right) &= \mathbb{P}\left(W \leq \frac{\alpha}{\sqrt{1+\beta^2\sigma^2}}\right) \\ &= \Phi\left(\frac{k(\beta_0+\beta_2\gamma_0)+k(\beta_1+\beta_2\gamma_1)x_1}{\sqrt{1+k^2\beta_2^2\sigma^2}}\right).\end{aligned}$$

By applying the approximation used above in reverse order we obtain

$$\mathbb{P}(Y=1 \mid X_1=x_1) \simeq L\left(\frac{\beta_0+\beta_2\gamma_0+(\beta_1+\beta_2\gamma_1)x_1}{\sqrt{1+k^2\beta_2^2\sigma^2}}\right). \tag{4.3.6}$$

Therefore we have a modified form of Cochran's formula, namely

$${\beta_0}^* = \frac{\beta_0+\beta_2\gamma_0}{\sqrt{1+k^2\beta_2^2\sigma^2}}, \quad {\beta_1}^* = \frac{\beta_1+\beta_2\gamma_1}{\sqrt{1+k^2\beta_2^2\sigma^2}}, \tag{4.3.7}$$

where ${\beta_0}^*$, ${\beta_1}^*$ are such that $\mathbb{P}(Y=1 \mid X_1=x_1) \simeq L({\beta_0}^*+{\beta_1}^*x_1)$.

Therefore if X_1 is the variable the effect of which on the outcome is of interest and we omit the other variable X_2 from a logistic regression model the estimated effect $\hat{\beta}_1^*$ of X_1 will be given by (4.3.7). The coefficients of the model omitting the intermediate variable are pushed closer to zero compared to the corresponding coefficients for linear regression, given by (4.1.1). Here the new coefficients will be deflated by the term $\sqrt{1+k^2\beta_2^2\sigma^2}$, which will be close to one if $\beta_2^2\sigma^2$ is small.

4.3.2 An alternative approximation

We now give a different approach which has the advantage of being applicable in principle to fairly general link functions, not only the logistic and others close to the Gaussian. An alternative approximation for the coefficient of x_1 in a logistic regression model after omitting a continuous intermediate variable x_2 from the model is shown. Let $\mathbb{P}(Y = 1 \mid X_1 = x_1, X_2 = x_2) = L(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ and $X_2 = \gamma_0 + \gamma_1 x_1 + \epsilon$, where $\mathbb{E}(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2$. Then

$$\mathbb{P}(Y = 1 \mid X_1 = x_1) = \mathbb{E}\left\{L(\alpha_0 + \alpha_1 x_1 + \beta_2 \epsilon)\right\},$$

where $\alpha_0 = \beta_0 + \beta_2\gamma_0$ and $\alpha_1 = \beta_1 + \beta_2\gamma_1$. We want β_0^* and β_1^* such that $\mathbb{P}(Y = 1 \mid X_1 = x_1) = L(\beta_0^* + \beta_1^* x_1)$. Using a Taylor expansion in $\beta_2\epsilon$ about 0,

$$\mathbb{P}(Y = 1 \mid X_1 = x_1) \simeq L(\alpha_0 + \alpha_1 x_1) + \frac{1}{2}\beta_2^2\sigma^2 L^{''}(\alpha_0 + \alpha_1 x_1). \qquad (4.3.8)$$

Suppose that we write (4.3.8) approximately as $L(\alpha_0 + \alpha_1 x_1 + \Delta)$. Using a first order Taylor expansion in Δ about 0,

$$L(\alpha_0 + \alpha_1 x_1 + \Delta) \simeq L(\alpha_0 + \alpha_1 x_1) + \Delta L^{'}(\alpha_0 + \alpha_1 x_1). \qquad (4.3.9)$$

Equating (4.3.8) and (4.3.9) we have that

$$\Delta = \frac{1}{2}{\beta_2}^2\sigma^2\frac{L^{''}(\alpha_0 + \alpha_1 x_1)}{L^{'}(\alpha_0 + \alpha_1 x_1)}.$$

Using the relationships

$$L^{'}(x) = L_0(x)L(x)$$

and

$$\frac{L^{''}(x)}{L^{'}(x)} = L_0(x) - L(x),$$

where $L_0(x) = 1 - L(x)$, we can write Δ as

$$\Delta = \frac{1}{2}{\beta_2}^2\sigma^2\left\{1 - 2L(\alpha_0 + \alpha_1 x_1)\right\}.$$

Thus from (4.3.8) we have that

$$\operatorname{logit}\mathbb{P}(Y = 1 \mid X_1 = x_1) \simeq \alpha_0 + \alpha_1 x_1 + \frac{1}{2}{\beta_2}^2\sigma^2\left\{1 - 2L(\alpha_0 + \alpha_1 x_1)\right\}$$

and β_1^*, the coefficient of x_1 in the model without x_2, will be

$$\beta_1^* = \frac{\partial}{\partial x_1} \operatorname{logit} \mathbb{P}(Y = 1 \mid X_1 = x_1) \simeq \alpha_1 \left\{ 1 - {\beta_2}^2 \sigma^2 \frac{e^{\alpha_0 + \alpha_1 x_1}}{(1 + e^{\alpha_0 + \alpha_1 x_1})^2} \right\}. \tag{4.3.10}$$

The function $L^{'}(x)$ has a maximum of $1/4$ at $x = 0$, therefore

$$\frac{e^{\alpha_0 + \alpha_1 x_1}}{(1 + e^{\alpha_0 + \alpha_1 x_1})^2} \leq \frac{1}{4}$$

and

$$1 - \beta_2^2 \sigma^2 \frac{\alpha_0 + \alpha_1 x_1}{(1 + e^{\alpha_0 + \alpha_1 x_1})^2} \geq 1 - \frac{1}{4} \beta_2^2 \sigma^2.$$

Then

$$\frac{\beta_1^*}{\alpha_1} \simeq 1 - \beta_2^2 \sigma^2 \frac{e^{\alpha_0 + \alpha_1 x_1}}{(1 + e^{\alpha_0 + \alpha_1 x_1})^2} \geq 1 - \frac{1}{4} \beta_2^2 \sigma^2,$$

that is,

$$\frac{\beta_1^*}{\beta_1 + \beta_2 \gamma_1} \gtrsim 1 - \frac{1}{4} \beta_2^2 \sigma^2,$$

where '$\gtrsim$' means 'greater than or approximately equal'. Because the function $L^{'}(\cdot)$ is always positive, we have that

$$\frac{\beta_1^*}{\beta_1 + \beta_2 \gamma_1} < 1$$

and thus the following inequality holds for β_1^*

$$|\beta_1 + \beta_2\gamma_1| \left(1 - \frac{1}{4}\beta_2^2\sigma^2\right) \lesssim |\beta_1^*| < |\beta_1 + \beta_2\gamma_1| \,. \tag{4.3.11}$$

This relates β_1^* to the corresponding coefficient obtained for linear least squares regression, $\beta_1 + \beta_2\gamma_1$.

4.3.3 Comparison of approximations

Two rather different approximations form the basis of the previous discussion. In their most simple form they concern

$$A(\mu, \sigma) = \mathbb{E}\left\{L(\mu + Z)\right\},$$

where Z is a normally distributed random variable with zero mean and standard deviation σ. We can write

$$A(\mu, \sigma) = \int_{-\infty}^{\infty} L(\mu + \sigma z)\phi(z)dz,$$

where $\phi(z)$ is the standard normal density function. Numerical integration was used to evaluate the integral $A(\mu, \sigma)$ and the two approximations

$$A_1(\mu, \sigma) = L\left(\frac{\mu}{\sqrt{1 + k^2\sigma^2}}\right) \tag{4.3.12}$$

and

$$A_2(\mu, \sigma) = L\left\{\mu + \frac{1}{2}\sigma^2\left(1 - 2L(\mu)\right)\right\} \tag{4.3.13}$$

were calculated. Equation (4.3.12) corresponds to the approximation given in Section 4.3.1 and (4.3.13) corresponds to that of Section 4.3.2. The following values of μ and σ were considered: $\mu = 0,\ 0.5,\ 1,\ 1.5,\ 2,\ 2.5,\ 3$ and $\sigma = 0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9,\ 1.1,\ 1.3,\ 1.5$. In Figure 4.3.1 the approximation to the difference between $\mathbb{E}\{L(\mu + \epsilon)\} - L(\mu)$ is plotted against the true value, for A_1 and A_2, respectively.

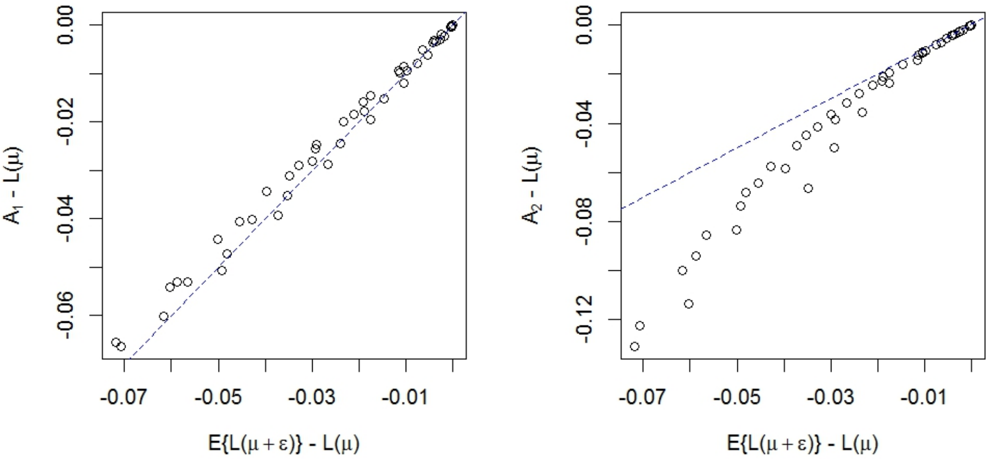

Figure 4.3.1: *Plot of the approximation to the difference between* $\mathbb{E}\{L(\mu + \epsilon)\} - L(\mu)$ *against the true value; the left plot corresponds to the approximation described in Section 4.3.1 and the right to that of Section 4.3.2.*

From the comparison of the exact values of $A(\mu, \sigma)$ with the two approximations we conclude that the first approximation A_1 gives reasonably accurate results over a range of values of $\mathbb{E}\{L(\mu + \epsilon)\} - L(\mu)$, whereas the second approximation A_2 overestimates the magnitude of that difference, seriously so if the difference $\mathbb{E}\{L(\mu+\epsilon)\} - L(\mu)$ exceeds about 0.05. The approximation could be improved either by including further terms in the expansion or by an empirical correction based on Figure 4.3.1.

The distinction between the two approaches is essentially that the second has high accuracy for very small σ^2, whereas the first can by choice of k be made to give reasonable accuracy over a much wider range.

4.3.4 Logistic regression – Multivariate case

Logistic regression with multivariate exposure and a single mediator

The decomposition given in Section 4.3.1 can easily be generalized to a multivariate exposure $\mathbf{x}$. Suppose now that $\mathbf{x}$ and $\boldsymbol{\beta}$ are $(p+1)\times 1$ vectors, $\mathbf{x} = (1, x_1, \dots, x_p)^{\mathrm{T}}$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^{\mathrm{T}}$ and that we have the multivariate logistic regression model

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = L(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p) = L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \qquad (4.3.14)$$

and let the mediator X_p depend on the rest of the remaining components of $\mathbf{X}$ through the linear model

$$X_p = \gamma_0 + \gamma_1 x_1 + \ldots + \gamma_{p-1} x_{p-1} + U,$$

where $U \sim N(0, \sigma^2)$, that is,

$$X_p = \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_{-p} + U, \tag{4.3.15}$$

where $\boldsymbol{\gamma}$ is the $p \times 1$ vector $(\gamma_0, \gamma_1, \dots, \gamma_{p-1})^{\mathrm{T}}$ and $\mathbf{x}_{-p}$ is a $p \times 1$ vector which consists of the first p elements of $\mathbf{x}$ (a constant term and $p-1$ variables), i.e. $\mathbf{x}_{-p} = (1, x_1, \dots, x_{p-1})^{\mathrm{T}}$. Let $\boldsymbol{\beta}_{-p}$ be the $p \times 1$ vector consisting of the first p elements of $\boldsymbol{\beta}$, that is, $\boldsymbol{\beta}_{-p} = (\beta_0, \beta_1, \dots, \beta_{p-1})^{\mathrm{T}}$.

The logistic regression model of Y on the first p elements of $\mathbf{x}$ is

$$\begin{aligned} \mathbb{P}(Y = 1 \mid \mathbf{X}_{-p} = \mathbf{x}_{-p}) &= \mathbb{E}_U\big\{L\big((\boldsymbol{\beta}_{-p} + \beta_p\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{x}_{-p} + \beta_p U\big)\big\}. \\ &\simeq \mathbb{E}_U\big\{\Phi\big(k(\boldsymbol{\beta}_{-p} + \beta_p\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{x}_{-p} + k\beta_p U\big)\big\}, \end{aligned}$$

where $k = 0.607$, approximating the logistic function by the normal cumulative distribution function as previously. Let $a = k(\boldsymbol{\beta}_{-p} + \beta_p\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{x}_{-p}$ and $b = k\beta_p$. Then

$$\mathbb{P}(Y = 1 \mid \mathbf{X}_{-p} = \mathbf{x}_{-p}) \simeq \mathbb{E}_U\big\{\Phi(a + bU)\big\} \tag{4.3.16}$$

and as in Section 4.3.1, we have that (4.3.16) is equal to

$$\mathbb{P}\left(W \le \frac{a}{\sqrt{1 + b^2\sigma^2}}\right) = \Phi\left(\frac{k(\boldsymbol{\beta}_{-p} + \beta_p\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{x}_{-p}}{\sqrt{1 + k^2\beta_p^2\sigma^2}}\right),$$

where W is a standard normal random variable, and by approximating the normal cumulative distribution function by the logistic function as before, we have that the new model averaged over the mediator is

$$\mathbb{P}(Y = 1 \mid \mathbf{X}_{-p}) \simeq L\left(\frac{(\boldsymbol{\beta}_{-p} + \beta_p \boldsymbol{\gamma})^{\mathrm{T}} \mathbf{x}_{-p}}{\sqrt{1 + k^2 \beta_p^2 \sigma^2}}\right). \tag{4.3.17}$$

Thus the multivariate version of the analogue of the path formula for logistic regression is

$$\boldsymbol{\beta}^* = \frac{\boldsymbol{\beta}_{-p} + \beta_p \boldsymbol{\gamma}}{\sqrt{1 + k^2 \beta_p^2 \sigma^2}}, \tag{4.3.18}$$

which is essentially the formula that holds for linear regression divided by the term $\sqrt{1 + k^2 \beta_p^2 \sigma^2}$. If the coefficient β_p of the mediator in the model in which the mediator is included and the error variance σ^2 of the linear regression model of the mediator on the remaining explanatory variables are small, the value of the new vector of coefficients will be approximated by the usual path formula for linear regression. As the term $\beta_p^2 \sigma^2$ becomes large, the new vector of coefficients will be pushed towards zero.

Logistic regression with multivariate exposure and multivariate mediator

This can be generalized to the case where the mediator is also a vector. Let Y be a binary response, $\mathbf{X}$ a $p \times 1$ vector and $\mathbf{Z}$ a $q \times 1$ vector, assumed to be associated in the way shown in Figure 4.3.2.

Suppose we now have the models

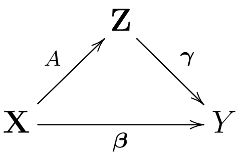

Figure 4.3.2: *Path diagram – multivariate exposure and mediator.*

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = L(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{z}) \tag{4.3.19}$$

and let the mediator $\mathbf{Z}$ depend on $\mathbf{x}$ according to the linear model

$$\mathbf{Z} = \mathbf{A}\mathbf{x} + \mathbf{U}, \tag{4.3.20}$$

where $\mathbf{U} \sim N(0, \Sigma)$, $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of coefficients, $\boldsymbol{\gamma}$ is a $q \times 1$ vector of coefficients and $\mathbf{A}$ is a $q \times p$ matrix of coefficients. We then obtain

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) = \mathbb{E}_{\mathbf{U}} \left\{ L\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{U}\right) \right\}. \tag{4.3.21}$$

Following the same procedure as before, the probability of Y being equal to 1 conditional on $\mathbf{x}$ only is

$$\mathbb{P}(Y = 1 \mid \mathbf{X} = \mathbf{x}) \simeq L \left\{ \frac{(\boldsymbol{\beta}^{\mathrm{T}} + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{A})\mathbf{x}}{\sqrt{1 + k^2 \boldsymbol{\gamma}^{\mathrm{T}} \Sigma \boldsymbol{\gamma}}} \right\}. \tag{4.3.22}$$

As before, the coefficient of the exposure obtained after averaging over the model which relates the mediator to the exposure is that obtained for least squares regression divided by a term which is greater than or equal to one, the coefficient decreasing as the variance of $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{U}$ increases.

4.3.5 Logistic regression with binary mediator

An analogous calculation to that of Section 4.3.1 is presented for the case in which the mediator also is a binary variable and its dependence on the exposure is described by a logistic regression model. Let Y, X_2 be binary random variables with values in $\{0, 1\}$ and let X_1 be either a continuous or discrete random variable. The probability of Y being equal to 1 given the values of X_1 and X_2 is assumed to be

$$\mathbb{P}(Y = 1 \mid X_1 = x_1, X_2 = x_2) = L(\beta_0 + \beta_1 x_1 + \beta_2 x_2) \tag{4.3.23}$$

and the probability of $X_2 = 1$ conditional on the value of X_1 is

$$\mathbb{P}(X_2 = 1 \mid X_1 = x_1) = L(\gamma_0 + \gamma_1 x_1). \tag{4.3.24}$$

We want to find the probability of Y being equal to 1 conditionally only on the value of X_1. This is

$$\begin{aligned}\mathbb{P}(Y=1\mid X_1=x_1) &= \mathbb{E}_{X_2}\{\mathbb{P}(Y=1\mid X_1,X_2)\}\\ &= \mathbb{P}(X_2=1\mid X_1=x_1)\mathbb{P}(Y=1\mid X_1=x_1,X_2=1)\\ &\quad +\mathbb{P}(X_2=0\mid X_1=x_1)\mathbb{P}(Y=1\mid X_1=x_1,X_2=0),\end{aligned}$$

by the law of total probability. This is then equal to

$$L(\gamma_0+\gamma_1x_1)L(\beta_0+\beta_1x_1+\beta_2)+\Big\{1-L(\gamma_0+\gamma_1x_1)\Big\}L(\beta_0+\beta_1x_1).$$

Therefore the probability of Y being equal to 1 given x_1 is

$$\mathbb{P}(Y=1\mid X_1=x_1)=L(\gamma_0+\gamma_1x_1)\Big\{L(\beta_0+\beta_1x_1+\beta_2)-L(\beta_0+\beta_1x_1)\Big\}+L(\beta_0+\beta_1x_1). \tag{4.3.25}$$

If β_2 is small, then $L(\beta_0+\beta_1x_1+\beta_2)\simeq L(\beta_0+\beta_1x_1)+\beta_2L'(\beta_0+\beta_1x_1)$. Thus in this case

$$\mathbb{P}(Y=1\mid X_1=x_1)\simeq\beta_2L(\gamma_0+\gamma_1x_1)L'(\beta_0+\beta_1x_1)+L(\beta_0+\beta_1x_1).$$

Let $p=L(\gamma_0+\gamma_1x_1)$. Then

$$\mathbb{P}(Y=1 \mid X_1=x_1) = pL(\beta_0+\beta_1x_1+\beta_2) + (1-p)L(\beta_0+\beta_1x_1). \quad (4.3.26)$$

Suppose that we can write (4.3.26) as $L(\beta_0+\beta_1x_1+\alpha\beta_2)$. Then for small $\alpha\beta_2$,

$$\mathbb{P}(Y=1 \mid X_1=x_1) \simeq L(\beta_0+\beta_1x_1) + \alpha\beta_2 L(\beta_0+\beta_1x_1) \qquad (4.3.27)$$

and from (4.3.26),

$$\mathbb{P}(Y=1 \mid X_1=x_1) \simeq p\beta_2\left\{L'(\beta_0+\beta_1x_1) + \frac{1}{2}\beta_2^2L''(\beta_0+\beta_1x_1)\right\} + L(\beta_0+\beta_1x_1). \qquad (4.3.28)$$

Equating (4.3.27) and (4.3.28) we find that

$$\alpha \simeq L(\gamma_0+\gamma_1x_1) + \frac{\beta_2}{2}\left\{1-2L(\beta_0+\beta_1x_1)\right\},$$

therefore

$$\mathbb{P}(Y=1 \mid X_1=x_1) = L\left(\beta_0+\beta_1x_1+\beta_2\left\{L(\gamma_0+\gamma_1x_1)+\frac{\beta_2}{2}\left\{1-2L(\beta_0+\beta_1x_1)\right\}\right\}\right) \qquad (4.3.29)$$

and

$$\begin{aligned}\beta_1^* &= \frac{\partial}{\partial x_1}\text{logit}\mathbb{P}(Y = 1 \mid X_1 = x_1) \\ &\simeq \beta_1 + \beta_2 \left\{\gamma_1 L'(\gamma_0 + \gamma_1 x_1) - \beta_1\beta_2 L'(\beta_0 + \beta_1 x_1)\right\}, \qquad (4.3.30)\end{aligned}$$

which is the coefficient of x_1 after omitting the binary mediator from the model.

If $L'(\cdot)$ takes values close to $1/4$, we have

$$\beta_1^* \simeq \beta_1 + \frac{\beta_2}{4}\left\{\gamma_1 - \beta_1\beta_2\right\}. \qquad (4.3.31)$$

This is an approximate result which relates the coefficient of an exposure in a logistic regression model for the outcome regressed on the exposure only, to the coefficient obtained by a logistic regression model which also includes the mediator, assuming that the mediator is related to the exposure via a logistic regression model.

4.4 Linear regression model with binary mediator

Now the case of linear least squares regression with a binary mediator is examined and an approximate relationship analogous to Cochran's formula is given. Let Y, X_1 and X_2 be random variables. Let Y be continuous, X_2

binary with values in $\{0, 1\}$ and X_1 be either a continuous or discrete random variable. Suppose that Y depends on x_1 and x_2 through a linear model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{4.4.1}$$

and that the probability of X_2 taking the value 1 conditional on the value of X_1 is

$$\mathbb{P}(X_2 = 1 \mid X_1 = x_1) = L(\gamma_0 + \gamma_1 x_1). \tag{4.4.2}$$

To find an expression for Y conditional only on the value of X_1, we take the expectation of (4.4.1) with respect to X_2 conditionally on $X_1 = x_1$,

$$\begin{aligned} Y &= \mathbb{E}_{X_2|x_1}(Y) = \mathbb{E}_{X_2|x_1}(\beta_0 + \beta_1 x_1 + \beta_2 X_2 + \epsilon) \\ &= (\beta_0 + \beta_1 x_1 + \beta_2 + \epsilon)\mathbb{P}(X_2 = 1 \mid x_1) + (\beta_0 + \beta_1 x_1 + \epsilon)\mathbb{P}(X_2 = 0 \mid x_1) \\ &= (\beta_0 + \beta_1 x_1 + \beta_2 + \epsilon)L(\gamma_0 + \gamma_1 x_1) + (\beta_0 + \beta_1 x_1 + \epsilon)\left\{1 - L(\gamma_0 + \gamma_1 x_1)\right\} \\ &= \beta_0 + \beta_1 x_1 + \beta_2 L(\gamma_0 + \gamma_1 x_1) + \epsilon. \qquad (4.4.3) \end{aligned}$$

If $\mathbb{E}(x_1) = 0$, then using a first-order Taylor expansion of (4.4.3) in $\gamma_1 x_1$ about 0 we have

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 x_1 + \beta_2 \left\{L(\gamma_0) + L'(\gamma_0)\gamma_1 x_1)\right\} + \epsilon \\
&= \left\{\beta_0 + \beta_2 L(\gamma_0)\right\} + \left\{\beta_1 + \beta_2\gamma_1 L'(\gamma_0)\right\} x_1 + \epsilon.
\end{aligned}
$$

Thus we have that in this case

$$\beta_0^* \simeq \beta_0 + \beta_2 L(\gamma_0)$$

and

$$\beta_1^* \simeq \beta_1 + \beta_2\gamma_1 L'(\gamma_0), \tag{4.4.4}$$

where β_0^* and β_1^* are respectively the intercept and coefficient of x_1 in the linear regression model of Y on x_1.

For the general case in which X_1 need not have mean zero, let $\mathbb{E}(X_1) = \mu$. Then (4.4.3) can be written as

$$
\begin{aligned}
Y &= \beta_0 + \beta_1\{\mu + (x_1 - \mu)\} + \beta_2 L\left\{\gamma_0 + \gamma_1\left(\mu + (x_1 - \mu)\right)\right\} + \epsilon \\
&= (\beta_0 + \beta_1\mu) + \beta_1(x_1 - \mu) + \beta_2 L\left\{(\gamma_0 + \gamma_1\mu) + \gamma_1(x_1 - \mu)\right\} + \epsilon.
\end{aligned}
$$

We have that $\mathbb{E}(X_1 - \mu) = 0$, therefore $\gamma_1(x_1 - \mu)$ can be assumed to be small and we can approximate $L\left\{(\gamma_0 + \gamma_1\mu) + \gamma_1(x_1 - \mu)\right\}$ by its first-order

Taylor expansion in $\gamma_1(x_1 - \mu)$ about zero. Then

$$\begin{aligned} Y &\simeq \beta_0 + \beta_1\{\mu + (x_1 - \mu)\} + \beta_2 \left\{L(\gamma_0 + \gamma_1\mu) + L'(\gamma_0 + \gamma_1\mu)\gamma_1(x_1 - \mu)\right\} + \epsilon \\ &\simeq \beta_0 + \beta_2 L(\gamma_0 + \gamma_1\mu) - \beta_2\gamma_1\mu L'(\gamma_0 + \gamma_1\mu) + \left\{\beta_1 + \beta_2\gamma_1 L'(\gamma_0 + \gamma_1\mu)\right\} x_1 + \epsilon. \end{aligned}$$

Therefore

$$\beta_0^* \simeq \beta_0 + \beta_2 L(\gamma_0 + \gamma_1\mu) - \beta_2\gamma_1\mu L'(\gamma_0 + \gamma_1\mu) \tag{4.4.5}$$

and

$$\beta_1^* \simeq \beta_1 + \beta_2\gamma_1 L'(\gamma_0 + \gamma_1\mu), \tag{4.4.6}$$

which is similar to Cochran's formula but here the indirect effect $\beta_2\gamma_1$ is multiplied by the term $L'(\gamma_0 + \gamma_1\mu)$. The function $L'(\cdot)$ has a maximum at $(0, \frac{1}{4})$ and takes positive values, thus when $\gamma_0 + \gamma_1\mu$ is close to zero, $\beta_2\gamma_1$ is multiplied by a number close to 0.25. As $\gamma_0 + \gamma_1\mu$ becomes larger in absolute value, $\beta_2\gamma_1$ is multiplied by a smaller number and thus the coefficient β_1^* of x_1 from the regression omitting x_2 gets closer to the coefficient β_1 from the larger model (4.4.1).

4.5 Two ‘causally ordered’ mediators

4.5.1 Introduction

Suppose we now have a more complicated setting with two mediators, one of which depends stochastically on the other, as illustrated by Figure 4.5.1.

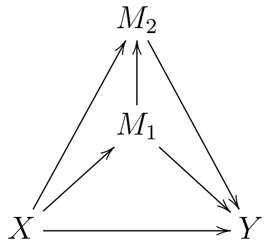

Figure 4.5.1: *Path diagram – two ‘causally ordered’ mediators.*

Thus there are three paths from X to Y. The direct effect from X to Y with respect to both mediators M_1 and M_2 (i.e. not mediated by either M_1 or M_2), the path through M_2 but not M_1, and the path through M_1, which itself consists of two parts.

A more complex version of this has been studied in the causal inference framework by Daniel et al. (2013).

4.5.2 Continuous outcome and mediators

The simplest case is when the outcome Y and the mediators M_1, M_2 are continuous and we use least squares linear regression to model the relationships between them as follows

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 m_1 + \beta_2 m_2 + \beta_3 x + \epsilon, && (4.5.1)\\
M_2 &= \gamma_0 + \gamma_1 m_1 + \gamma_2 x + \eta, && (4.5.2)\\
M_1 &= \alpha_0 + \alpha_1 x + \zeta, && (4.5.3)
\end{aligned}
$$

where $\epsilon \sim N(0, \sigma^2_\epsilon)$, $\eta \sim N(0, \sigma^2_\eta)$ and $\zeta \sim N(0, \sigma^2_\zeta)$.

From (4.5.2) and (4.5.3) we obtain

$$
\begin{aligned}
M_2 &= \gamma_0 + \gamma_1(\alpha_0 + \alpha_1 x + \zeta) + \gamma_2 x + \eta\\
&= (\gamma_0 + \gamma_1\alpha_0) + (\gamma_1\alpha_1 + \gamma_2)x + \gamma_1\zeta + \eta && (4.5.4)
\end{aligned}
$$

thus using (4.5.1), (4.5.3) and (4.5.4) we obtain the following expression for Y given x:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1(\alpha_0 + \alpha_1 + \zeta) + \beta_2\left\{(\gamma_0 + \gamma_1\alpha_0) + (\gamma_1\alpha_1 + \gamma_2)x + \gamma_1\zeta + \eta\right\} + \beta_3 x + \epsilon\\
&= \left\{\beta_0 + \beta_1\alpha_0 + \beta_2(\gamma_0 + \gamma_1\alpha_0)\right\} + \left\{\beta_1\alpha_1 + \beta_2(\gamma_1\alpha_1 + \gamma_2) + \beta_3\right\} x\\
&\qquad + (\beta_1 + \beta_2\gamma_1)\zeta + \beta_2\eta + \epsilon. && (4.5.5)
\end{aligned}
$$

Therefore the new model (4.5.5) is a linear regression model and the effect of x on Y is the sum of three contributions, $\beta_1\alpha_1 + \beta_2(\gamma_1\alpha_1 + \gamma_2) + \beta_3$, one of which consists of two parts which correspond to the two paths from X to M_2.

4.5.3 Binary outcome and continuous mediators

Consider now the case where the outcome Y is binary and the mediators M_1 and M_2 are continuous and suppose that we have a logistic regression of Y on M_1, M_2 and X and linear models for M_1 and M_2,

$$\mathbb{P}(Y \mid M_2 = m_2, M_1 = m_1, X = x) = L(\beta_0 + \beta_1 m_1 + \beta_2 m_2 + \beta_3 x), \quad (4.5.6)$$

$$M_2 = \gamma_0 + \gamma_1 m_1 + \gamma_2 x + \eta, \quad (4.5.7)$$

$$M_1 = \alpha_0 + \alpha_1 x + \zeta, \quad (4.5.8)$$

where $\eta \sim N(0, \sigma^2_\eta)$ and $\zeta \sim N(0, \sigma^2_\zeta)$.

From (4.5.7) and (4.5.8) we obtain the linear regression model of M_2 on x only,

$$\begin{aligned} M_2 &= \gamma_0 + \gamma_1(\alpha_0 + \alpha_1 x + \zeta) + \gamma_2 x + \eta \\ &= (\gamma_0 + \gamma_1\alpha_0) + (\gamma_1\alpha_1 + \gamma_2)x + \gamma_1\zeta + \eta \qquad (4.5.9) \end{aligned}$$

and then using (4.5.6), (4.5.8) and (4.5.9) we obtain

$$\begin{aligned}
\mathbb{P}(Y \mid X = x) &= \mathbb{E}\Big\{L\big(\beta_0 + \beta_1(\alpha_0 + \alpha_1 + \zeta) + \beta_2\{(\gamma_0 + \gamma_1\alpha_0) + (\gamma_1\alpha_1 + \gamma_2)x \\
&\qquad + \gamma_1\zeta + \eta\} + \beta_3 x\big)\Big\} \\
&= \mathbb{E}\Big\{L\big(\beta_0 + \beta_1\alpha_0 + \beta_2(\gamma_0 + \gamma_1\alpha_0) + \{\beta_1\alpha_1 + \beta_2(\gamma_1\alpha_1 + \gamma_2) + \beta_3\}x \\
&\qquad + (\beta_1 + \beta_2\gamma_1)\zeta + \beta_2\eta\big)\Big\}. \qquad (4.5.10)
\end{aligned}$$

Let $U = (\beta_1 + \beta_2\gamma_1)\zeta + \beta_2\eta$. Then $U \sim N(0, \tau^2)$, where $\tau^2 = (\beta_1 + \beta_2\gamma_1)^2\sigma_\zeta^2 + \beta_2^2\sigma_\eta^2$.

Let $\alpha_0^* = \beta_0 + \beta_1\alpha_0 + \beta_2(\gamma_0 + \gamma_1\alpha_0)$ and $\alpha_1^* = \beta_1\alpha_1 + \beta_2(\gamma_1\alpha_1 + \gamma_2) + \beta_3$. Then from (4.5.10),

$$\begin{aligned}
\mathbb{P}(Y = 1 \mid X = x) &= \mathbb{E}_U\left\{L(\alpha_0^* + \alpha_1^* x + U)\right\} \\
&\simeq L\left(\frac{\alpha_0^* + \alpha_1^* x}{\sqrt{1 + k^2\tau^2}}\right).
\end{aligned}$$

Thus

$$\beta_0^* = \frac{\beta_0 + \beta_1\alpha_0 + \beta_2(\gamma_0 + \gamma_1\alpha_0)}{\sqrt{1 + k^2\{(\beta_1 + \beta_2\gamma_1)^2\sigma_\zeta^2 + \beta_2^2\sigma_\eta^2\}}} \qquad (4.5.11)$$

and

$$\beta_1^* = \frac{\beta_1\alpha_1 + \beta_2(\gamma_1\alpha_1 + \gamma_2) + \beta_3}{\sqrt{1 + k^2\{(\beta_1 + \beta_2\gamma_1)^2\sigma_\zeta^2 + \beta_2^2\sigma_\eta^2\}}}. \tag{4.5.12}$$

Here the new coefficient β_1^* representing the effect of x on Y is again a sum of the three contibutions but pushed closer to zero compared to the corresponding coefficient yielded for linear regression. The smaller the variances of the error terms ζ and η, the closer the coefficient is to $\beta_1\alpha_1 + \beta_2(\gamma_1\alpha_1 + \gamma_2) + \beta_3$. The formulae given here can be generalized to several 'causally ordered' mediators.

4.5.4 Continuous outcome and binary mediators

Suppose now that the outcome Y is continuous and the mediators M_1, M_2 are binary. We have the models

$$Y = \beta_0 + \beta_1 m_1 + \beta_2 m_2 + \beta_3 x + \epsilon, \tag{4.5.13}$$

$$\mathbb{P}(M_2 = 1 \mid M_1 = m_1, X = x) = L(\gamma_0 + \gamma_1 m_1 + \gamma_2 x), \tag{4.5.14}$$

$$\mathbb{P}(M_1 = 1 \mid X = x) = L(\alpha_0 + \alpha_1 x), \tag{4.5.15}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$.

From (4.5.14) and (4.5.15) we obtain

$$\begin{aligned}\mathbb{P}(M_2 = 1 \mid X = x) &= \mathbb{E}_{M_1}\left\{\mathbb{P}(M_2 = 1 \mid M_1, X = x)\right\} \\ &= \mathbb{P}(M_1 = 1 \mid X = x)\mathbb{P}(M_2 = 1 \mid M_1 = 1, X = x) \\ &\quad +\mathbb{P}(M_1 = 0 \mid X = x)\mathbb{P}(M_2 = 1 \mid M_1 = 0, X = x) \\ &= L(\alpha_0 + \alpha_1 x)L(\gamma_0 + \gamma_1 + \gamma_2 x) + \left\{1 - L(\alpha_0 + \alpha_1 x)\right\} L(\gamma_0 + \gamma_2 x)\end{aligned} \tag{4.5.16}$$

thus using (4.5.13), (4.5.15) and (4.5.16) we obtain the following model for Y conditionally on x:

$$\begin{aligned}Y \mid x &= \mathbb{E}_{M_1,M_2}\left(\beta_0 + \beta_1 M_1 + \beta_2 M_2 + \beta_3 x + \epsilon\right) \\ &= \beta_0 + \beta_1\mathbb{P}(M_1 = 1 \mid X = x) + \beta_2\mathbb{P}(M_2 = 1 \mid X = x) + \beta_3 x + \epsilon \\ &= \beta_0 + \beta_1 L(\alpha_0 + \alpha_1 x) + \beta_2\Big\{L(\alpha_0 + \alpha_1 x)L(\gamma_0 + \gamma_1 + \gamma_2 x) \\ &\qquad + \left[1 - L(\alpha_0 + \alpha_1 x)\right] L(\gamma_0 + \gamma_2 x)\Big\} + \beta_3 x + \epsilon. \quad (4.5.17)\end{aligned}$$

Let $\mathbb{E}(X) = \mu$. Then

$$
\begin{aligned}
Y \mid x &= \beta_0 + \beta_1 L\left\{\alpha_0 + \alpha_1\mu + \alpha_1(x-\mu)\right\} \\
&\quad + \beta_2 L\left\{\alpha_0 + \alpha_1\mu + \alpha_1(x-\mu)\right\} L\left\{\gamma_0 + \gamma_1 + \gamma_2\mu + \gamma_2(x-\mu)\right\} \\
&\quad + \beta_2\left\{1 - L\left\{\alpha_0 + \alpha_1\mu + \alpha_1(x-\mu)\right\}\right\} L\left\{\gamma_0 + \gamma_1 + \gamma_2\mu + \gamma_2(x-\mu)\right\} + \beta_3 x + \epsilon.
\end{aligned}
\tag{4.5.18}
$$

We now have that $\mathbb{E}(X-\mu) = 0$, thus we can assume that $\alpha_1(x-\mu)$ and $\gamma_2(x-\mu)$ are sufficiently small, so that we can use a Taylor expansion in $\alpha_1(x-\mu)$ and $\gamma_2(x-\mu)$ about zero. We then have

$$
\begin{aligned}
Y \mid x &\simeq \beta_0 + \beta_1\left\{L(\alpha_0 + \alpha_1\mu) + L'(\alpha_0 + \alpha_1\mu)\alpha_1(x-\mu)\right\} \\
&\quad + \beta_2\left\{L(\alpha_0 + \alpha_1\mu) + L'(\alpha_0 + \alpha_1\mu)\alpha_1(x-\mu)\right\} \\
&\qquad \cdot\left\{L(\gamma_0 + \gamma_1 + \gamma_2\mu) + L'(\gamma_0 + \gamma_1 + \gamma_2\mu)\gamma_2(x-\mu)\right\} \\
&\quad + \beta_2\left\{1 - L(\alpha_0 + \alpha_1\mu) - L'(\alpha_0 + \alpha_1\mu)\alpha_1(x-\mu)\right\} \\
&\qquad \cdot\left\{L(\gamma_0 + \gamma_2\mu) + L'(\gamma_0 + \gamma_2\mu)\gamma_2(x-\mu)\right\} + \beta_3 x + \epsilon.
\end{aligned}
\tag{4.5.19}
$$

Assuming the terms including $(x-\mu)^2$ are negligible, we find that the model above can be simplified to

$$
Y \mid x = \beta_0^* + \beta_1^* x + \epsilon,
$$

where the new coefficients are

$$\begin{aligned}\beta_0^* \simeq\ & \beta_0 + \beta_1 L(\alpha_0+\alpha_1\mu) - \beta_1 L'(\alpha_0+\alpha_1\mu)\alpha_1\mu \\ & +\beta_2\Big\{L(\alpha_0+\alpha_1\mu)\{L(\gamma_0+\gamma_1+\gamma_2\mu) - L(\gamma_0+\gamma_2\mu)\} + L(\gamma_0+\gamma_1+\gamma_2\mu) \\ & -\mu\{L(\alpha_0+\alpha_1\mu)L'(\gamma_0+\gamma_1+\gamma_2\mu)\gamma_2 + L'(\alpha_0+\alpha_1\mu)L(\gamma_0+\gamma_1+\gamma_2\mu)\alpha_1 \\ & +L'(\gamma_0+\gamma_2\mu)\gamma_2 - L(\alpha_0+\alpha_1\mu)L'(\gamma_0+\gamma_2\mu)\gamma_2 - L'(\alpha_0+\alpha_1\mu)L(\gamma_0+\gamma_2\mu)\alpha_1\}\Big\}\end{aligned} \tag{4.5.20}$$

and

$$\begin{aligned}\beta_1^* \simeq\ & \beta_1 + L'(\alpha_0+\alpha_1\mu)\alpha_1 \\ & +\beta_2\Big\{L(\alpha_0+\alpha_1\mu)L'(\gamma_0+\gamma_1+\gamma_2\mu)\gamma_2 + L'(\alpha_0+\alpha_1\mu)L(\gamma_0+\gamma_1+\gamma_2\mu)\alpha_1 \\ & +L'(\gamma_0+\gamma_2\mu)\gamma_2 - L(\alpha_0+\alpha_1\mu)L'(\gamma_0+\gamma_2\mu)\gamma_2 - L'(\alpha_0+\alpha_1\mu)L(\gamma_0+\gamma_2\mu)\alpha_1\Big\} \\ & +\beta_3.\end{aligned} \tag{4.5.21}$$

Thus the effect of X on Y can be approximately written as a sum of the effects of the three paths from X to Y, but the coefficients attached to the different paths are specified in a way more complicated than in least squares regression.

4.5.5 Binary outcome and binary mediators

Suppose now that the outcome Y and the mediators M_1, M_2 are binary. We have the models

$$\mathbb{P}(Y=1 \mid M_2=m_2, M_1=m_1, X=x) = L(\beta_0+\beta_1 m_1+\beta_2 m_2+\beta_3 x), \tag{4.5.22}$$

$$\mathbb{P}(M_2=1 \mid M_1=m_1, X=x) = L(\gamma_0+\gamma_1 m_1+\gamma_2 x), \tag{4.5.23}$$

$$\mathbb{P}(M_1=1 \mid X=x) = L(\alpha_0+\alpha_1 x). \tag{4.5.24}$$

Then

$$\begin{aligned}
\mathbb{P}(Y=1 \mid X=x) &= \mathbb{E}_{M_1}\Big\{\mathbb{E}_{M_2}\left\{L(\beta_0+\beta_1 M_1+\beta_2 M_2+\beta_3 x)\right\}\Big\} \\
&= \mathbb{E}_{M_1}\Big\{L(\gamma_0+\gamma_1 m_1+\gamma_2 x)L(\beta_0+\beta_1 m_1+\beta_2+\beta_3 x) \\
&\quad +\left\{1-L(\gamma_0+\gamma_1 m_1+\gamma_2 x)\right\}L(\beta_0+\beta_1 m_1+\beta_3 x)\Big\} \\
&= \mathbb{P}(M_1=1 \mid X=x)\Big\{L(\gamma_0+\gamma_1+\gamma_2 x)L(\beta_0+\beta_1+\beta_2+\beta_3 x) \\
&\qquad +\left\{1-L(\gamma_0+\gamma_1+\gamma_2 x)\right\}L(\beta_0+\beta_1+\beta_3 x)\Big\} \\
&\quad +\mathbb{P}(M_1=0 \mid X=x)\Big\{L(\gamma_0+\gamma_2 x)L(\beta_0+\beta_2+\beta_3 x) \\
&\qquad +\left\{1-L(\gamma_0+\gamma_2 x)\right\}L(\beta_0+\beta_3 x)\Big\}.
\end{aligned} \tag{4.5.25}$$

Therefore we have that

$$\begin{aligned}\mathbb{P}(Y=1 \mid X=x) &= L(\alpha_0+\alpha_1 x)\Big\{L(\gamma_0+\gamma_1+\gamma_2 x)L(\beta_0+\beta_1+\beta_2+\beta_3 x) \\ &\qquad + \{1-L(\gamma_0+\gamma_1+\gamma_2 x)\}\, L(\beta_0+\beta_1+\beta_3 x)\Big\} \\ &\quad + \{1-L(\alpha_0+\alpha_1 x)\}\Big\{L(\gamma_0+\gamma_2 x)L(\beta_0+\beta_2+\beta_3 x) \\ &\qquad + \{1-L(\gamma_0+\gamma_2 x)\}\, L(\beta_0+\beta_3 x)\Big\}. \quad (4.5.26)\end{aligned}$$

The effect of an exposure on an outcome, mediated by two variables which are assumed to depend on each other in a particular way, has been decomposed into path-specific effects for the cases in which the mediators and/or the outcome are either continuous or binary. The approach described here can be extended to more complicated situations with more variables.

4.6 Approximate path formula for Generalized Linear Models

4.6.1 Univariate case

Let

$$\mathbb{E}(Y) = g^{-1}(\beta_1 w + \beta_2 z), \qquad (4.6.1)$$

where g is a link function and

$$\mathbb{E}(Z) = f^{-1}(\alpha w), \tag{4.6.2}$$

where f is a link function and W, Z and Y are assumed to be associated in the way shown in Figure 4.6.1.

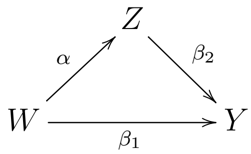

Figure 4.6.1: *Path diagram.*

Then $Z = f^{-1}(\alpha w) + \epsilon$, where ϵ is an error term with zero mean. Thus

$$\mathbb{E}(Y) = g^{-1}\left\{\beta_1 w + \beta_2 f^{-1}(\alpha w) + \beta_2 \epsilon\right\}.$$

The exact version of this would be

$$\begin{aligned} \mathbb{E}(Y) &= \mathbb{E}\left\{g^{-1}\left\{\beta_1 w + \beta_2 f^{-1}(\alpha w) + \beta_2 \epsilon\right\}\right\} \\ &= \int g^{-1}\left\{\beta_1 w + \beta_2 f^{-1}(\alpha w) + \beta_2 \epsilon\right\} h_\epsilon(\epsilon)\, d\epsilon, \end{aligned}$$

where $h_\epsilon(\cdot)$ is the density of ϵ (specified by f) and the integral is taken over the support of $h_\epsilon(\cdot)$.

If $\beta_2 \epsilon$ is small,

$$\begin{aligned}\mathbb{E}\left\{g^{-1}\left\{\beta_1 w+\beta_2 f^{-1}(\alpha w)+\beta_2\epsilon\right\}\right\} &\simeq g^{-1}\left\{\beta_1 w+\beta_2 f^{-1}(\alpha w)\right\}\\ &\quad+\frac{1}{2}\beta_2^2\sigma_\epsilon^2(g^{-1})''\left\{\beta_1 w+\beta_2 f^{-1}(\alpha w)\right\},\end{aligned}$$

where $\sigma_\epsilon^2=\mathrm{var}(\epsilon)$.

Let $\mathbb{E}(W)=\mu$. Then $f^{-1}(aw)$ can be written as $f^{-1}(\mu\alpha+(w-\mu)\alpha)$. We can assume that $(w-\mu)\alpha$ is 'small' and thus can use a Taylor expansion of $f^{-1}(\mu\alpha+(w-\mu)\alpha)$ in $(w-\mu)\alpha$ about 0. We have that

$$f^{-1}(\mu\alpha+(w-\mu)\alpha)\simeq f^{-1}(\mu\alpha)+(f^{-1})'(\mu\alpha)(w-\mu)\alpha.$$

Therefore

$$\begin{aligned}\mathbb{E}(Y) &\simeq g^{-1}\left\{\beta_1 w+(f^{-1})'(\mu\alpha)(w-\mu)\alpha\beta_2+f^{-1}(\mu\alpha)\beta_2+\beta_2\epsilon\right\}\\ &\simeq g^{-1}\Big\{\left\{\beta_1+\alpha\beta_2(f^{-1})'(\mu\alpha)\right\}w+\left\{f^{-1}(\mu\alpha)-\mu\alpha(f^{-1})'(\mu\alpha)\right\}\beta_2+\beta_2\epsilon\Big\}.\end{aligned} \tag{4.6.3}$$

That is,

$$\mathbb{E}(Y)\simeq g^{-1}(\beta_0^*+\beta_1^* w+\epsilon^*), \tag{4.6.4}$$

where

$$\beta_1^* = \beta_1 + \alpha\beta_2(f^{-1})'(\mu\alpha), \tag{4.6.5}$$

$$\beta_0^* = \left\{f^{-1}(\mu\alpha) - \mu\alpha(f^{-1})'(\mu\alpha)\right\}\beta_2$$

and

$$\epsilon^* = \beta_2\epsilon.$$

If the initial model is a generalized linear model (GLM), after marginalizing over the mediator the resulting model is not a GLM. If the resulting model is approximated by a GLM of the same family, the new coefficient of the exposure is by a first order approximation given by (4.6.5).

For example if (4.6.1) is a linear model and (4.6.2) a logistic regression model, thus g is the identity link and f the logit link, we obtain the approximation given by (4.4.6) in Section 4.4. If g is not the identity link, (4.6.5) implies a linear approximation to g^{-1}.

If for example (4.6.1) and (4.6.2) are exponential GLMs with $\mathbb{E}(Y \mid W = w, Z = z) = e^{\beta_1 w + \beta_2 z}$ and $\mathbb{E}(Z|W = w) = e^{\alpha w}$, then β_1^* will be approximately $\beta_1 + \alpha\beta_2 e^{\mu\alpha}$.

4.6.2 Multivariate case

Suppose we have a generalized linear model

$$\mathbb{E}(Y) = g^{-1}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}), \tag{4.6.6}$$

where g is the link function, Y is the response variable, $\mathbf{X}$ is a $(p+1) \times 1$ vector of explanatory variables including a constant term and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of coefficients. Suppose that $\mathbf{X}$ can be partitioned into a $q \times 1$ vector $\mathbf{W}$, an $r \times 1$ vector $\mathbf{Z}$ and a constant term, where $q + r = p$, that is,

$$\mathbf{X} = \left[1 \vdots \mathbf{W}^{\mathrm{T}} \vdots \mathbf{Z}^{\mathrm{T}}\right]^{\mathrm{T}},$$

where $\mathbf{Z}$ denotes the set of mediators and $\mathbf{W}$ the exposures. Suppose that $\mathbf{Z}$ depends on $\mathbf{W}$ through the generalized linear model

$$\mathbb{E}(\mathbf{Z}) = f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}). \tag{4.6.7}$$

Then

$$\mathbf{Z} = f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}) + \boldsymbol{\epsilon}, \tag{4.6.8}$$

where $\boldsymbol{\epsilon}$ is an $r \times 1$ error term with zero mean and $\boldsymbol{\alpha}$ is a $q \times r$ matrix of coefficients. Then

$$
\begin{aligned}
\mathbb{E}(Y) &= g^{-1}\left([\beta_0 \vdots \boldsymbol{\beta}_{\mathbf{1}}^{\mathrm{T}} \vdots \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}]\left[1 \vdots \mathbf{W}^{\mathrm{T}} \vdots \mathbf{Z}^{\mathrm{T}}\right]^{\mathrm{T}}\right) \\
&= g^{-1}\left([\beta_0 \vdots \boldsymbol{\beta}_{\mathbf{1}}^{\mathrm{T}} \vdots \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}]\left[1 \vdots \mathbf{W}^{\mathrm{T}} \vdots \{f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}) + \boldsymbol{\epsilon}\}^{\mathrm{T}}\right]^{\mathrm{T}}\right) \\
&= g^{-1}\left(\beta_0 + \boldsymbol{\beta}_{\mathbf{1}}^{\mathrm{T}}\mathbf{W} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}} f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}) + \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}\boldsymbol{\epsilon}\right).
\end{aligned}
$$

Let $\mathbb{E}(\mathbf{W}) = \boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is a $q \times 1$ vector. Then $\mathbf{W}$ can be written as $\boldsymbol{\mu} + (\mathbf{W} - \boldsymbol{\mu})$ and $f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{W}) = f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}[\boldsymbol{\mu} + (\mathbf{W} - \boldsymbol{\mu})])$ and the elements of $\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{W} - \boldsymbol{\mu})$ can be assumed to be small. Then using a Taylor expansion in $\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{W} - \boldsymbol{\mu})$ about 0,

$$
f^{-1}\left(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu} + \boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{W} - \boldsymbol{\mu})\right) \simeq f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu}) + \left\{\nabla f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu})\right\}\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{W} - \boldsymbol{\mu}), \tag{4.6.9}
$$

where $\nabla f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu})$ is the $r \times r$ diagonal matrix of partial derivatives of f^{-1}. Then

$$
\begin{aligned}
\mathbb{E}(Y) &\simeq g^{-1}\left(\beta_0 + \boldsymbol{\beta}_{\mathbf{1}}^{\mathrm{T}}\mathbf{W} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}\left\{f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu}) + \left[\nabla f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu})\right]\boldsymbol{\alpha}^{\mathrm{T}}(\mathbf{W} - \boldsymbol{\mu})\right\} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}\boldsymbol{\epsilon}\right) \\
&\simeq g^{-1}\Big(\beta_0 + \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}\left\{f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu}) - \left[\nabla f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu})\right]\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu}\right\} \\
&\quad + \left\{\boldsymbol{\beta}_{\mathbf{1}}^{\mathrm{T}} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}\left[\nabla f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu})\right]\boldsymbol{\alpha}^{\mathrm{T}}\right\}\mathbf{W} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathrm{T}}\boldsymbol{\epsilon}\Big).
\end{aligned} \tag{4.6.10}
$$

Thus

$$g\left\{\mathbb{E}(Y)\right\} = \beta_0^* + \boldsymbol{\beta}^{*\mathrm{T}}\mathbf{W} + \epsilon^*,$$

where the new vector of coefficients is given by

$$\boldsymbol{\beta}^* \simeq \boldsymbol{\beta_1} + \boldsymbol{\alpha}\left(\nabla f^{-1}(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\mu})\right)\boldsymbol{\beta_2}. \tag{4.6.11}$$

Various generalizations of the path formula for non-linear models, in particular logistic regression models, have been presented. A more approximate version for other link functions has been given.

4.7 Proportional hazards model

4.7.1 Proportional hazards model with continuous mediator

We now consider a time-to-event response and describe an analogous decomposition into direct and indirect effects for proportional hazards models. This was done for additive hazard models, where there is a close parallel with standard linear regression, by Aalen et al. (2008). We first consider the simplest case with a continuous mediator which depends via a linear least squares regression model on another variable. This is then generalized to a multivariate exposure and then the case in which a binary mediator is related to the exposure via a logistic regression model is discussed.

Let X_1, X_2 be explanatory variables in a proportional hazards model

$$h(t; X_1 = x_1, X_2 = x_2) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2) \tag{4.7.1}$$

and let X_2 depend on X_1 via a linear model

$$X_2 = \gamma_0 + \gamma_1 x_1 + U, \tag{4.7.2}$$

where U is an error term with mean zero. Note that this is different from inserting a random effect directly as a factor in (4.7.1) leading to a frailty model and deflation of the effect of the explanatory variable as t increases (Aalen et al., 2008). The survival function of T given x_1 and x_2 is given by

$$\begin{aligned} S(t; x_1, x_2) &= \exp\left\{ - \int_0^t h(\tau; x_1, x_2) d\tau \right\} \\ &= \exp\Big\{ - H_0(t) \exp\{\beta_1 x_1 + \beta_2 x_2\} \Big\}. \end{aligned}$$

where $H_0(t) = \int_0^t h_0(\tau)\, d\tau$ is the integrated baseline hazard. By substituting for x_2 from the corresponding linear model we have

$$S(t, u; x_1) = \exp\Big\{ - H_0(t) \exp\{\gamma_0\beta_2 + (\beta_1 + \beta_2\gamma_1)x_1 + \beta_2 u\} \Big\}.$$

Then averaging over U we have that the survival function given only x_1 is

$$\begin{aligned} S(t;x_1) &= \mathbb{E}_U\Big\{\exp\left\{-H_0(t)e^{\gamma_0\beta_2+(\beta_1+\beta_2\gamma_1)x_1}e^{\beta_2 U}\right\}\Big\} \\ &= \mathbb{E}_U\left\{e^{-\kappa e^{\beta U}}\right\}, \end{aligned} \tag{4.7.3}$$

where $\kappa = H_0(t)e^{\gamma_0\beta_2+(\beta_1+\beta_2\gamma_1)x_1}$ and $\beta = \beta_2$.

Let $V = e^{\beta U}$. We assume that V follows a Gamma distribution with parameters θ and ρ, so that U has a log gamma distribution. The density of V is given by

$$f_V(v) = \frac{\rho(\rho v)^{\theta-1}e^{-\rho v}}{\Gamma(\theta)}, \quad \text{where} \ \ \rho,\, \theta > 0, \quad v > 0. \tag{4.7.4}$$

Note that if $\beta = 0$, then $V \equiv 1$ and $\theta = \rho \to \infty$. From equation (4.7.3) we have

$$\begin{aligned} S(t) &= \int_0^\infty e^{-\kappa v}\frac{\rho(\rho v)^{\theta-1}e^{-\rho v}}{\Gamma(\theta)}\,dv \\ &= \frac{\rho^\theta}{\Gamma(\theta)}\int_0^\infty v^{\theta-1}e^{-(\kappa+\rho)v}\,dv. \end{aligned}$$

Making the substitution $(\kappa+\rho)v = w$ we obtain

$$
\begin{aligned}
S(t) &= \frac{\rho^\theta}{\Gamma(\theta)} \int_0^\infty \frac{1}{(\kappa+\rho)^{\theta-1}} w^{\theta-1} e^{-w} \frac{1}{\kappa+\rho}\, dw \\
&= \left(\frac{\rho}{\kappa+\rho}\right)^\theta,
\end{aligned}
$$

since $\Gamma(\theta) = \int_0^\infty w^{\theta-1}e^{-w}dw$. By substituting the expression for κ, we have that the survival function given x_1 is

$$
S(t;x_1) = \left(\frac{\rho}{H_0(t)e^{\gamma_0\beta_2+(\beta_1+\beta_2\gamma_1)x_1} + \rho}\right)^\theta . \tag{4.7.5}
$$

Given that the error U of the linear model has zero mean, a relationship between the parameters θ and ρ of the Gamma distribution can be found. We have that $V = e^{\beta U}$, thus $U = \frac{1}{\beta}\log V$. Therefore since $\mathbb{E}(U) = 0$ we have that $\mathbb{E}\left\{\frac{1}{\beta}\log V\right\} = 0$ and

$$
\int_0^\infty \frac{1}{\beta} \log v f_V(v)\, dv = 0,
$$

thus

$$
\frac{\rho^\theta}{\beta\Gamma(\theta)} \int_0^\infty v^{\theta-1}e^{-\rho v} \log v\, dv = 0.
$$

On making the substitution $\rho v = r$, the above equation becomes

$$
\frac{1}{\beta\Gamma(\theta)} \int_0^\infty r^{\theta-1}e^{-r} \log\left(\frac{r}{\rho}\right)\, dr = 0
$$

thus

$$\frac{1}{\beta\Gamma(\theta)}\left\{\int_0^\infty r^{\theta-1}e^{-r}\log r\,dr-\log\rho\int_0^\infty r^{\theta-1}e^{-r}\,dr\right\}=0.$$

But $\int_0^\infty r^{\theta-1}e^{-r}\log r\,dr=\Gamma'(\theta)$ and $\int_0^\infty r^{\theta-1}e^{-r}\,dr=\Gamma(\theta)$, hence

$$\frac{1}{\beta\Gamma(\theta)}\left\{\Gamma'(\theta)-(\log\rho)\Gamma(\theta)\right\}=0$$

and thus the relationship between ρ and θ is

$$\rho=e^{\psi(\theta)}, \tag{4.7.6}$$

where $\psi(\theta)=\Gamma'(\theta)/\Gamma(\theta)$ is the digamma function (Abramowitz and Stegun, 1964, p. 260). The hazard function that corresponds to the survival function given by (4.7.5) is

$$\begin{aligned} h(t;x_1) &= -\frac{d}{dt}\log S(t;x_1)\\ &= -\frac{d}{dt}\Big\{\theta\left\{\log\rho-\log\left[H_0(t)e^{\gamma_0\beta_2+(\beta_1+\beta_2\gamma_1)x_1}+\rho\right]\right\}\Big\}, \end{aligned} \tag{4.7.7}$$

thus the hazard function of T given x_1 is

$$h(t;x_1)=\frac{\theta h_0(t)e^{\gamma_0\beta_2+(\beta_1+\beta_2\gamma_1)x_1}}{H_0(t)e^{\gamma_0\beta_2+(\beta_1+\beta_2\gamma_1)x_1}+e^{\psi(\theta)}}. \tag{4.7.8}$$

We thus see that after marginalizing over the mediator the model we obtain is not a proportional hazards model. As the parameter $\rho = e^{\psi(\theta)}$ becomes close to zero, the dependence on x_1 decreases.

Let $\mathrm{var}(U) = \sigma^2$. We then have that

$$\begin{aligned}
\sigma^2 &= \mathrm{var}\left\{\frac{1}{\beta}\log V\right\} \\
&= \frac{1}{\beta^2}\mathbb{E}\left\{(\log V)^2\right\} \\
&= \frac{1}{\beta^2}\psi'(\theta). && (4.7.9)
\end{aligned}$$

This defines θ as a function of $\beta\sigma$.

For large θ, i.e. small σ^2,

$$\psi(\theta) \simeq \log\theta - \frac{1}{2\theta}$$

and

$$\psi'(\theta) \simeq \frac{1}{\theta} + \frac{1}{2\theta^2}$$

(Abramowitz and Stegun, 1964, p. 260). Since $\psi(\theta) = \log\rho$ and $\psi'(\theta) = \beta^2\sigma^2$,

$$\beta^2\sigma^2 \simeq \frac{1}{\theta} + \frac{1}{2\theta^2}$$

and approximately

$$\rho = \theta = \frac{1}{\beta^2\sigma^2}. \tag{4.7.10}$$

Thus for small $\beta_2^2\sigma^2$, (4.7.8) becomes

$$h(t; x_1) = \frac{h_0(t)e^{\lambda}}{1 + \beta_2^2\sigma^2 H_0(t)e^{\lambda}},$$

where $\lambda = \gamma_0\beta_2 + (\beta_1 + \beta_2\gamma_1)x_1$. Thus for small t, when $H_0(t)$ is small, the hazard takes a simple proportional form with the same modification of the regression coefficient as in a least squares analysis. For larger values of t the hazard is decreased and the influence of x_1 more complicated.

4.7.2 Proportional hazards model with continuous mediator – Multivariate case

The previous relationship in then generalized to a multivariate exposure. Let $\mathbf{x}$ be a $p \times 1$ vector of explanatory variables, $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$, and let $\boldsymbol{\beta}$ be a $p \times 1$ vector of coefficients, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$. We have the proportional hazards model

$$h(t; X_1 = x_1, \ldots, X_p = x_p) = h_0(t)\exp(\beta_1 x_1 + \ldots + \beta_p x_p),$$

that is,

$$h(t; \mathbf{X} = \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{x}). \tag{4.7.11}$$

Let X_p depend on the remaining elements of $\mathbf{x}$ via the linear model

$$X_p = \gamma_0 + \gamma_1 x_1 + \ldots + \gamma_{p-1} x_{p-1} + U,$$

that is,

$$X_p = \gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}} \mathbf{x}_{-p} + U, \tag{4.7.12}$$

where U is the error, assumed to have zero mean, and $\boldsymbol{\gamma}$ is the $(p-1) \times 1$ vector $(\gamma_1, \ldots, \gamma_{p-1})^{\mathrm{T}}$.

The survival function given $\mathbf{x}$ is

$$\begin{aligned} S(t; \mathbf{x}) &= \exp\left\{ -\int_0^t h(\tau; \mathbf{x}) d\tau \right\} \\ &= \exp\left\{ -H_0(t) \exp\left\{ \boldsymbol{\beta}^{\mathrm{T}} \mathbf{x} \right\} \right\}. \end{aligned}$$

Let $\mathbf{x}_{-p}$ and $\boldsymbol{\beta}_{-p}$ be $(p-1) \times 1$ vectors consisting of the first $p-1$ elements of $\mathbf{x}$ and $\boldsymbol{\beta}$, respectively, that is, $\mathbf{x}_{-p} = (x_1, \ldots, x_{p-1})^{\mathrm{T}}$ and $\boldsymbol{\beta}_{-p} = (\beta_1, \ldots, \beta_{p-1})^{\mathrm{T}}$. By writing x_p as a function of $\mathbf{x}_{-p}$ we have

$$S(t,u;\mathbf{x}_{-p}) = \exp\big\{ - H_0(t)\exp\big\{{\boldsymbol{\beta}_{-p}}^{\mathrm{T}}\mathbf{x}_{-p} + \beta_p(\gamma_0 + \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{x}_{-p} + u)\big\}\big\}$$

and averaging over U we have

$$S(t;\mathbf{x}_{-p}) = \mathbb{E}_U\left\{e^{-\kappa e^{\beta U}}\right\}, \tag{4.7.13}$$

where $\kappa = H_0(t)e^{\beta_p\gamma_0+(\boldsymbol{\beta}_{-p}+\beta_p\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{x}_{-p}}$ and $\beta = \beta_p$.

Let $V = e^{\beta U}$ have a Gamma(θ, ρ) density, given by

$$f_V(v) = \frac{\rho(\rho v)^{\theta-1}e^{-\rho v}}{\Gamma(\theta)}, \quad \text{where} \;\; \rho,\, \theta > 0, \quad v > 0. \tag{4.7.14}$$

Then from (4.7.13) we have, as previously,

$$S(t;\mathbf{x}_{-p}) = \left(\frac{\rho}{\kappa+\rho}\right)^{\theta},$$

that is,

$$S(t;\mathbf{x}_{-p}) = \left(\frac{\rho}{H_0(t)e^{\beta_p\gamma_0+(\boldsymbol{\beta}_{-p}+\beta_p\boldsymbol{\gamma})^{\mathrm{T}}\mathbf{x}_{-p}} + \rho}\right)^{\theta}. \tag{4.7.15}$$

As previously, we specify the distribution of V by requiring $\mathbb{E}(U) = 0$, which leads to

$$\rho = e^{\psi(\theta)}. \tag{4.7.16}$$

The hazard function which corresponds to the survival function given by (4.7.15) is

$$h(t; \mathbf{x}_{-p}) = \frac{\theta h_0(t) e^{\beta_p \gamma_0 + (\boldsymbol{\beta}_{-p} + \beta_p \boldsymbol{\gamma})^{\mathrm{T}} \mathbf{x}_{-p}}}{H_0(t) e^{\beta_p \gamma_0 + (\boldsymbol{\beta}_{-p} + \beta_p \boldsymbol{\gamma})^{\mathrm{T}} \mathbf{x}_{-p}} + e^{\psi(\theta)}}. \tag{4.7.17}$$

This is the hazard obtained after marginalizing over the mediator. When t is small, the effect of the exposure on the mediator is small or the residual variance of the model for the mediator is small, this can be approximated by a proportional hazards model with the coefficient of the explanatory variable approximately equal to that given by Cochran's formula.

4.7.3 Proportional hazards model with binary mediator

The case in which the mediator is a binary variable and depends on the exposure via a logistic regression model is considered.

Let T be a continuous non-negative time variable, X_2 a binary variable with values in $\{0, 1\}$ and X_1 either a continuous or discrete random variable. Suppose that the survival function of T given the values of X_1 and X_2 is

$$S(t; X_1 = x_1, X_2 = x_2) = \exp\left\{-H_0(t) e^{\beta_1 x_1 + \beta_2 x_2}\right\} \tag{4.7.18}$$

and the probability of $X_2 = 1$ given the value of X_1 is given by the logistic regression model

$$\mathbb{P}(X_2 = 1 \mid X_1 = x_1) = L(\gamma_0 + \gamma_1 x_1). \tag{4.7.19}$$

We want the distribution of T given only the value of X_1. The survival function conditionally on only X_1 is

$$\begin{aligned}
&S(t; X_1 = x_1) = \mathbb{E}_{X_2|X_1}\left\{S(t; X_1 = x_1, X_2)\right\} \\
&= \mathbb{E}_{X_2}\left\{\mathbb{P}(T > t \mid X_1 = x_1, X_2)\right\} \\
&= \mathbb{P}(X_2 = 1 \mid X_1 = x_1)\mathbb{P}(T > t \mid X_1 = x_1, X_2 = 1) \\
&\qquad\qquad + \mathbb{P}(X_2 = 0 \mid X_1 = x_1)\mathbb{P}(T > t \mid X_1 = x_1, X_2 = 0) \\
&= L(\gamma_0 + \gamma_1 x_1)\exp\left\{-H_0(t)e^{\beta_1 x_1 + \beta_2}\right\} + \left\{1 - L(\gamma_0 + \gamma_1 x_1)\right\}\exp\left\{-H_0(t)e^{\beta_1 x_1}\right\}.
\end{aligned} \tag{4.7.20}$$

The corresponding hazard function is then

$$\begin{aligned}
h(t; x_1) &= -\frac{d}{dt}\log S(t; x_1) \\
&= -\frac{-h_0(t)e^{\beta_1 x_1 + \beta_2}A - h_0(t)e^{\beta_1 x_1}B}{A + B} \\
&= h_0(t)e^{\beta_1 x_1}\left\{(e^{\beta_2} - 1)\frac{A}{A + B} + 1\right\}.
\end{aligned}$$

where $A = L(\gamma_0 + \gamma_1 x_1) \exp\left\{-H_0(t) e^{\beta_1 x_1 + \beta_2}\right\}$ and $B = \{1 - L(\gamma_0 + \gamma_1 x_1)\} \exp\left\{-H_0(t) e^{\beta_1 x_1}\right\}$. We have that

$$\frac{A}{A+B} = L\left\{\gamma_0 + \gamma_1 x_1 - H_0(t) e^{\beta_1 x_1}(e^{\beta_2} - 1)\right\}.$$

Therefore the hazard given x_1 is

$$h(t; x_1) = h_0(t) e^{\beta_1 x_1}\Big\{(e^{\beta_2} - 1) L\Big[\gamma_0 + \gamma_1 x_1 - H_0(t) e^{\beta_1 x_1}(e^{\beta_2} - 1)\Big] + 1\Big\}. \quad (4.7.21)$$

Suppose without essential loss of generality that X_1 is measured in such a way that $\beta_1 \geq 0$. If $\beta_1 = 0$, a dependence on x_1 is induced in (4.7.21) unless $\gamma_1 = 0$. That is, for $\beta_1 = 0$ the hazard is $h_0(t)\{(e^{\beta_2} - 1)L(\gamma_0 + \gamma_1 x_1 - H_0(t)(e^{\beta_2} - 1)) + 1\}$. If $\beta_1 > 0$, the dependence on x_1 for small t is determined by $e^{\beta_1 x_1 + \beta_2} L(\gamma_0 + \gamma_1 x_1)$, the hazard decreasing with t as $H_0(t)$ increases.

Assuming without loss of generality that $\beta_2 \geq 0$, because $0 < L(x) < 1$ for all x, we have that $0 \leq (e^{\beta_2} - 1)L(w) < e^{\beta_2} - 1$, where $w = \gamma_0 + \gamma_1 x_1 - H_0(t) e^{\beta_1 x_1}(e^{\beta_2} - 1)$. Thus $1 \leq (e^{\beta_2} - 1)L(w) + 1 < e^{\beta_2}$ and from (4.7.21) we have that

$$h_0(t) e^{\beta_1 x_1} \leq h(t; x_1) < h_0(t) e^{\beta_1 x_1 + \beta_2}.$$

When $\beta_2 = 0$, i.e. there is no effect of X_2, the hazard reduces to $h(t; x_1) = h_0(t) e^{\beta_1 x_1}$ as expected.

If $\beta_2 >> 0$ and so $e^{\beta_2} >> 1$, then

$$h(t; x_1) \simeq h_0(t)e^{\beta_1 x_1 + \beta_2} L\left(\gamma_0 + \gamma_1 x_1 - H_0(t)e^{\beta_1 x_1 + \beta_2}\right)$$

or equivalently

$$h(t; x_1) \simeq h_0^*(t)e^{\beta_1 x_1} L\left(\gamma_0 + \gamma_1 x_1 - H_0^*(t)e^{\beta_1 x_1}\right),$$

where $h_0^*(t) = e^{\beta_2} h_0(t)$ and $H_0^*(t) = e^{\beta_2} H_0(t)$.

Let $\alpha = e^{\beta_2} - 1$. Then

$$h(t; x_1) = h_0(t)e^{\beta_1 x_1}\Big\{\alpha L\Big[\gamma_0 + \gamma_1 x_1 - \alpha H_0(t)e^{\beta_1 x_1}\Big] + 1\Big\}.$$

- If t is suffieciently close to 0 and we assume that $H_0(t) \simeq 0$, we have that

  $$h(t; x_1) = h_0(t)g_1(x_1),$$

  where $g_1(x_1) = e^{\beta_1 x_1}\Big\{\alpha L(\gamma_0 + \gamma_1 x_1) + 1\Big\}$.

- If $H_0(t) \simeq 1$, then

  $$h(t; x_1) = h_0(t)g_2(x_1),$$

  where $g_2(x_1) = e^{\beta_1 x_1}\Big\{\alpha L(\gamma_0 + \gamma_1 x_1 - \alpha e^{\beta_1 x_1}) + 1\Big\}$.

- If t is large so that $H_0(t) >> 1$, then $L(\gamma_0 + \gamma_1 x_1 - \alpha H_0(t) e^{\beta_1 x_1}) \simeq 0$ and

$$h(t; x_1) \simeq h_0(t) e^{\beta_1 x_1}.$$

Exponential model

In the simplest special case of fiting an exponential model to T given x_1 and x_2 with constant baseline hazard $h_0(t) = \lambda$ and therefore cumulative baseline hazard $H_0(t) = \lambda t$, the hazard of T given only x_1 is

$$h(t; x_1) = \lambda e^{\beta_1 x_1} \Big\{ (e^{\beta_2} - 1) L \big[\gamma_0 + \gamma_1 x_1 - \lambda t e^{\beta_1 x_1} (e^{\beta_2} - 1) \big] + 1 \Big\} \tag{4.7.22}$$

and the corresponding survival function

$$S(t; x_1) = L(\gamma_0 + \gamma_1 x_1) e^{-\lambda t e^{\beta_1 x_1 + \beta_2}} + [1 - L(\gamma_0 + \gamma_1 x_1)] \, e^{-\lambda t e^{\beta_1 x_1}}. \tag{4.7.23}$$

For large t the time dependent part of the hazard function $L\big[\gamma_0 + \gamma_1 x_1 - \lambda t e^{\beta_1 x_1} (e^{\beta_2} - 1)\big]$ will be close to zero and thus the dependence of the hazard on x_2 will become very weak.

The model obtained by marginalizing over a mediator in a proportional hazards model has been derived for the cases in which the mediator is either continuous or binary.

Some of these results are illustrated using a dataset in Chapter 5.

Chapter 5

DNA methylation and prostate cancer mortality

5.1 Introduction

In this chapter the analysis of a particular set of data is described and an illustration of some aspects of the work described in Chapter 4 is provided.

Data on prostate cancer mortality were analysed to assess the relationships between the DNMT3b gene (as measured by the single nucleotide polymorphism (SNP) rs406193), DNA methylation in the tumour tissue, tumour aggressiveness (as measured by the Gleason score) and long-term prostate cancer mortality. The data, previously analysed by Richiardi et al. (2009) and Gillio-Tos et al. (2012), were on 438 prostate cancer patients diagnosed between 1982–1988 and between 1993–1996 in Turin, Italy. The primary outcome of interest was mortality from prostate cancer.

DNA methylation is a biochemical process that is important for normal development and cellular differentiation in higher organisms. The Gleason Grading system is used to help evaluate prostate cancer prognosis. Cancers with a higher Gleason score are more aggressive and have a worse prognosis.

Richiardi et al. (2009) found evidence that methylation in the APC gene is associated with prostate cancer mortality, particularly among individuals with a highly to moderately differentiated tumour. Gillio-Tos et al. (2012) found that the rs406193 single-nucleotide polymorphism (SNP) is associated with Gleason score and, via this effect, to prostate cancer mortality.

In Section 5.2 the data are described, in Section 5.3 some preliminary analyses and results are given and in Section 5.4 the results of Chapter 4 are applied to these data, to illustrate the relationships between the effects of an exposure on the outcome when adjusting for an intermediate variable and when not.

5.2 Data

The data, collected and originally analysed by Richiardi et al. (2009), consist of two cohorts of prostate cancer patients of any age diagnosed at a single pathology ward of the San Giovanni Battista Hospital in Turin, Italy (216 patients diagnosed from 1982 to 1988 and 243 patients diagnosed from 1993 to 1996). Patients enrolled were restricted to have residence in Turin and surroundings and the population includes mostly Caucasians. Patients received a biopsy of the prostate, transurethral resection of the prostate (TURP), or radical prostatectomy.

DNA was obtained from paraffin-embedded tumour tissue and evaluated for promoter methylation status in glutathione S-transferase (GSTP1), adenomatous polyposis coli (APC) and runt-related transcription factor 3 (RUNX3). Patients for which DNA extraction was not successful were removed from the analysis.

Death certificates were coded for prostate cancer-specific mortality blindly to tumour characteristics of the patients. Prostate cancer mortality assigned on the basis of death certificates may, for example, be overestimated if patients who died from other causes have their cause of death erroneously attributed to prostate cancer because of the underlying diagnosis, or the opposite.

Information including age, tumour grade and, limited to the 1990s cohort, Gleason score, was obtained from the pathology report for each patient. Patients with incorrect demographic information were excluded from the study. Diagnostic slides for patients in the 1980s cohort were traced and re-evaluated by a pathologist, who assigned the Gleason score. For eight tumours for which the slides could not be traced, information on tumour grade that was available in the pathology report was used; well-differentiated tumours were translated to a Gleason score of 6 or less, moderately differentiated tumours were assigned a score of 7, and poorly differentiated tumours were assigned a score of 8+.

Patients were followed up from the date of the pathology report to February

13, 2006 for the 1980s cohort and to January 15, 2007 for the 1990s cohort. Patients with no follow-up information were excluded from the study. Patients were censored on their date of death. Information on vital status and copies of the death certificates were obtained from demographic offices. Patients for which death certificates were not retrievable were removed from the analysis.

The final dataset used contains information on 438 patients. Information on methylation status of all three genes was available for 393 patients, while 45 patients had incomplete information. The exact numeric value of the Gleason score was missing for 10 patients, but we have information on whether it was less than 8 or not.

We have data on the rs406193 genotype (CC, CT or TT), the methylation status of each of the three genes APC, GSTP1 and RUNX (unmethylated or methylated), from which the number of methylated genes out of these three was derived for each patient, Gleason score (5–10), cohort (1982–1988 or 1993–1996), source of tumour tissue (biopsy (B), transurethral resection of the prostate (TURP), radical prostatectomy (RP)), age at prostate cancer diagnosis, date of birth, censoring date (date of death or date of last follow-up), vital status at last follow-up and whether death was due to prostate cancer.

Further details on the data and genotyping are given by Richiardi et al. (2009) and Gillio-Tos et al. (2012).

Analyses were performed based on the hypothesized causal relationships between the four main variables of interest: the DNMT3b variant (rs406193), DNA methylation, Gleason score and prostate cancer mortality. It was hypothesized that (i) DNMT3b activity may affect the methylation status of the three assessed genes (as well as that of several other genes), (ii) methylation status may affect the tumour morphology and thus the Gleason score, but not vice versa, and (iii) DNA methylation may affect mortality. The assumed pathways of dependence between the variables of interest are shown in Figure 5.2.1.

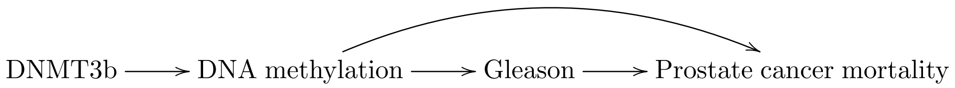

Figure 5.2.1: *Path diagram of the assumed relationships between the variables of interest.*

Descriptive characteristics of the data are shown in Table 5.2.1. Out of the 438 patients in the dataset, 189 patients (43%) died from prostate cancer. The genotype frequencies of rs406193 were CC: 54.6%, CT: 42.7% and TT: 2.7%. 188 (43%) of the patients had a Gleason score of at least 8 and 80 (18%) had promoter methylation in all three evaluated genes (APC, GSTP1 and RUNX3).

Survival time (years) – range	0.02–22.32
Median survival time (years)	4.95
Mortality	
Overall	368 (84 %)
Due to prostate cancer	189 (43 %)
Due to other causes	179 (41 %)
Mean age at diagnosis	71.59
Source of tumour tissue	
Biopsy	325 (74 %)
TURP	56 (13 %)
Radical prostatectomy	57 (13 %)
Period of diagnosis	
1982–1986	197 (44 %)
1993–1996	241 (55 %)
Number of methylated genes	
0–1	136 (31 %)
2	177 (40%)
3	80 (18 %)
missing	45
DNMT3b genotype	
CC	239 (55 %)
CT	187 (43 %)
TT	12 (3 %)
Gleason score	
< 8	250 (57 %)
≥ 8	188 (43 %)

Table 5.2.1: *Descriptive characteristics of the data.*

5.3 Preliminary analyses

To assess the effect of DNMT3b, DNA methylation and Gleason score on mortality, semiparametric proportional hazards models were fitted, with all possible combinations of the three explanatory variables of interest. The aim of this preliminary analysis was to investigate which of the variables of interest have a significant effect on prostate cancer mortality. Interactions were fitted one at a time but none was significant after correcting for multiple testing. The effects of DNMT3b and methylation on the Gleason score were assessed using linear and logistic regression models. The effect of DNMT3b on methylation was investigated by fitting an ordinal logistic regression model. In all models age, cohort and source of tumour tissue were included as covariates. The models fitted are shown in the top panel of Table 5.3.1. The genotype of rs406193 was used to give per-allele hazard ratios and the number of methylated genes was used as a three-level categorical variable, with levels 0–1, 2 and 3. Gleason score was used both as a numeric and as a binary variable indicating whether it is greater than or equal to 8. Time in years since prostate cancer diagnosis was used as the time scale of the survival models.

The effect of rs406193 on prostate cancer mortality was evaluated using a Cox proportional hazards model to estimate hazard ratios per copy of the rare allele. Individuals who died from causes other than prostate cancer were censored at death. To check whether the proportional hazards assumption was plausible, scaled Schoenfeld residuals (Schoenfeld, 1982) were examined graphically and a χ^2 test based on Schoenfeld residuals was used (Grambsch and Therneau, 1994).

Linear regression was used to estimate the effects of rs406193 and DNA methylation on Gleason score, as well as logistic regression models where Gleason score was dichotomised. Ordinal logistic regression was used to estimate the effect of rs406193 on the number of methylated genes.

Gleason score was used in the models either as a numeric variable, which takes integer values from 5 to 10, or as a binary variable, taking the value 1 if the Gleason score is greater than or equal to 8 and zero otherwise.

The results from the fitted models are summarized in the bottom panel of Table 5.3.1. No significant association was found between mortality and the rs406193 variant, or between methylation status and rs406193. An increase in the number of methylated genes is associated with increased prostate cancer mortality and increased Gleason score. Each copy of the T allele of rs406193 is associated with a decrease in Gleason score. There is no evidence of any significant association between rs406193 and the number of methylated genes. Gleason score is strongly associated with mortality. The associations between variables that were established from the data are shown in Figure 5.3.1.

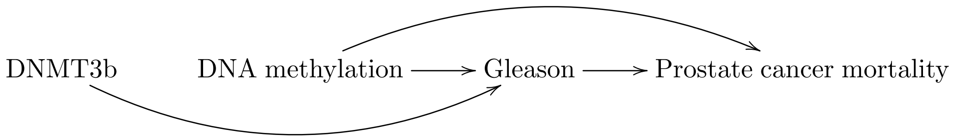

Figure 5.3.1: *Path diagram of the relationships between the variables as established by the data.*

A proportional odds model (ordinal logistic regression) was used to estimate the effect of the variant rs406193 on the number of methylated genes (Model 11). We used three levels for the dependent variable (0–1, 2 or 3 methylated genes out of APC, GSTP1 and RUNX3) and the odds ratio estimated is interpreted as the effect of one copy of the T allele of the rs406193 variant on each increase in the number of methylated genes. The results are shown in more detail in Table 5.3.2.

Adjusting for age, cohort and source of tumour tissue.	
1	mortality $\sim$ DNMT3b + methylation + Gleason
2	mortality $\sim$ DNMT3b + methylation
3	mortality $\sim$ DNMT3b
4	mortality $\sim$ DNMT3b + Gleason
5	mortality $\sim$ methylation + Gleason
6	mortality $\sim$ methylation
7	mortality $\sim$ Gleason
8a	Gleason $\sim$ DNMT3b + methylation
8b	Gleason (binary) $\sim$ DNMT3b + methylation
9a	Gleason $\sim$ methylation
9b	Gleason (binary) $\sim$ methylation
10a	Gleason $\sim$ DNMT3b
10b	Gleason (binary) $\sim$ DNMT3b
11	methylation $\sim$ DNMT3b

Model	per allele of DNMT3b		number of methylated genes				Gleason (numeric)	
			2		3			
	HR	p	HR	p	HR	p	HR	p
1	0.89	0.45	1.60	0.020	1.92	0.005	1.79	2.1×10^{-10}
2	0.81	0.16	1.73	0.006	2.12	0.001		
3	0.82	0.15						
4	0.93	0.63					1.78	1.3×10^{-11}
5			1.59	0.022	1.91	0.006	1.80	1.3×10^{-10}
6			1.72	0.007	2.11	0.001		
7							1.79	6.6×10^{-12}
	coef.	p	coef.	p	coef.	p		
8a	−0.18	0.040	0.31	0.005	0.29	0.039		
8b	−0.53	0.008	0.64	0.009	0.69	0.024		
9a			0.32	0.004	0.30	0.029		
9b			0.64	0.009	0.71	0.018		
10a	−0.23	0.006						
10b	−0.57	0.002						
11	−0.13	0.71						

Table 5.3.1: *Models fitted (top panel) and results from the fitted models (bottom panel). Time since diagnosis in years was used as the time scale in the survival models.*

DNMT3b rs406193	number of methylated genes 0–1 N (%)	 2 N (%)	 3 N (%)	OR	95% CI	p
CC	74 (54.4 %)	96 (54.2 %)	46 (57.5 %)	1		
CT	59 (43.4 %)	76 (42.9 %)	32 (40.0 %)	0.83	(0.56, 1.22)	0.63
TT	3 (2.2 %)	5 (2.8 %)	2 (2.5 %)	1.01	(0.31, 3.26)	
per allele				0.87	(0.62, 1.24)	0.44

Table 5.3.2: *Odds ratio of each increase in the number of methylated genes per copy of the T allele of DNMT3b rs406193 (OR obtained from proportional odds model (ordinal logistic regression) adjusted for age at diagnosis, cohort and source of tumour tissue).*

Logistic regression was used to estimate the odds ratio of having a high Gleason score (greater than or equal to 8). The results are shown in Table 5.3.3. The number of methylated genes was not included in this model so that the total effect is estimated, as the number of methylated genes was originally assumed to be an intermediate variable in the path between DNMT3b activity and Gleason score (Figure 5.2.1). If the number of methylated genes was adjusted for, the estimated effect of rs406193 on the dichotomised Gleason score would be the direct effect of rs406193 on Gleason.

DNMT3b rs406193	Gleason < 8 N (%)	Gleason ≥ 8 N (%)	OR	95% CI	p
CC	125 (50.0 %)	114 (60.6 %)	1		
CT	115 (46.0 %)	72 (38.3 %)	0.61	(0.41, 0.91)	0.016
TT	10 (4.0 %)	2 (1.1 %)	0.18	(0.04, 0.87)	0.033
per allele			0.57	(0.39, 0.82)	0.002

Table 5.3.3: *Odds ratio of a Gleason score of at least 8 for DNMT3b rs406193 (OR obtained from logistic regression adjusted for age at diagnosis, period of diagnosis and source of tumour tissue).*

Scaled Schoenfeld residuals (Grambsch and Therneau, 1994) for the model including rs406193, Gleason score, age at diagnosis, cohort and source of tumour tissue as covariates are shown in Figure 5.3.2. The curve plotted is a fitted natural spline and the dashed lines show the corresponding 95% confidence band. Since the curves are close to a horizontal line, the proportionality assumption seems plausible. An alternative approach would be to fit interactions with time.

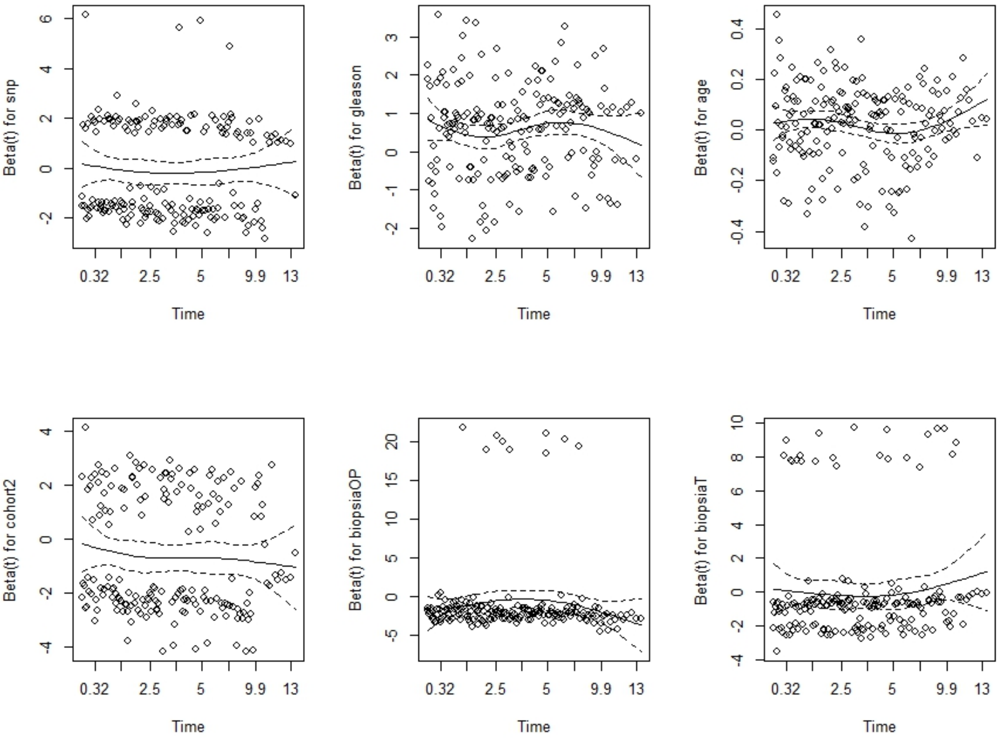

Figure 5.3.2: *Schoenfeld residuals for the model including DNMT3b, Gleason score and age at diagnosis as covariates, also adjusting for cohort and source of tumour tissue, using time since diagnosis (in years) as the time scale.*

The results summarized in Table 5.3.1 (Model 3) suggest that carriers of the rs406193 T allele might have had a (non-significantly) decreased risk of dying from prostate cancer (per allele hazard ratio (HR): 0.82, $p = 0.15$). After adjusting for Gleason score (Model 4), the HR increased to 0.93 ($p = 0.63$), suggesting that the effect of rs406193 on mortality might be partly explained by its effect on the Gleason score, that is, it might be mediated by Gleason score. The linear model for the effect of rs406193 on the Gleason score (Model 10a) shows an expected decrease in Gleason score for each copy of the T allele ($\hat{\beta} = -0.23$, $p = 0.006$). Similarly, the corresponding logistic regression model (Model 10b) yields significantly smaller odds of a high Gleason score for each copy of the T allele.

A model for the effect of rs406193 on mortality was fitted with separate parameters for the heterozygotes (CT) and the rare homozygotes (TT), the HR for CT was 0.80 ($p = 0.14$) and that for TT 0.78 ($p = 0.59$). When logistic regression was used to estimate the effect of being a CT or a TT carrier on the Gleason score ($<$ or ≥ 8), the odds ratios of having a Gleason score of at least 8 were 0.61 ($p = 0.016$) and 0.18 ($p = 0.033$), respectively, which supports the use of a multiplicative (per allele) model.

Various sensitivity analyses were done. An analysis with death from all causes as the event of interest was performed and the results are shown in Table 5.3.4. Also, an analysis with event of interest death from causes other than prostate cancer was performed. The results are shown in Table 5.3.5. The analysis on death from causes other than prostate cancer yields no association between any of the variables, as expected.

An alternative is to use age as the time scale in the proportional hazards model. However, when time is measured from birth, the risk of death of each individual changes at the time of disease initiation. This time point

Adjusting for age, cohort and source of tumour tissue.	
1	mortality $\sim$ DNMT3b + methylation + Gleason
2	mortality $\sim$ DNMT3b + methylation
3	mortality $\sim$ DNMT3b
4	mortality $\sim$ DNMT3b + Gleason
5	mortality $\sim$ methylation + Gleason
6	mortality $\sim$ methylation
7	mortality $\sim$ Gleason

Model	per allele of DNMT3b		number of methylated genes 2		3		Gleason (numeric)	
	HR	p	HR	p	HR	p	HR	p
1	1.04	0.72	1.15	0.282	1.36	0.057	1.30	4.7×10^{-5}
2	0.98	0.83	1.25	0.084	1.51	0.010		
3	0.94	0.53						
4	1.01	0.91					1.32	4.6×10^{-6}
5			1.16	0.276	1.37	0.055	1.30	5.0×10^{-5}
6			1.25	0.084	1.51	0.010		
7							1.31	4.3×10^{-6}

Table 5.3.4: *Results of the analysis of death from all causes, using time since diagnosis as the time scale.*

is different for each individual, therefore their hazards cannot reasonably be assumed to be proportional. This is why the primary analysis was performed with time measured from diagnosis to censoring (in years).

The analysis was repeated with time from birth in years (age) as the time scale. When using this time scale, the age at diagnosis variable does not seem to satisfy the proportional hazards assumption as before, as shown by plots of Schoenfeld residuals. Also the proportional hazards assumption for Gleason score seems to be problematic.

As there appears to be a problem with the proportional hazards assumption for age at diagnosis and possibly for the binary Gleason variable, a model

Adjusting for age, cohort and source of tumour tissue.	
1	mortality $\sim$ DNMT3b + methylation + Gleason
2	mortality $\sim$ DNMT3b + methylation
3	mortality $\sim$ DNMT3b
4	mortality $\sim$ DNMT3b + Gleason
5	mortality $\sim$ methylation + Gleason
6	mortality $\sim$ methylation
7	mortality $\sim$ Gleason

Model	per allele of DNMT3b		number of methylated genes				Gleason (numeric)	
			2		3			
	HR	p	HR	p	HR	p	HR	p
1	1.18	0.28	0.92	0.63	1.08	0.75	0.94	0.50
2	1.16	0.23	0.96	0.81	1.12	0.63		
3	1.08	0.60						
4	1.10	0.51					0.96	0.64
5			0.91	0.61	1.08	0.76	0.93	0.45
6			0.95	0.78	1.12	0.64		
7							0.95	0.61

Table 5.3.5: *Results of the analysis of death from causes other than prostate cancer, using time since diagnosis as the time scale.*

with an interaction between the two variables was fitted. This model seemed to fit well and the Schoenfeld residuals looked much closer to what is expected when the proportional hazard assumption is valid. To look into this further, the data were divided according to age at diagnosis, ≤ 70 and > 70, and by Gleason score, < 8 and ≥ 8. The four possible combinations were formed. Figure 5.3.3 shows the Kaplan–Meier estimates of the survival curves for the four groups.

The square of age at diagnosis was also included in all models and the results were compared to the ones without the squared term. The results did not change much after the inclusion of the squared term.

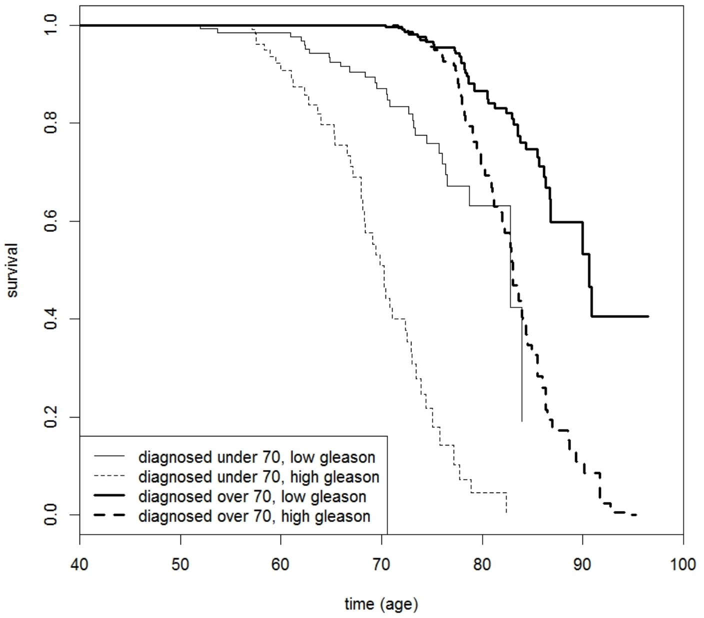

Figure 5.3.3: *Kaplan–Meier curves for model with time from birth (age) as the time scale.*

Figure 5.3.4 shows the Kaplan–Meier estimates of the survival curves for the four groups, divided by age at diagnosis and Gleason score as above, with time since diagnosis as the time scale. The Kaplan–Meier curves confirm the desirability of using time since diagnosis as the time scale.

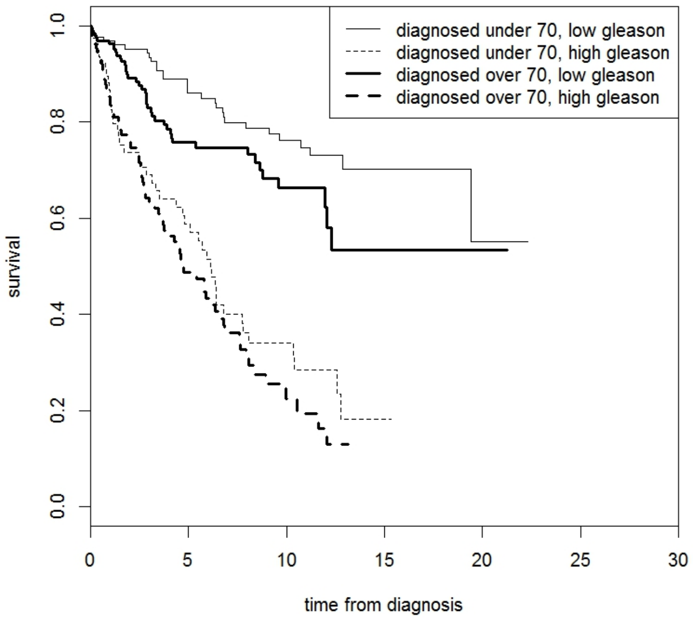

Figure 5.3.4: *Kaplan–Meier curves for model with time since diagnosis (in years) as the time scale.*

5.4 Further analyses

So far the analysis was based entirely on well-established methods. In this Section the results derived in Chapter 4 are illustrated using the prostate cancer data. Section 5.4.1 illustrates the relationship between the coefficients of logistic regression models with and without adjusting for an intermediate variable which was described in Section 4.3, while Sections 5.4.2 and 5.4.3 illustrate the corresponding relationship for proportional hazards models, the derivation of which was given in Section 4.7.

5.4.1 Application of result for logistic regression

In Section 4.3.1 an analogue of Cochran's formula for logistic regression with a continuous intermediate variable was derived. Here this relationship is illustrated using the prostate cancer data. We focus on the relationship between DNMT3b (rs406193), the number of methylated genes and Gleason score. The assumed relationships between these three variables are shown in Figure 5.4.1.

$$\begin{tikzcd} \text{DNMT3b} \arrow[r] \arrow[rr, bend left] & \text{DNA methylation} \arrow[r] & \text{Gleason score} \\ \mathbf{X_1} & \mathbf{X_2} & \mathbf{Y} \end{tikzcd}$$

Figure 5.4.1: *Path diagram of the relationship between DNMT3b, DNA methylation and Gleason score.*

Here, to illustrate the relationship that is analogous to the path formula for binary outcomes, Gleason score was treated as a binary variable (dichotomised between < 8 and ≥ 8), while DNMT3b (number of copies of the rare allele) and DNA methylation (number of methylated genes) were treated as numeric. To calculate the coefficients using equation (4.3.18), which is the analogue of the path formula for logistic regression with a continuous mediator, that is, it gives the coefficient of a vector $\mathbf{x}^*$ of explanatory variables which does not include the mediator, as a function of the coefficient of the mediator and of the coefficient of the full vector $\mathbf{x}$ which includes the mediator, we fitted a logistic regression model for Gleason score with both DNMT3b and DNA methylation as covariates and a linear regression model with DNA methylation as the response and DNMT3b as an explanatory variable. We also fitted a logistic regression model for Gleason score with DNA methylation omitted to compare the results. Age at diagnosis, cohort and type of biopsy were adjusted for in all models. That is, the following models were fitted:

$$\operatorname{logit}\{\mathbb{P}(Y=1 \mid \mathbf{x})\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \boldsymbol{\beta}_{3:6}^{\mathrm{T}} \mathbf{x}_{3:6} \tag{5.4.1}$$

where Y is the Gleason score, x_1 is the number of copies of the rare allele of DNMT3b, x_2 is the number of methylated genes and $\mathbf{x}_{3:6}$ is the vector $(x_3, \ldots, x_6)$ consisting of other covariates we adjust for (age at diagnosis, cohort and type of biopsy),

$$X_2 = \gamma_0 + \gamma_1 x_1 + \boldsymbol{\gamma}_{3:6}^{\mathrm{T}} \mathbf{x}_{3:6} + \epsilon \tag{5.4.2}$$

and

$$\operatorname{logit}\{\mathbb{P}(Y=1 \mid \mathbf{x}_{-2})\} = \beta_0^* + \beta_1^* x_1 + \boldsymbol{\beta}^{*\mathrm{T}}_{3:6} \mathbf{x}_{3:6}, \tag{5.4.3}$$

where $\mathbf{x}_{-2} = (x_1, x_3, \ldots, x_6)$. It is assumed that there are no interactions between the explanatory variables.

Using equation (4.3.18) with the estimated coefficients from models (5.4.1) and (5.4.2), we find the values of $(\beta_0^*, \beta_1^*, \beta_3^*, \ldots, \beta_6^*)$. The results are shown in Table 5.4.1, together with those obtained by fitting model (5.4.3) directly.

	Coefficients estimated from	
	Equation (4.3.18)	Model (5.4.3)
(Intercept)	-0.57	-0.10
x_1 (DNMT3b)	-0.53	-0.57
x_3 (age at diagnosis)	0.01	0.01
x_4 (cohort: 2)	-0.35	-0.39
x_5 (biopsy: RP)	-0.86	-0.82
x_6 (biopsy: TURP)	0.78	0.79

Table 5.4.1: *Coefficients calculated using the path decomposition for a logistic regression model with a continuous mediator and from fitting the corresponding logistic model without the mediator.*

The results obtained from equation (4.3.18) are very similar to those from fitting model (5.4.3), as expected, since if model (5.4.2) is correctly specified, the coefficients calculated using (4.3.18) should be approximately equal to those obtained from fitting model (5.4.3), from which X_2 is omitted. Thus in this context, (4.3.18) is reasonably adequate.

The total effect of the DNMT3b SNP on having a high Gleason score, adjusting for age at diagnosis, cohort and biopsy, is -0.53. The direct effect of DNMT3b on Gleason score is -0.51 and the indirect effect, mediated through DNA methylation is -0.02. Thus only a small proportion of the effect of the SNP on Gleason score is via DNA methylation.

5.4.2 Application of decomposition for proportional hazards models

The decomposition for proportional hazards models described in Section 4.7 was also applied to the prostate cancer data. We now consider the relationship between DNA methylation, Gleason score and mortality due to prostate cancer. The assumed relationships between the three variables are shown in Figure 5.4.2.

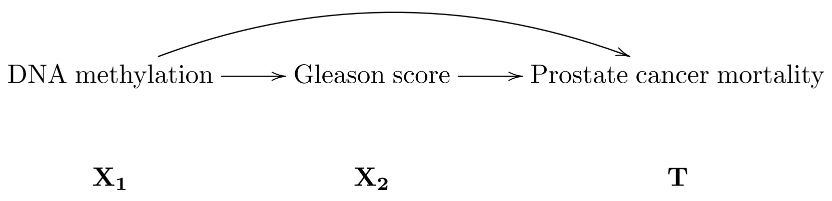

Figure 5.4.2: *Path diagram of the relationships between DNA methylation, Gleason score and mortality due to prostate cancer.*

The cumulative baseline hazard after fitting the model with the number of methylated genes, x_1, and Gleason score, x_2, as explanatory variables shown in Figure 5.4.3 is roughly a straight line, which supports the use of an exponential model. An exponential model has a constant baseline hazard $h_0(t) = \lambda$ and therefore $H_0(t) = \lambda t$. The hazard function is $h(t; \mathbf{x}) = \lambda e^{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}}$. The results obtained by fitting the parametric exponential model were very similar to those from the semi-parametric model. The results obtained from the two models are shown in Table 5.4.2.

	Exponential model			Semi-parametric model		
	HR	$\hat{\beta}$	s.e.($\hat{\beta}$)	HR	$\hat{\beta}$	s.e.($\hat{\beta}$)
(Intercept)	1.22×10^{-4}	−9.01	1.01			
Number of methylated genes	1.39	0.33	0.11	1.37	0.32	0.11
Gleason score	1.83	0.60	0.09	1.81	0.59	0.09
Age at diagnosis	1.02	0.02	0.01	1.02	0.02	0.01
Cohort: 2	0.59	−0.53	0.17	0.60	−0.52	0.17
Biopsy: RP	0.33	−1.10	0.37	0.34	−1.07	0.37
Biopsy: TURP	0.95	−0.05	0.24	0.95	−0.06	0.24

Table 5.4.2: *Results from exponential and semi-parametric model.*

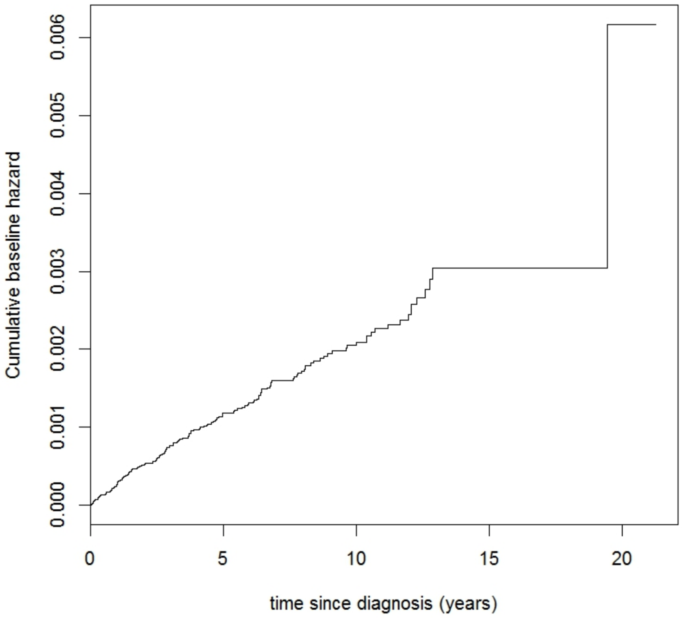

Figure 5.4.3: *Estimated cumulative baseline hazard for prostate cancer mortality.*

As described in Section 4.7, when we have time T as the outcome, an explanatory variable X_1 and a mediator X_2, after marginalizing over X_2 the hazard function can be written as

$$h^*(t) = \frac{\theta h_0(t) e^{\beta^* x_1}}{H_0(t) e^{\beta^* x_1} + e^{\psi(\theta)}},$$

where $\beta^* = \beta_1 + \beta_2\gamma_1$ and θ is the parameter related to the error term of the linear regression model of the mediator on the exposure. For small t $H_0(t)e^{\beta^* x_1}$ is negligible compared to $e^{\psi(\theta)}$. Therefore in this case the hazard will be

$$h(t) \simeq \frac{\theta}{e^{\psi(\theta)}} h_0(t) e^{\beta^* x_1}$$

which is a proportional hazards model with a different baseline hazard function and a coefficient for x_1 which is given by the usual path formula. If t is large, $e^{\psi(\theta)}$ will become small compared to $H_0(t)e^{\beta^* x_1}$ and then the dependence of the hazard on x_1 becomes small and eventually disappears.

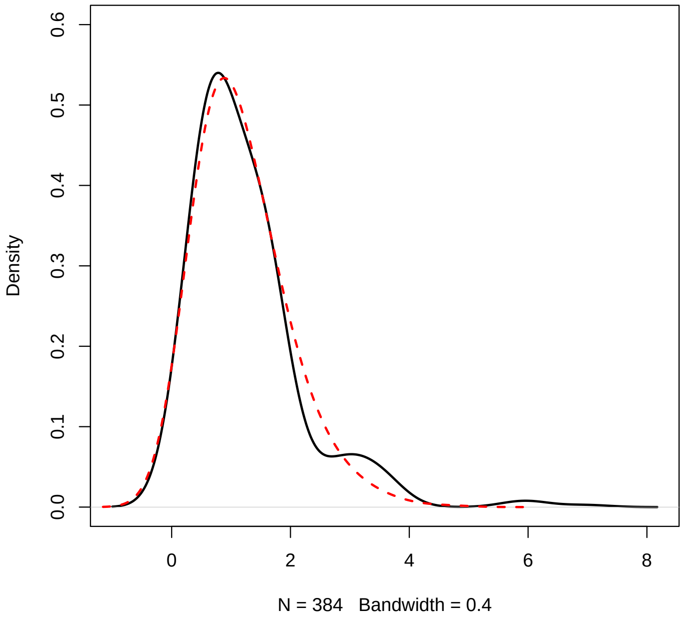

Figure 5.4.4: *Smoothed density of $e^{\beta_2 U}$ (black line) and Gamma density with estimated parameters (red dashed line).*

We had assumed in Section 4.7 that $V \sim \text{Gamma}(\theta, \rho)$, where $V = e^{\beta_2 U}$, U is the error of the linear regression model $X_2 = \gamma_0 + \gamma_1 x_1 + U$ and β_2 is the coefficient of x_2 from the proportional hazards model for the outcome T, $h(t) = h_0(t)e^{\beta_1 x_1 + \beta_2 x_2}$. To check whether the assumption that $e^{\beta_2 U}$ follows a

Gamma distribution is plausible, the distribution of the observed values of $e^{\beta_2 U}$ was compared to the theoretical Gamma$(\tilde{\theta}, \tilde{\rho})$ density, where $\tilde{\theta}$ and $\tilde{\rho}$ are the estimated Gamma parameters. The variance of the residuals of the linear regression model was used to estimate the parameters of the Gamma distribution using (4.7.10). Figure 5.4.4 shows the smoothed observed density of $e^{\beta_2 U}$ and the corresponding theoretical Gamma density. Figure 5.4.5 is a QQ plot of the quantiles of $e^{\beta_2 U}$ against the theoretical Gamma quantiles. These show reasonable agreement of the approximation with the observed density.

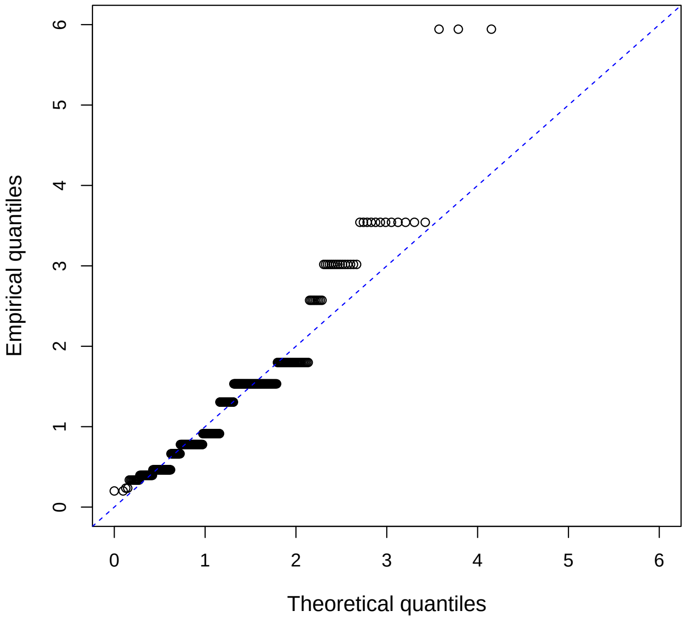

Figure 5.4.5: *QQ plot of quantiles of $e^{\beta_2 u}$ against theoretical quantiles of Gamma density with estimated parameters. The 1% point of the theoretical distribution is 3.08.*

In one approach we used an exponential model with hazard function $h(t; x_1, x_2) =$

$\lambda e^{\beta_1 x_1 + \beta_2 x_2}$. The parameter λ of the exponential model was estimated to be 1.84×10^{-4}.

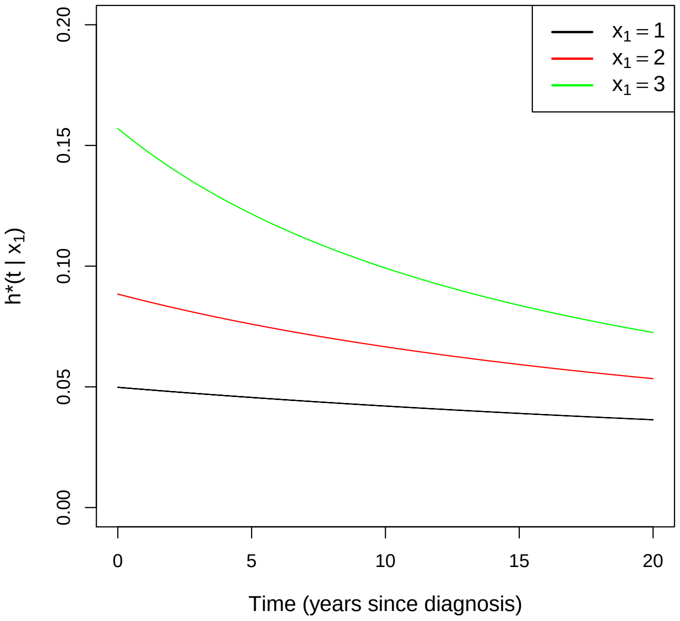

Figure 5.4.6: *Estimated hazard function after marginalizing over a continuous mediator (Gleason score) in an exponential model, calculated using equation* (5.4.4) *for each value of X_1, the number of methylated genes.*

A plot of the hazard function (Figure 5.4.6) given by

$$h(t; x_1) = \frac{\theta \lambda e^{\gamma_0 \beta_2 + (\beta_1 + \beta_2 \gamma_1) x_1}}{\lambda t e^{\gamma_0 \beta_2 + (\beta_1 + \beta_2 \gamma_1) x_1} + e^{\psi(\theta)}} \tag{5.4.4}$$

for each value of X_1, where X_1 is the number of methylated genes, and takes values in $\{1, 2, 3\}$, shows how the dependence on x_1 changes over time. The dependence of the hazard on x_1 decreases very quickly and eventually the three hazards come very close to each other. If a proportional hazards model

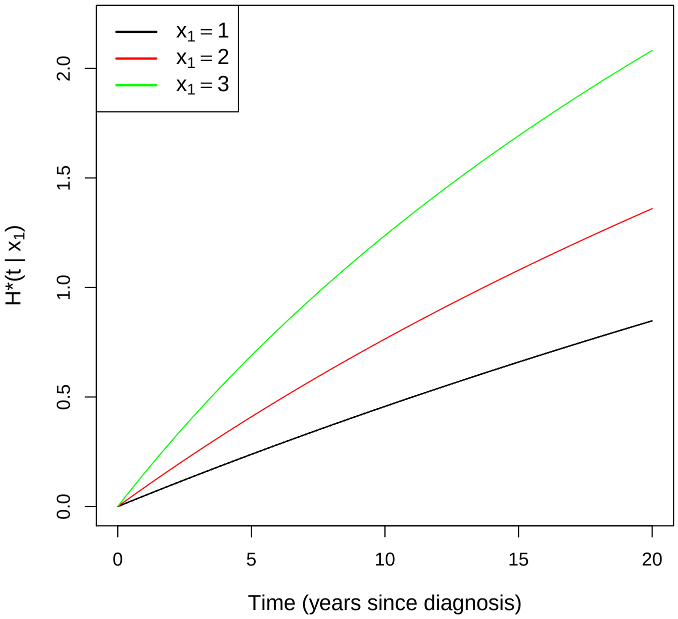

Figure 5.4.7: *Estimated cumulative hazard function for each x_1 (number of methylated genes) after marginalizing over a continuous mediator (Gleason score) in an exponential model, given by equation* (5.4.5)*.*

was fitted ignoring the mediator X_2, the main assumption would be that the log hazards have a constant vertical deviation over time. This is no longer the case after applying the result of Section 4.7 and the hazard function is appreciably different especially for small values of t. For a particular set of data if the hazards are proportional for the effect of X_1 conditional on X_2, they will not be exactly so for the marginal effect of X_1 and vice versa.

Figures 5.4.7 and 5.4.8 show the cumulative hazard and survival function, respectively, for each value of X_1. The new cumulative hazard function $H(t; x_1)$ (Figure 5.4.7) given by

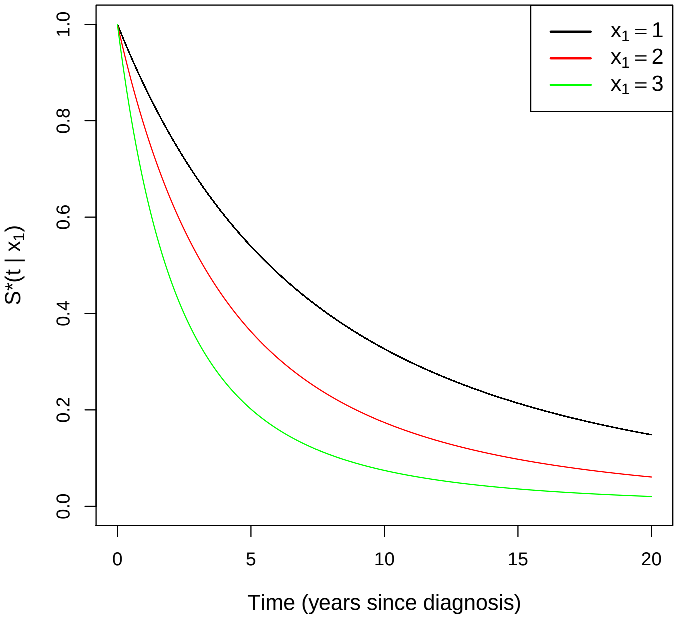

Figure 5.4.8: *Estimated survival function for each x_1 (number of methylated genes) after marginalizing over a continuous mediator (Gleason score) in an exponential model, corresponding to the hazard function given by equation* (5.4.4).

$$H(t;x_1) = \theta\left\{\log\left(\lambda t e^{\gamma_0\beta_2+(\beta_1+\beta_2\gamma_1)x_1} + e^{\psi(\theta)}\right) - \psi(\theta)\right\} \tag{5.4.5}$$

gives three curves instead of three straight lines with different slopes as we would obtain by fitting an exponential model adjusting only for x_1.

For small t the hazard function after marginalizing over the mediator will be approximately

$$\frac{\theta}{e^{\psi(\theta)}}\lambda e^{\beta^* x_1},$$

where $\beta^* = \beta_1 + \beta_2\gamma_1$. In this case the total effect β^* of the number of methylated genes on mortality was estimated to be 0.57 (HR: 1.78). The direct effect for small t is 0.41 (HR: 1.51) and the indirect effect, mediated by the Gleason score, is 0.17 (HR: 1.18). Thus about 29% of the effect of DNA methylation on prostate cancer mortality is mediated by the Gleason score.

Approximation of the lognormal by a Gamma distribution

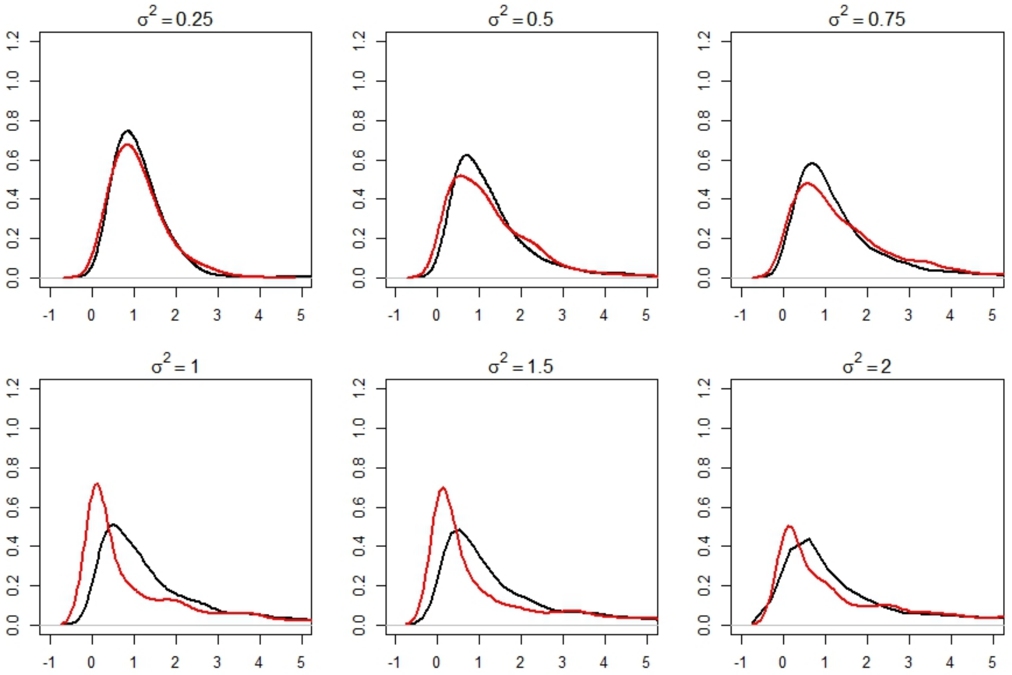

Figure 5.4.9: *Smoothed density of $W = e^{\beta U}$ (black line) and Gamma density (red line); σ^2 is the variance of W.*

Let $W = e^{\beta U}$, where $U \sim \mathcal{N}(0, \tau^2)$. Then $\log W \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \beta^2\tau^2$. Figure 5.4.9 shows the smoothed estimate of the density of W and the corresponding theoretical Gamma density for the following values of σ^2:

$0.25, 0.5, 0.75, 1, 1.5, 2$. For small values of σ^2 the two densities are very close. As σ^2 increases, the tail of the lognormal distribution becomes heavier and the approximation becomes less satisfactory.

Thus approximating the exponential of the product of the regression coefficient and the error term of the linear regression model for the mediator on the exposure by a Gamma distribution is expected to lead to satisfactory results over a reasonable range of values of the variance of that product.

5.4.3 Proportional hazards model with binary mediator

The intermediate variable, Gleason score, is next treated as a binary variable to illustrate the results of Section 4.7.3 for a proportional hazards model with a binary mediator. In this case we do not obtain a simple decomposition into a direct and an indirect effect. Figures 5.4.10 and 5.4.11 show the hazard and survival function, respectively, for each value of X_1. These were obtained from Equations (4.7.22) and (4.7.23), respectively.

As previously, we begin with an exponential model, which has a constant hazard function. After marginalizing over the binary mediator, the hazard function, shown in Figure 5.4.10 is no longer constant, but is decreasing with time.

Figure 5.4.12 shows the fitted survival functions as given by equation (4.7.20) (solid smooth curves) for each value of x_1 (1, 2 and 3). The dashed lines are the corresponding fitted survival curves from the model that ignores X_2

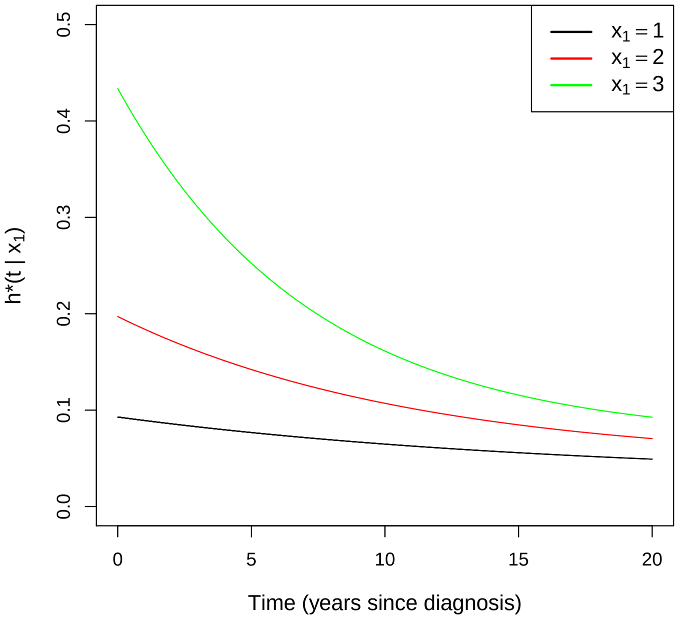

Figure 5.4.10: *Estimated hazard function after marginalizing over a binary mediator (Gleason score) in an exponential model, obtained from equation* (4.7.22) *for each value of* X_1*, the number of methylated genes.*

(Gleason score), that is, the model with hazard function $h(t; x_1) = \lambda^* e^{\beta_1^* x_1}$. The step functions are the Kaplan–Meier curves estimated separately within each stratum, after dividing the data according to the value of x_1.

As shown by Figure 5.4.12 the survival curves after marginalizing over the binary mediator agree with the Kaplan–Meier curves fitted to the data. For small values of time, the model in which the intermediate variable X_2 is ignored and the model obtained after integrating out X_2 have survival functions which are essentially the same. For larger values of time, the model which ignores X_2 gives smaller survival probabilities, but the data become sparse for larger values of t.

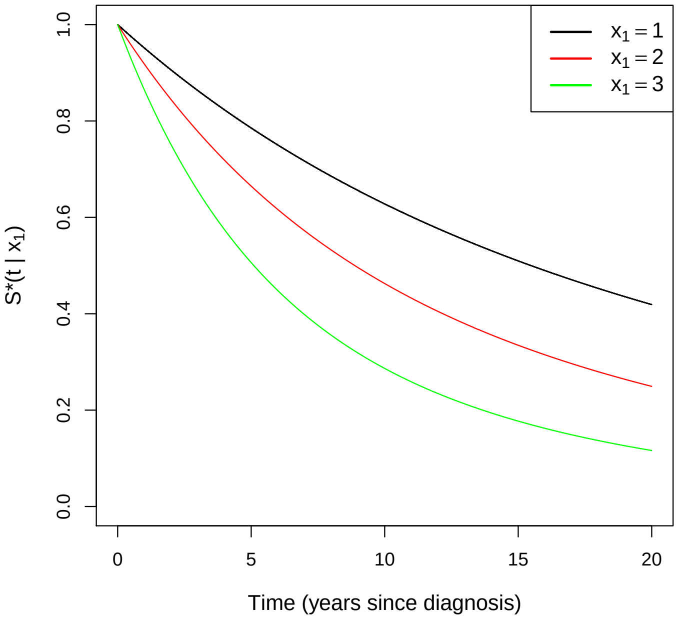

Figure 5.4.11: *Estimated survival function for each x_1 (number of methylated genes) after marginalizing over a binary mediator (Gleason score) in an exponential model, calculated using* (4.7.23)*.*

5.4.4 Discussion

In this Chapter the approximations derived in Chapter 4 for the effect of an explanatory variable after marginalizing over an intermediate variable were applied to the data and compared to the effect estimated by a model from which the intermediate variable is omitted. The cases in which a logistic regression and a proportional hazards model are fitted were illustrated. There is overall good agreement between the estimates calculated directly and those implied by the theory.

In particular the qualitative implications of the theoretical discussion for sur-

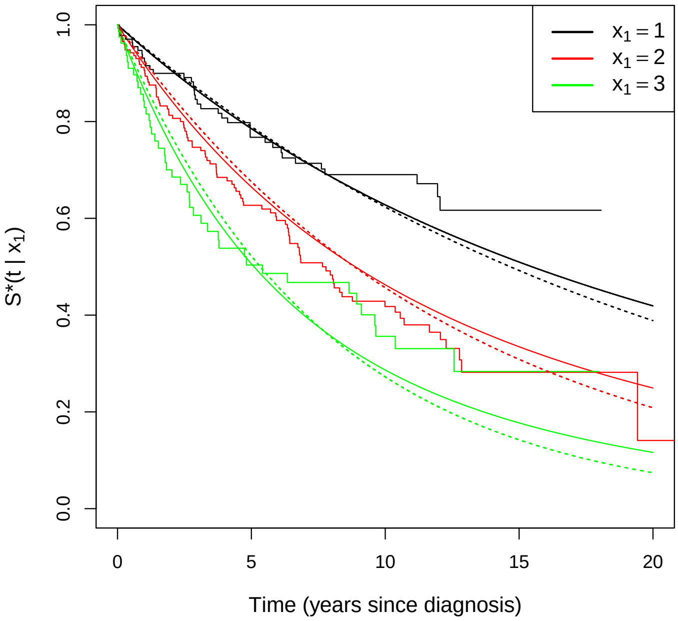

Figure 5.4.12: *Estimated survival function for each x_1 from equation* (4.7.20) *(solid smooth curves) with Kaplan–Meier curves (step functions) and estimated survival function from the model in which x_2 (Gleason score) is omitted (dashed lines).*

vival data are broadly confirmed. The broad agreement between the results of direct estimation and use of the general formulae suggests that the latter are useful for the general behaviour.

Chapter 6

Binary matched pairs

6.1 Introduction

Consider individuals which are paired, the pairing usually being such that the two individuals in any one pair tend to be similar. In each pair one individual is assigned at random to treatment 0, the other to treatment 1. On each individual a binary response is observed. Let n be the number of pairs (and $k = 2n$ the total number of binary responses). For the i^{th} pair, the observations are represented by random variables $Y_{i0}, Y_{i1},\ i = 1, \ldots, n$. Hence the possible observations on a pair, that for group 0 being written first, are: (0, 0), (0, 1), (1, 0), (1, 1). Let $R^{00}, R^{01}, R^{10}, R^{11}$ denote the numbers of pairs with the four types of response. Then $\sum R^{uv} = n$, the number of pairs.

The data can be summarized in a 2×2 contingency table:

	group 0	group 1
0		
1		
	n	n

The usual χ^2 significance test for such a table is invalid, because it ignores the correlation induced by pairing (McNemar, 1947). The significance of the difference between groups 0 and 1 should be tested using McNemar's test, that is, by rejecting the pairs (0, 0) and (1, 1), and by examining whether the proportion of (1, 0)'s among the discordant pairs, i.e. pairs with 'mixed' responses (0, 1) and (1, 0), is consistent with binomial variation with probability $\frac{1}{2}$ (Cox, 1958).

Cox (1958) introduced a logistic model that is analogous to the normal linear model commonly used for paired data with continuous responses. For the i^{th} pair, the logistic transforms for treatments 0 and 1 are respectively

$$\alpha_i \text{ and } \alpha_i + \theta \tag{6.1.1}$$

where α_i is a nuisance parameter characteristic of the i^{th} pair and θ is a treatment effect assumed constant on the logistic scale. That is, θ is the logistic difference between the two individuals in a pair considered conditionally on the aspects represented by the α_i. The number of model parameters increases with n and consequently estimates of θ obtained by maximum likelihood estimation are inconsistent (Neyman and Scott, 1948). Because of the large number of nuisance parameters a conditional likelihood approach is used.

The statistics associated with the nuisance parameters, and hence used for conditioning, are the pair totals $Y_{i0} + Y_{i1}$, $i = 1, \ldots, n$, and the statistic associated with the parameter θ is the total number of successes for group 1, $T = R^{01} + R^{11}$.

To examine the conditional distribution of $T = \sum_{i=1}^{n} Y_{i1}$ given $\{(Y_{i0}+Y_{i1}), i = 1, \ldots, n\}$, note that any pair for which $Y_{i0}+Y_{i1} = 0$ contributes zero to T and any pair for which $Y_{i0} + Y_{i1} = 2$ contributes one to T. Only discordant pairs, for which $Y_{i0}+Y_{i1} = 1$, contribute to T an amount to be regarded as random. Therefore the conditional distribution considered is that of the number R^{01} of pairs 01, given that $R^{01} + R^{10} = m$, the total number of discordant pairs.

Thus the conditional probability that the i^{th} pair contributes one to R^{01}, given that it is discordant, is given by

$$\mathbb{P}(Y_{i0} = 0, Y_{i1} = 1 \mid Y_{i0} + Y_{i1} = 1) = \frac{e^{\theta}}{1 + e^{\theta}}. \qquad (6.1.2)$$

This is the same for all pairs, thus the conditional distribution of R^{01} is Binomial$\left(m,\ \frac{e^{\theta}}{1+e^{\theta}}\right)$, with probability mass function

$$\mathbb{P}(R = r|M = m) = \binom{m}{r} \left(\frac{e^{\theta}}{1 + e^{\theta}}\right)^{r} \left(1 - \frac{e^{\theta}}{1 + e^{\theta}}\right)^{m-r}.$$

Under the null hypothesis that $\theta = 0$, the binomial parameter is equal to 0.5. Hence it is possible to test hypotheses about, or to obtain confidence limits for, θ, in the usual way for a binomial parameter.

A pair with response 00 might have a very large negative α_i disguising the presence of a treatment effect θ. If, however, some restriction is placed on the variation of the α_i's, some relevant information may be contained in the numbers of 00 and 11 pairs. The occurence of a large number of such pairs is, however, often evidence that θ is small and often also that some other major source of variability is present (Cox and Snell, 1989, p. 54).

If each pair is characterized by one or more explanatory variables, a possible approach is to replace α_i by a function, probably linear, of those explanatory variables and to fit the resulting model, which would in this case contain a much smaller number of unknown parameters. An alternative approach is to group the pairs into sets with the same or similar values of the explanatory variables and to use the analysis for the combination of several 2×2 tables. Both these analyses assume that the explanatory variables account for all or most of the correlation present (Cox and Snell, 1989, p. 54).

Alternatively it could be assumed that all individuals respond independently with probabilities of success θ_0 and θ_1, for treatments 0 and 1. Then only the total numbers of successes $R^{10} + R^{11}$ and $R^{01} + R^{11}$ in the two treatment groups are considered, but in this analysis the correlation between the two individuals in a pair is ignored.

Another approach given by Cox and Snell (1989) which does not require explanatory variables is as follows. Suppose each pair yields one of the 4 possible responses $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$ with probabilities

$$\theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}, \quad \Sigma\theta_{ij} = 1. \tag{6.1.3}$$

Here the probabilities of success under the two treatments are

$$\theta_{01} + \theta_{11} \ \text{ and } \ \theta_{10} + \theta_{11}.$$

The hypothesis that the success probabilities are equal, allowing for arbitrary dependency, is equivalent to the hypothesis that $\theta_{01} = \theta_{10}$. If θ_{11} and θ_{00} are arbitrary, this is equivalent to $\theta = 0$ in (6.1.2).

However the estimation procedure is different. Model (6.1.1) leads to estimation of θ via (6.1.2), which can be interpreted as the logit difference for an arbitrary individual conditional on the α_i for the pair, while under the multinomial model (6.1.3), some contrast of probabilities of success in the two treatment groups is estimated, for instance $\theta_{10} - \theta_{01}$ or, on a logistic scale,

$$\log\left(\frac{\theta_{10} + \theta_{11}}{\theta_{01} + \theta_{00}}\right) - \log\left(\frac{\theta_{01} + \theta_{11}}{\theta_{10} + \theta_{00}}\right) = \log\left(\frac{\theta_{1\cdot}}{\theta_{0\cdot}}\right) - \log\left(\frac{\theta_{\cdot 1}}{\theta_{\cdot 0}}\right).$$

Confidence intervals for these parameters are calculated either exactly, or via a profile log likelihood, or via the maximum likelihood estimates, for which large sample variances can be calculated. This approach involves all the data and not only the discordant pairs.

Neuhaus et al. (1994) considered a mixed-effects logistic model for binary matched pairs, that is, a model in which the α_is are treated as a sample from some essentially arbitrary distribution, and gave conditions for consistent estimation. They showed that under these conditions, the mixed-effects model estimator of the treatment effect is identical to the conditional likelihood estimator for matched pairs.

For most practical purposes the simple analysis based on the rather general model (6.1.1) is likely to be preferred. A limitation of the model is that there is no check from the data on its adequacy, unless further information is introduced, e.g. grouping of pairs into two types. As a test of significance of $\theta = 0$, the binomial test has the correct probability properties whenever there is no treatment effect.

A study design using dependent samples can help improve the precision of inferences for within-subject effects; the improvement is substantial when samples are highly correlated (Agresti, 2002, p. 412). Matched-pair designs also provide an effective method to control for potential confounding effects of covariates in studies of the effect of a binary risk factor.

If θ is small compared to the α_is, a pair with a large positive α_i has a high success probability for both individuals in the pair, while a pair with a large negative α_i has a high failure probability for both individuals in the pair. For any value of θ, the greater the variability in the α_is, the greater the overall positive association between responses.

The conditional analysis based on model (6.1.1) might lead to a large variance for the estimate of the treatment effect if the pairs have a large probability of being concordant, in which case most data will be discarded and the size of the sample used in the analysis will be a small proportion of the initial sample.

An alternative to the analysis based on (6.1.1) is to use an unconditional analysis, in which the probabilities of success of each group are averaged over the observations in the group, that is, the matching is ignored, and in which all observations are being used.

An important issue is the relation between logistic differences θ in different models. The issue is partly concerned with the informativeness of concordant pairs. There is a major difficulty in such comparisons in ensuring the comparability of parameters when comparisons are made between different essentially nonlinear models for the same data.

Gail (1988), Robinson and Jewell (1991), and Begg and Lagakos (1993) considered precision in logistic regression. Robinson and Jewell (1991) showed that adjusting for nonconfounding covariates, that is, covariates which are independent of the exposure given the response or independent of the response given the exposure, in logistic regression never results in a reduction in the variance of the estimate of the treatment effect. Variability increases except when the covariates are jointly independent of the outcome and treatment.

In Section 6.2 the conditional analysis for binary matched pairs corresponding to model (6.1.1) is described and in Section 6.3 the unconditional analysis in which the pairing is ignored is set out. Then in Section 6.4 the efficiency of the two types of analysis is compared. In Section 6.5 simulation results are presented and compared with the results obtained theoretically. In Section 6.6 conclusions are discussed.

6.2 Conditional analysis

Consider n binary matched pairs (Y_{i0}, Y_{i1}) such that for the i^{th} pair

$$\mathbb{P}(Y_{i0} = 1) = L_1(\alpha_i), \quad \mathbb{P}(Y_{i1} = 1) = L_1(\alpha_i + \theta), \tag{6.2.1}$$

where $L_1(\cdot) = L(\cdot)$ is the logistic function. Given a sample of binary observations on matched pairs, the conditional analysis disregards all concordant pairs and uses only the discordant ones for estimating the treatment effect θ. The probability that a pair is discordant is, treating α_i as a random variable A,

$$\begin{aligned} \pi_d &= \mathbb{P}(Y_{i0} = 0, Y_{i1} = 1) + \mathbb{P}(Y_{i0} = 1, Y_{i1} = 0) \\ &= \mathbb{E}_A \left\{ L_0(A) L_1(A + \theta) + L_1(A) L_0(A + \theta) \right\}, \end{aligned} \tag{6.2.2}$$

where $L_0(x) = 1 - L_1(x)$. Let ϕ be the conditional probability that a discordant pair is (0, 1). Then

$$\begin{aligned}\phi &= \frac{\mathbb{P}(Y_{i0}=0,Y_{i1}=1)}{\mathbb{P}(Y_{i0}=1,Y_{i1}=0)+\mathbb{P}(Y_{i0}=0,Y_{i1}=1)} \\ &= \frac{L_0(A)L_1(A+\theta)}{L_1(A+\theta)L_0(A)+L_0(A+\theta)L_1(A)} \\ &= \frac{e^{\theta}}{1+e^{\theta}}. \qquad (6.2.3)\end{aligned}$$

Thus the conditional probability that a discordant pair is (1, 0) is equal to $1-\phi$ and from (6.2.3) we have that

$$\theta = \log\frac{\phi}{1-\phi}. \qquad (6.2.4)$$

Let $\hat{\theta}_C$ denote the estimate of θ from the conditional analysis. This is

$$\hat{\theta}_C = \log\left(\frac{R^{01}}{R^{10}}\right).$$

We now calculate the theoretical asymptotic variance of this estimate. Let R be a random variable following a Binomial(m,ϕ) distribution and let

$$S = \log\frac{R}{m-R}.$$

Then

$$\frac{dS}{dR} = \frac{1}{R} + \frac{1}{m-R} = \frac{m}{R(m-R)} \simeq \frac{m}{m\phi(m-m\phi)} = \frac{1}{m\phi(1-\phi)}$$

and hence by the delta method

$$\mathrm{var}(S) \simeq \left(\frac{dS}{dR}\right)^2 \mathrm{var}(R) \simeq \frac{1}{m^2\phi^2(1-\phi)^2} m\phi(1-\phi) = \frac{1}{m\phi(1-\phi)}. \quad (6.2.5)$$

From (6.2.4) and (6.2.5), we have that

$$\mathrm{var}(\hat{\theta}_C) = \mathrm{var}(S) \simeq \frac{1}{m\phi(1-\phi)},$$

where m is the number of discordant pairs ($m \leq n$) and thus $m = n\pi_d$. Then

$$\mathrm{var}(\hat{\theta}_C) \simeq \frac{1}{n\pi_d\phi(1-\phi)}.$$

It follows that in the conditional analysis asymptotically

$$\mathrm{var}(\hat{\theta}_C) = \frac{1}{n\pi_d}\frac{(1+e^{\theta})^2}{e^{\theta}}. \quad (6.2.6)$$

To calculate theoretically the variance of $\hat{\theta}_C$ using (6.2.6), we thus need to calculate the value of π_d, the probability of a pair being discordant, given by (6.2.2). Let μ and σ^2 be the mean and variance, respectively, of the random variable A and let $\epsilon = A - \mu$. Then from (6.2.2), we have that

$$\pi_d = \mathbb{E}_\epsilon \left\{ L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta) + L_1(\mu+\epsilon)L_0(\mu+\epsilon+\theta) \right\}. \qquad (6.2.7)$$

We assume that ϵ is small, or equivalently that σ^2 is small. We then expand f_1 and f_2 in ϵ about 0, where $f_1(\epsilon) = L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)$ and $f_2(\epsilon) = L_1(\mu+\epsilon)L_0(\mu+\epsilon+\theta)$.

The Taylor expansion of $f_1(\epsilon)$ is

$$\begin{aligned} f_1(\epsilon) &= L_0(\mu)L_1(\mu+\theta) + \left\{L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)\right\}' \big|_{\epsilon=0}\epsilon \\ &\quad + \frac{1}{2}\left\{L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)\right\}'' \big|_{\epsilon=0}\epsilon^2 + \dots\ . \end{aligned}$$

We have that

$$L_1'(x) = L_1(x)L_0(x),$$

$$L_0'(x) = -L_1(x)L_0(x),$$

$$L_1''(x) = L_0(x)L_1(x)\left\{L_0(x) - L_1(x)\right\}$$

and

$$L_0''(x) = L_0(x)L_1(x)\left\{L_1(x) - L_0(x)\right\}.$$

Then

$$\frac{d}{d\epsilon}\left\{L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)\right\} = L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)\left\{L_0(\mu+\epsilon+\theta) - L_1(\mu+\epsilon)\right\}$$

and

$$\begin{aligned}\frac{d^2}{d\epsilon^2}\left\{L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)\right\} &= L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)\{L_0^2(\mu+\epsilon+\theta) \\ &\quad -L_0(\mu+\epsilon+\theta)L_1(\mu+\epsilon+\theta) - 2L_0(\mu+\epsilon+\theta)L_1(\mu+\epsilon) \\ &\quad +L_1^2(\mu+\epsilon) - L_0(\mu+\epsilon)L_1(\mu+\epsilon)\}.\end{aligned}$$

Evaluated at $\epsilon = 0$,

$$\begin{aligned}\frac{d^2}{d\epsilon^2}\left\{L_0(\mu+\epsilon)L_1(\mu+\epsilon+\theta)\right\}\Big|_{\epsilon=0} &= L_0(\mu)L_1(\mu+\theta)\{L_0^2(\mu+\theta) - L_0(\mu+\theta)L_1(\mu+\theta) \\ &\quad -2L_0(\mu+\theta)L_1(\mu) + L_1^2(\mu) - L_0(\mu)L_1(\mu)\}.\end{aligned}$$

Then

$$\begin{aligned} f_1(\epsilon) \;\simeq\;& L_0(\mu)L_1(\mu+\theta)+\epsilon f_1'(0)+\frac{1}{2}\epsilon^2 L_0(\mu)L_1(\mu+\theta)\{L_0^2(\mu+\theta) \\ & -L_0(\mu+\theta)L_1(\mu+\theta)-2L_0(\mu+\theta)L_1(\mu)+L_1^2(\mu)-L_0(\mu)L_1(\mu)\}. \end{aligned} \tag{6.2.8}$$

Similarly,

$$\begin{aligned} f_2(\epsilon) \;=\;& L_1(\mu)L_0(\mu+\theta)+\{L_1(\mu+\epsilon)L_0(\mu+\epsilon+\theta)\}'\big|_{\epsilon=0}\epsilon \\ & +\frac{1}{2}\{L_1(\mu+\epsilon)L_0(\mu+\epsilon+\theta)\}''\big|_{\epsilon=0}\epsilon^2+\dots\;. \end{aligned}$$

We have that

$$\frac{d}{d\epsilon}\{L_1(\mu+\epsilon)L_0(\mu+\epsilon+\theta)\}=L_1'(\mu+\epsilon)L_0(\mu+\epsilon+\theta)+L_1(\mu+\epsilon)L_0'(\mu+\epsilon+\theta)$$

and thus

$$\begin{aligned} \frac{d^2}{d\epsilon^2}\{L_1(\mu+\epsilon)L_0(\mu+\epsilon+\theta)\} \;=\;& L_1''(\mu+\epsilon)L_0(\mu+\epsilon+\theta)+L_1'(\mu+\epsilon)L_0'(\mu+\epsilon+\theta) \\ & +L_1'(\mu+\epsilon)L_0'(\mu+\epsilon+\theta)+L_1(\mu+\epsilon)L_0''(\mu+\epsilon+\theta) \end{aligned}$$

and

$$\begin{aligned}\frac{d^2}{d\epsilon^2}\left\{L_1(\mu+\epsilon)L_0(\mu+\epsilon+\theta)\right\}\Big|_{\epsilon=0} &= L_1(\mu)L_0(\mu+\theta)\{L_0^2(\mu)-L_0(\mu)L_1(\mu)\\ &\quad -2L_0(\mu)L_1(\mu+\theta)+L_1^2(\mu+\theta)\\ &\quad -L_0(\mu+\theta)L_1(\mu+\theta)\},\end{aligned}$$

thus

$$\begin{aligned}f_2(\epsilon) &\simeq L_1(\mu)L_0(\mu+\theta)+\epsilon f_2'(0)+\frac{1}{2}\epsilon^2L_1(\mu)L_0(\mu+\theta)\{L_0^2(\mu)+L_1^2(\mu+\theta)\\ &\quad -L_0(\mu)L_1(\mu)-2L_0(\mu)L_1(\mu+\theta)-L_0(\mu+\theta)L_1(\mu+\theta)\}. \quad (6.2.9)\end{aligned}$$

From (6.2.7), (6.2.8) and (6.2.9),

$$\begin{aligned}\pi_d &\simeq \mathbb{E}_\epsilon\Big\{L_0(\mu)L_1(\mu+\theta)+\epsilon f_1'(0)+\frac{1}{2}\epsilon^2L_0(\mu)L_1(\mu+\theta)\{L_0^2(\mu+\theta)+L_1^2(\mu)\\ &\quad -L_0(\mu+\theta)L_1(\mu+\theta)-2L_0(\mu+\theta)L_1(\mu)-L_0(\mu)L_1(\mu)\}\\ &\quad +L_1(\mu)L_0(\mu+\theta)+\epsilon f_2'(0)+\frac{1}{2}\epsilon^2L_1(\mu)L_0(\mu+\theta)\{L_0^2(\mu)+L_1^2(\mu+\theta)\\ &\quad -L_0(\mu)L_1(\mu)-2L_0(\mu)L_1(\mu+\theta)-L_0(\mu+\theta)L_1(\mu+\theta)\}\Big\}.\end{aligned}$$

Therefore the probability that a pair is discordant is

$$\begin{aligned}\pi_d \quad \simeq \quad & L_0(\mu)L_1(\mu+\theta) + L_1(\mu)L_0(\mu+\theta) \\ & +\frac{\sigma^2}{2}\Big\{L_0(\mu)L_1(\mu+\theta)\{L_0^2(\mu+\theta) + L_1^2(\mu) - L_0(\mu+\theta)L_1(\mu+\theta) \\ & \qquad\qquad -2L_0(\mu+\theta)L_1(\mu) - L_0(\mu)L_1(\mu)\} \\ & \quad +L_1(\mu)L_0(\mu+\theta)\{L_0^2(\mu) + L_1^2(\mu+\theta) - L_0(\mu)L_1(\mu) \\ & \qquad\qquad -2L_0(\mu)L_1(\mu+\theta) - L_0(\mu+\theta)L_1(\mu+\theta)\}\Big\}.\end{aligned} \tag{6.2.10}$$

Thus

$$\pi_d = A(\mu,\theta) + \frac{1}{2}\sigma^2 B(\mu,\theta), \tag{6.2.11}$$

say.

For a more symmetrical interpretation we now consider a formulation in which the pairs have parameters $(\nu - \frac{1}{2}\theta,\ \nu + \frac{1}{2}\theta)$, that is, $\nu = \mu + \frac{1}{2}\theta$. We consider values of ν between -2 and 2, θ from 0 to 2 and σ from 0 to 1.

Figure 6.2.1 shows π_d against ν and Figure 6.2.2 is a plot of π_d against ν and σ, with different colours showing different values of θ. The proportion of discordant pairs decreases as ν increases in absolute value and tends to be larger for larger values of θ.

Figures 6.2.3 and 6.2.4 are plots of $A(\nu,\theta)$ and $B(\nu,\theta)$, respectively, against ν, with different colours showing different values of θ. The range of the values

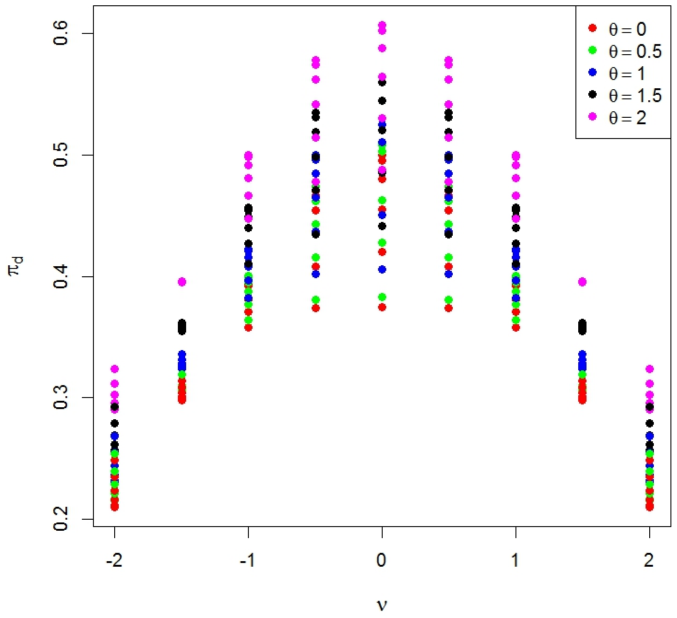

Figure 6.2.1: *Plot of the probability of a pair being discordant, π_d, against ν; colours represent different values of θ.*

that $B(\nu, \theta)$ takes is smaller than the range of $A(\nu, \theta)$. The quantity $A(\nu, \theta)$ decreases as the absolute value of ν increases and increases as θ increases, while $B(\nu, \theta)$ increases as ν increases in absolute value and decreases with increasing θ.

Figure 6.2.5 shows the differences $\pi_d - \pi_{d\text{avg}}$, where the average is taken at each value of ν, over the different values of σ, plotted against ν. This difference tends to be larger when ν is closer to zero and seems to be smallest when θ is 1 or 1.5.

A different approach, not explored in detail here, relates π_d to approximating bivariate normal integrals.

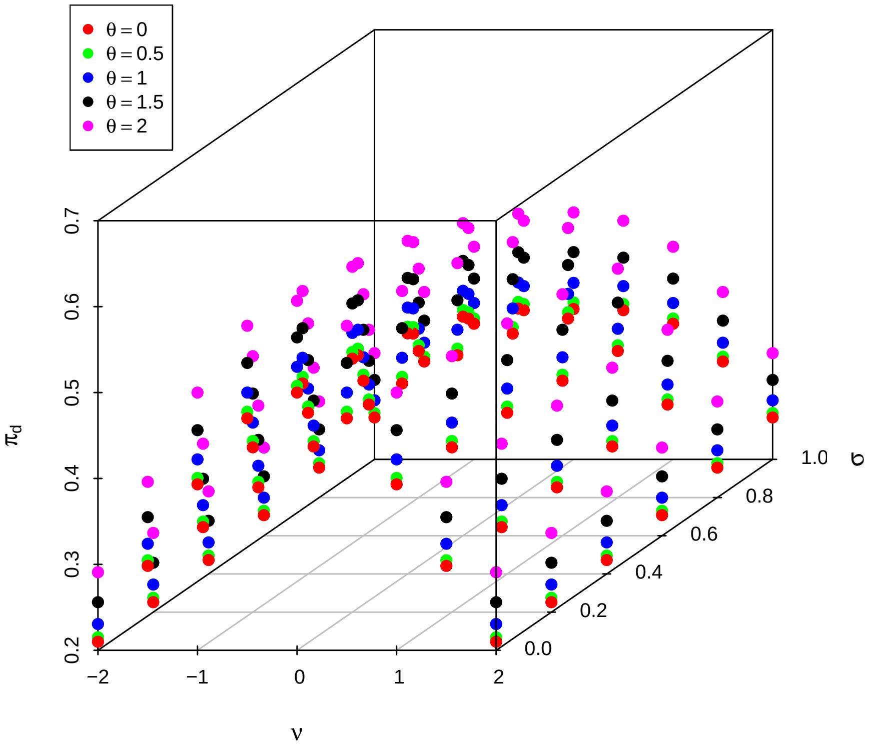

Figure 6.2.2: *Scatterplot of π_d against ν and σ; colours represent different values of θ.*

6.2.1 Simulation to check the calculation of π_d, the probability of a pair being discordant

To check the validity of the approximation to π_d, a simulation was performed. We generated $n = 1000$ binary matched pairs according to model (6.2.1) with A normally distributed and then calculated $\hat{\pi}_d$ as the proportion of discordant pairs. The following values were considered: $\theta = 0, 0.5, 1, 1.5, 2$, $\nu = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$ and $\sigma = 0, 0.2, 0.4, 0.6, 0.8, 1$, where

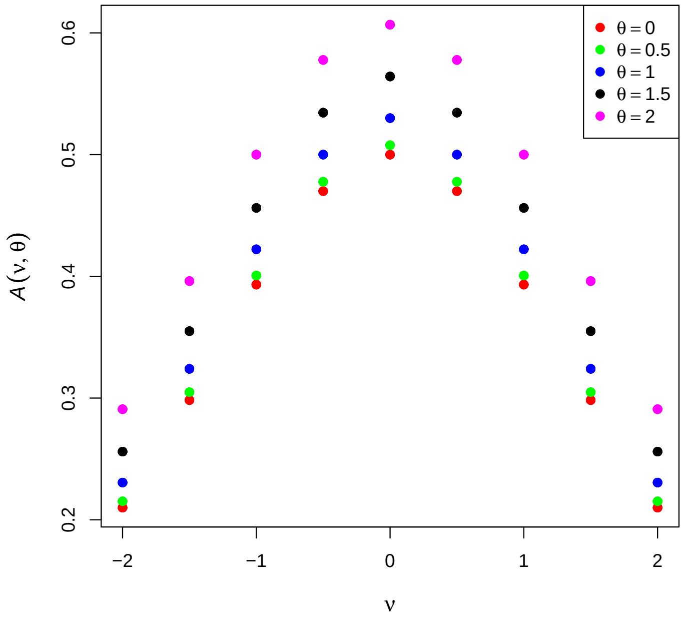

Figure 6.2.3: *Plot of $A(\nu, \theta)$ (from equation (6.2.11)) against ν; colours represent different values of θ.*

$\nu = \mu + \frac{1}{2}\theta$ and μ and σ are respectively the mean and standard deviation of A. The simulated proportions were very close to the probabilities obtained using the approximation based on the Taylor expansion and no systematic differences appeared between the two sets of values.

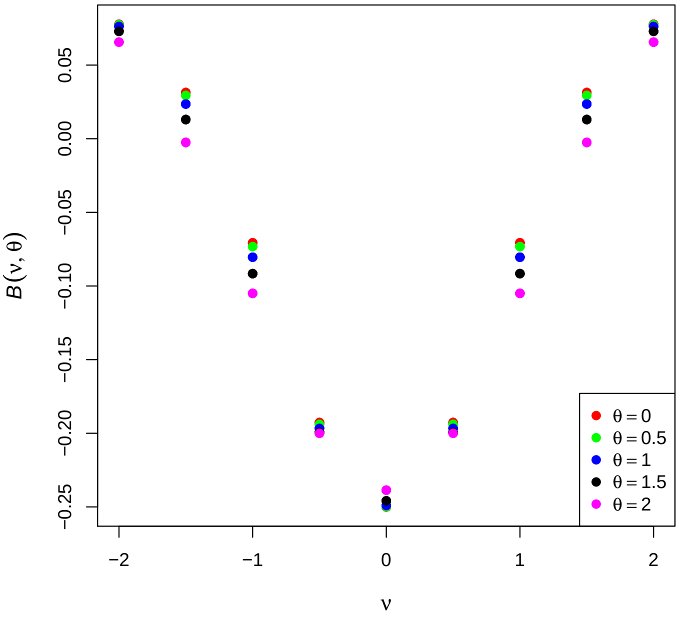

Figure 6.2.4: *Plot of $B(\nu, \theta)$ (from equation (6.2.11)) against ν; colours represent different values of θ.*

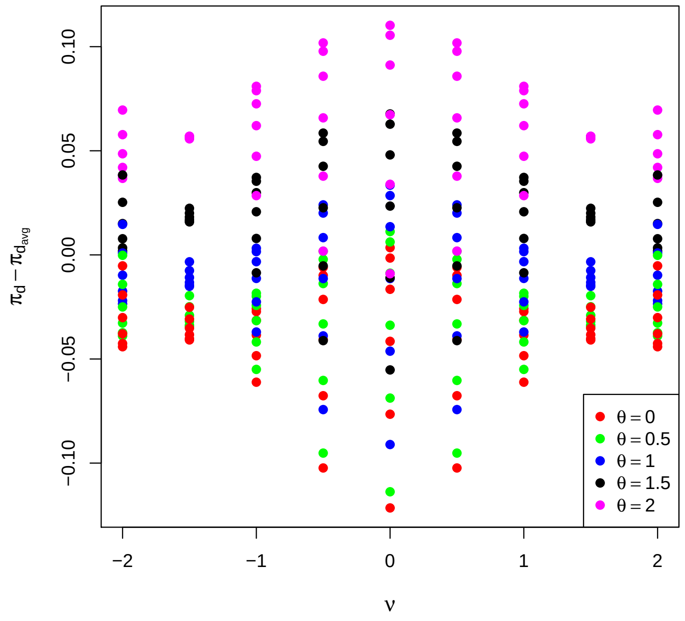

Figure 6.2.5: *Plot of $\pi_d - \pi_{d_{\text{avg}}}$, where the average is taken at each value of ν, against ν.*

6.3 Unconditional analysis

Suppose that the pairing is ignored, or equivalently that individuals are randomized to two groups, 0 (no treatment) and 1 (treatment), with probabilities of success

$$\mathbb{P}(Y_0 = 1) = \mathbb{E}\left\{L_1(A)\right\}, \quad \mathbb{P}(Y_1 = 1) = \mathbb{E}\left\{L_1(A+\theta)\right\}, \tag{6.3.1}$$

respectively. This unconditional analysis uses all pairs.

We have that

$$L_1(A+\theta) \simeq \Phi(kA + k\theta) = \mathbb{P}(Z < kA + k\theta),$$

where $Z \sim N(0,1)$. Thus, using the approximation described in Section 4.3,

$$\mathbb{E}\{L_1(A+\theta)\} \simeq \Phi\left(\frac{k\mu + k\theta}{\sqrt{1+k^2\sigma^2}}\right) \simeq L_1\left(\frac{\mu+\theta}{\sqrt{1+k^2\sigma^2}}\right).$$

Thus in the unconditional analysis, the probability of success for an individual in the 'control' group 0 is approximately

$$\phi_1 \simeq L_1\left(\frac{\mu}{\sqrt{1+k^2\sigma^2}}\right) \tag{6.3.2}$$

and that for an individual in the 'treatment' group 1 is

$$\phi_2 \simeq L_1\left(\frac{\mu+\theta}{\sqrt{1+k^2\sigma^2}}\right). \tag{6.3.3}$$

Therefore

$$\frac{\mu}{\sqrt{1+k^2\sigma^2}} = \text{logit}(\phi_1)$$

and

$$\frac{\mu+\theta}{\sqrt{1+k^2\sigma^2}} = \text{logit}(\phi_2).$$

To estimate the marginal log odds ratio, we calculate

$$\text{logit}(\phi_2) - \text{logit}(\phi_1),$$

which is related to the parameter θ via

$$\text{logit}(\phi_2) - \text{logit}(\phi_1) = \frac{\mu+\theta}{\sqrt{1+k^2\sigma^2}} - \frac{\mu}{\sqrt{1+k^2\sigma^2}}$$

and therefore we have

$$\frac{\theta}{\sqrt{1+k^2\sigma^2}} = \log\frac{\phi_2}{1-\phi_2} - \log\frac{\phi_1}{1-\phi_1}$$

and

$$\theta_U = \sqrt{1+k^2\sigma^2}\left\{\log\frac{\phi_2}{1-\phi_2} - \log\frac{\phi_1}{1-\phi_1}\right\}. \tag{6.3.4}$$

The test of significance of θ_U being different from zero is equivalent to testing whether the log odds contrast is different from zero.

Thus the variance of the estimate of the treatment effect θ in the unconditional analysis is, assuming σ^2 known,

$$\begin{aligned} \operatorname{var}(\hat{\theta}_U) &\simeq (1+k^2\sigma^2)\left\{\frac{1}{n\phi_1(1-\phi_1)} + \frac{1}{n\phi_2(1-\phi_2)}\right\} \\ &\simeq \frac{1+k^2\sigma^2}{n}\left\{\frac{1}{L_1\left(\frac{\mu}{\sqrt{1+k^2\sigma^2}}\right)L_0\left(\frac{\mu}{\sqrt{1+k^2\sigma^2}}\right)} + \frac{1}{L_1\left(\frac{\mu+\theta}{\sqrt{1+k^2\sigma^2}}\right)L_0\left(\frac{\mu+\theta}{\sqrt{1+k^2\sigma^2}}\right)}\right\}. \end{aligned} \tag{6.3.5}$$

The parameter σ^2 might possibly be estimated from the proportion of discordant pairs, although the resulting precision is likely to be low.

Figure 6.3.1 is a plot of $\operatorname{var}(\hat{\theta}_U)$ against ν. The variance of the estimate of θ from the unconditional analysis increases as the magnitude of ν increases. Figure 6.3.2 is a plot of $\operatorname{var}(\hat{\theta}_U)$ against σ, which shows that the relation of the variance of the estimate of the treatment effect from the unconditional analysis with σ is rather weak.

To aid interpretation of the plots, Table 6.3.1 shows the $100p\%$ point of the logistic function for probabilities p between 0.5 and 0.95. Suppose for instance that p varied with roughly 95% of values being between 0.6 and

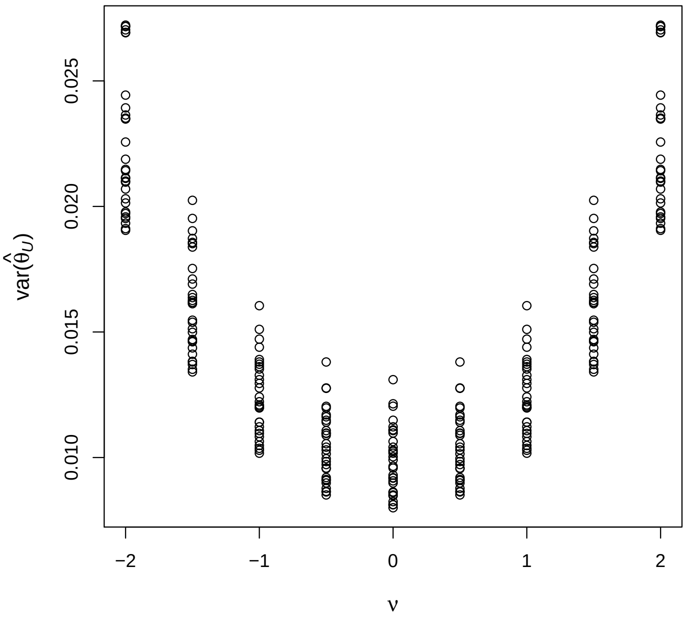

Figure 6.3.1: *Plot of* $\mathrm{var}(\hat{\theta}_U)$ *against* ν.

0.9 so that $L^{-1}(p)$ varies between 0.405 and 2.197 suggesting a σ of roughly 0.45. Then the factor $\sqrt{1 + (0.607^2 \cdot 0.45^2)}$ would be about 1.037, implying a relatively modest adjustment.

p	$L^{-1}(p)$
0.50	0.000
0.55	0.201
0.60	0.405
0.65	0.619
0.70	0.847
0.75	1.099
0.80	1.386
0.85	1.735
0.90	2.197
0.95	2.944

Table 6.3.1: *Probabilities* p *and values of* $L^{-1}(p)$.

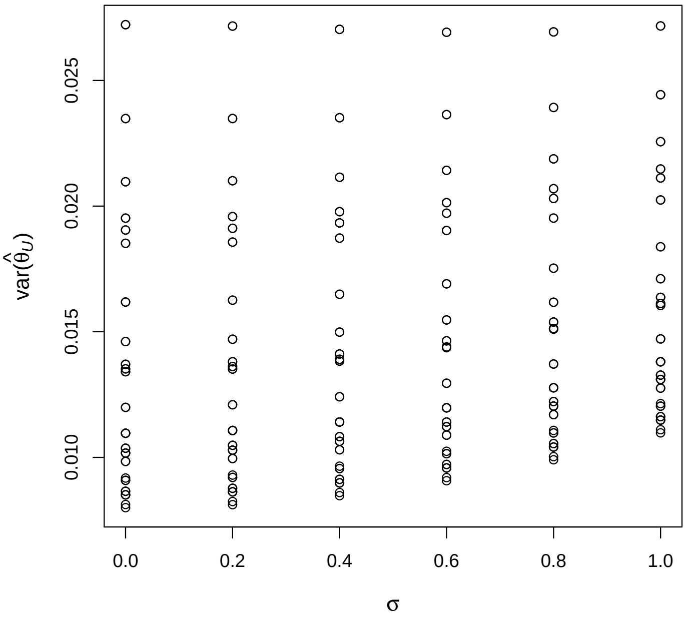

Figure 6.3.2: *Plot of* $\text{var}(\hat{\theta}_U)$ *against* σ*.*

6.4 Comparison of the efficiency of the conditional and unconditional analysis

The precisions of $\hat{\theta}_C$ and $\hat{\theta}_U$, the estimates yielded from the conditional and unconditional analysis, respectively, are compared. To investigate which of the two analyses is the more efficient and under which circumstances, we calculated the variance of the treatment effect θ using the two types of analysis, using (6.2.6) and (6.3.5) for the conditional and unconditional analysis, respectively. The parameter θ is defined in terms of the conditional formulation. Naive estimates of the log odds ratio are not directly comparable. The expressions obtained for the variance of θ were used to calculate variances for all combinations between the following values: $\theta = 0, 0.5, 1, 1.5, 2$, $\nu = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$, where $\nu = \mu + \frac{1}{2}\theta$, and $\sigma = 0, 0.2, 0.4, 0.6, 0.8, 1$. It should be noted that the value $\theta = 2$ corresponds to an extreme odds ratio, unlikely to be found in applications.

Figure 6.4.1 is a scatterplot of the ratio of the variance of the estimate of θ_C to the variance of θ_U against θ, ν and σ.

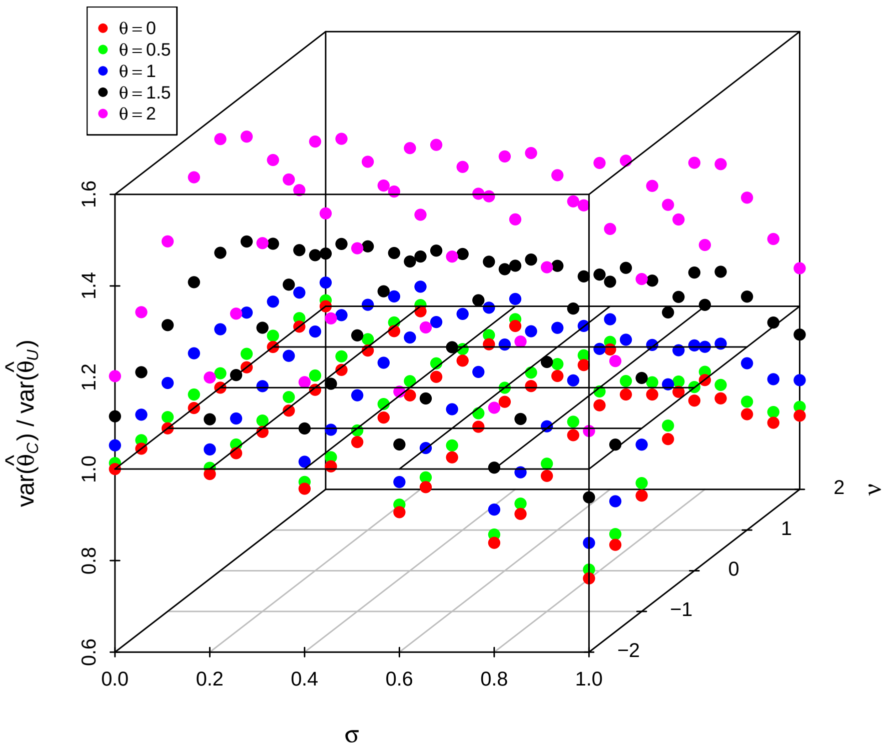

Figure 6.4.1: *Scatterplot of* $\mathrm{var}(\hat{\theta}_C)/\mathrm{var}(\hat{\theta}_U)$ *against* σ *and* ν*; colours represent different values of* θ*.*

The ratio $\mathrm{var}(\hat{\theta}_C)/\mathrm{var}(\hat{\theta}_U)$ is equal to one only in the trivial case of $\theta = \sigma =$

0. For 34.4% of the values considered here, $\mathrm{var}(\hat{\theta}_C)/\mathrm{var}(\hat{\theta}_U)$ was less than one, that is, the conditional analysis seems to be more efficient than the unconditional analysis. For 63.3% of the values considered, $\mathrm{var}(\hat{\theta}_C)/\mathrm{var}(\hat{\theta}_U)$ was greater than one, that is, for those values, the unconditional analysis seems to be more efficient than the conditional analysis.

When $\theta = 0$, $\mathrm{var}(\hat{\theta}_C)/\mathrm{var}(\hat{\theta}_U)$ is always less than or equal to one, that is, the conditional analysis yields a more precise estimate than the unconditional near $\theta = 0$. In this special case where the true treatment effect θ is zero, when $\sigma = 0$, the ratio is equal to one and as σ increases the ratio decreases, indicating that the conditional analysis becomes more precise compared to the unconditional. Also as ν increases in absolute value, the conditional analysis becomes more precise, although π_d decreases with increasing ν.

As θ increases, the ratio becomes larger, especially when σ and ν (or equivalently μ) are small. When $\theta = 2$, the unconditional analysis is almost always more efficient than the conditional. The lowest value of the ratio $\mathrm{var}(\hat{\theta}_C)/\mathrm{var}(\hat{\theta}_U)$ for $\theta = 2$ out of the cases considered here is obtained when $\nu = \pm 2$ and $\sigma = 1$, in which case it is very close to 1.

Figures 6.4.2, 6.4.3 and 6.4.4 show scatterplots of the ratio of the variance of the estimate of θ_C to the variance of the estimate of θ_U against θ, ν and σ, respectively.

Linear regression with $\mathrm{var}(\hat{\theta}_C)/\mathrm{var}(\hat{\theta}_U)$ as the response and θ, ν and σ as explanatory variables was used to investigate the relationship between the ratio of the variances and the three parameters. Figures 6.4.2, 6.4.3 and 6.4.4

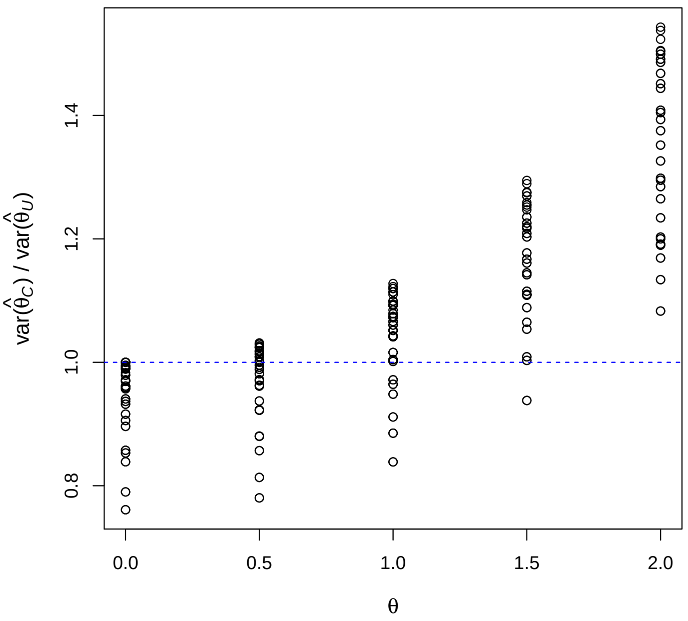

Figure 6.4.2: *Ratio of variances plotted against θ.*

suggest that the relationships of the ratio of the variances with θ, ν and σ are not linear, thus higher order terms were considered. Adding higher order terms one at a time and deciding whether to include them according to the t-tests, even though the variation involved is numerical rather than statistical, shows that θ^2, ν^2 and σ^2 should also be included. Table 6.4.1 shows the results from the fitted model. The estimated coefficients show that the ratio of the two variances is expected to increase as θ^2 increases, to decrease as ν^2 increases and to decrease as σ^2 increases.

The value of π_d, as calculated using the Taylor approximation (Equation (6.2.10)), varies between 0.21 and 0.61 for the values used here. There is no apparent pattern in the relationship between σ and π_d. Figure 6.2.1 shows

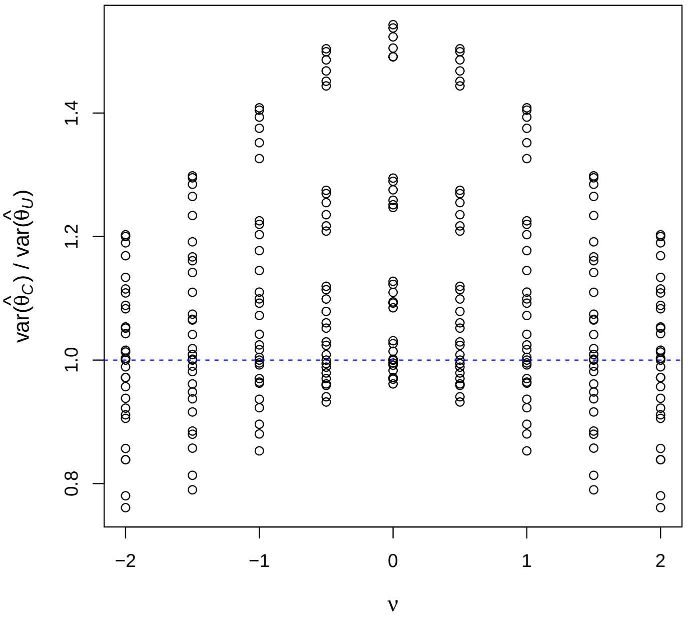

Figure 6.4.3: *Ratio of variances plotted against ν.*

	Estimate	Formal st. error
Intercept	1.06	0.009
θ	-0.009	0.014
θ^2	0.105	0.007
ν	1.1×10^{-17}	0.002
ν^2	-0.041	0.002
σ	-0.032	0.030
σ^2	-0.100	0.029

Table 6.4.1: *Results from linear regression with the ratio of the two variances as the response.*

how the value of π_d varies with ν and θ.

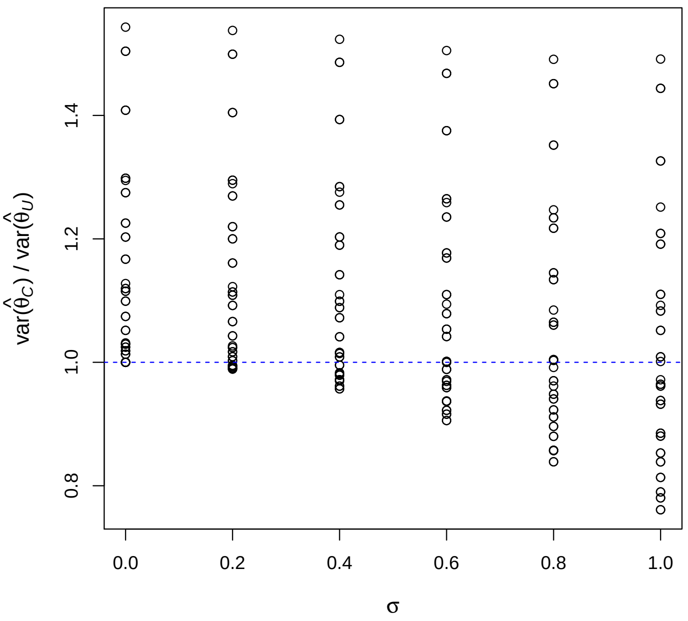

Figure 6.4.4: *Ratio of variances plotted against σ.*

6.4.1 Significance testing of the hypothesis of no treatment effect

We consider the case where the hypothesis of no difference between the two groups is to be tested. In comparing alternative formulations, such as conditional and unconditional analyses, it is important that the parameters estimated have directly comparable interpretations in the two formulations. In the simpler situation of testing the null hypothesis it is enough that the null hypotheses are the same and that evaluation of local behaviour near the null hypothesis also corresponds.

For the conditional analysis, described in Section 6.2, we take the test statistic to be $T_C = \log\{n_{01}/n_{10}\}$ and in the discussion to follow of the unconditional analysis we take $T_U = \log\{n_{.1}n_{0.}/n_{1.}n_{.0}\} = \log\{\hat{\phi}_2(1-\hat{\phi}_1)/(\hat{\phi}_1(1-\hat{\phi}_2))\}$.

T_C, interpreted as the logit difference between the two individuals in an arbitrary pair, has expected value $\mathbb{E}(T_C) = \theta$. At the null hypothesis $\theta = 0$ we have that, asymptotically,

$$\mathrm{var}(T_C) = \frac{1}{\frac{1}{2}n_d} + \frac{1}{\frac{1}{2}n_d} = \frac{4}{n_d} = \frac{4}{n\pi_d},$$

where n is the number of pairs, n_d the number of discordant pairs and π_d the probability of a pair being discordant. The Pitman efficacy (Cox and Hinkley, 1974, p. 337–338) for testing the hypothesis that $\theta = 0$ is

$$\mathcal{E}_C = \frac{\left\{\partial\mathbb{E}(T_C)/\partial\theta\big|_{\theta=0}\right\}^2}{n\mathrm{var}(T_C)\big|_{\theta=0}} = \frac{\pi_d}{4}. \qquad (6.4.1)$$

Under the null hypothesis, from (6.2.10), the probability of a pair being discordant is

$$\pi_d \simeq 2L_0(\mu)L_1(\mu)\left\{1 + \frac{1}{2}\sigma^2\left(1 - 6L_0(\mu)L_1(\mu)\right)\right\},$$

thus

$$\mathcal{E}_C \simeq \frac{1}{2}L_0(\mu)L_1(\mu)\left\{1 + \frac{1}{2}\sigma^2\left(1 - 6L_0(\mu)L_1(\mu)\right)\right\}. \qquad (6.4.2)$$

In the unmatched analysis of Section 6.3, we have that $\mathbb{P}(Y_0 = 1) = \mathbb{E}\left\{L_1(A)\right\}$

and $\mathbb{P}(Y_1 = 1) = \mathbb{E}\left\{L_1(A+\theta)\right\}$. These probabilities can be calculated approximately by using a Taylor expansion or by approximating $L_1(\cdot)$ by $\Phi(\cdot)$. For group 0 this is

$$\mathbb{P}(Y_0 = 1) \simeq L_1(\mu) + \frac{1}{2}\sigma^2 L_1(\mu) L_0(\mu)\left\{L_0(\mu) - L_1(\mu)\right\}$$

and

$$\mathbb{P}(Y_0 = 0) \simeq L_0(\mu) + \frac{1}{2}\sigma^2 L_1(\mu) L_0(\mu)\left\{L_1(\mu) - L_0(\mu)\right\}.$$

For group 1 analogous expressions hold, with μ is replaced by $\mu + \theta$. Using a further approximation,

$$\operatorname{logit}\left\{\mathbb{P}(Y_0 = 1)\right\} \simeq \mu + \frac{1}{2}\sigma^2\left\{L_0(\mu) - L_1(\mu)\right\},$$

thus T_U, the log odds contrast in the unconditional analysis, has asymptotic expected value

$$\begin{aligned}\mathbb{E}(T_U) &= \operatorname{logit}\left\{\mathbb{P}(Y_1 = 1)\right\} - \operatorname{logit}\left\{\mathbb{P}(Y_0 = 1)\right\} \\ &= \theta + \frac{1}{2}\sigma^2\left\{L_0(\mu+\theta) - L_1(\mu+\theta) - L_0(\mu) + L_1(\mu)\right\}.\end{aligned}$$

Then

$$\frac{\partial \mathbb{E}(T_U)}{\partial \theta} \simeq 1 - \sigma^2 L_0(\mu+\theta) L_1(\mu+\theta)$$

which under the null hypothesis is $1 - \sigma^2 L_0(\mu)L_1(\mu)$. The variance under the null hypothesis is that of the comparison of two independent logits, each based on n observations and thus is

$$\begin{aligned}
\operatorname{var}(T_U) &= \frac{2}{n} \frac{1}{\mathbb{P}(Y_0 = 0)\mathbb{P}(Y_0 = 1)} \\
&= \frac{2}{n} \frac{1}{L_0(\mu)\left\{1 + \frac{1}{2}\sigma^2 \frac{L_0''(\mu)}{L_0(\mu)}\right\} L_1(\mu)\left\{1 + \frac{1}{2}\sigma^2 \frac{L_1''(\mu)}{L_1(\mu)}\right\}} \\
&\simeq \frac{2}{nL_0(\mu)L_1(\mu)} \left\{1 - \frac{1}{2}\sigma^2 \left(L_0(\mu) - L_1(\mu)\right)^2\right\}, \qquad (6.4.3)
\end{aligned}$$

assuming σ^4 is negligible. Therefore the Pitman efficacy for T_U is

$$\begin{aligned}
\mathcal{E}_U &= \frac{\left\{1 - \sigma^2 L_0(\mu)L_1(\mu)\right\}^2 L_0(\mu)L_1(\mu)}{2\left\{1 - \frac{1}{2}\sigma^2 \left(L_0(\mu) - L_1(\mu)\right)^2\right\}} \\
&\simeq \frac{L_0(\mu)L_1(\mu)}{2} \left\{1 - 2\sigma^2 L_0(\mu)L_1(\mu) + \frac{1}{2}\sigma^2 \left(L_0(\mu) - L_1(\mu)\right)^2\right\} \\
&= \frac{L_0(\mu)L_1(\mu)}{2} \left\{1 + \sigma^2 \left(\frac{1}{2} - 4L_0(\mu)L_1(\mu)\right)\right\},
\end{aligned}$$

thus

$$\mathcal{E}_U \simeq \frac{L_0(\mu)L_1(\mu)}{2} \left\{1 + \frac{1}{2}\sigma^2 \left(1 - 8L_0(\mu)L_1(\mu)\right)\right\}, \qquad (6.4.4)$$

using $\{L_0(x) - L_1(x)\}^2 = 1 - 4L_0(x)L_1(x)$, $\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n$, $|x| < 1$ and ignoring terms of order σ^4 and above.

Therefore to assess the relative efficiency for $\theta = 0$, we compare (6.4.2) and (6.4.4). Since in this special case $\mathcal{E}_U$ is smaller than $\mathcal{E}_C$, near the null hypothesis of zero treatment effect the matched design tends to be slightly more efficient than the unmatched one, as is confirmed by the previous plots (Figures 6.4.1 and 6.4.2).

Often

$$L_0(\mu)L_1(\mu) \simeq \frac{1}{4}$$

and then

$$\mathcal{E}_C \simeq \frac{1}{8}\left(1 - \frac{1}{4}\sigma^2\right) \tag{6.4.5}$$

and for comparison

$$\mathcal{E}_U \simeq \frac{1}{8}\left(1 - \frac{1}{2}\sigma^2\right). \tag{6.4.6}$$

Thus for testing the hypothesis of no treatment effect the conditional analysis is slightly better than the unconditional analysis, depending on the amount of variability between pairs. However, this is an asymptotic calculation.

6.5 Simulation

As a check on the expressions for the estimates of the treatment effect and their variances obtained by the two different methods of analysis, a simulation was performed and the values calculated theoretically were compared to those obtained from the simulation.

To obtain simulated values of $\hat{\theta}_C$, $\hat{\theta}_U$ and their variances, $r = 1000$ simulations were performed as follows. At each simulation run, a sample of $n = 1000$ matched pairs was generated (that is, $k = 2n$ binary values in total), such that the first value (corresponding to group 0) is equal to 1 with probability $L_1(A)$ and the second one is 1 with probability $L_1(A + \theta)$, where A is drawn from a normal distribution with mean μ and variance σ^2. All combinations of the following values were considered: $\theta = 0, 0.5, 1, 1.5, 2$, $\nu = -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2$, where $\nu = \mu + \frac{1}{2}\theta$, and $\sigma = 0, 0.2, 0.4, 0.6, 0.8, 1$.

6.5.1 Simulation for conditional analysis

To obtain estimates of θ_C, the parameter representing the treatment effect in the conditional analysis, in each simulation run $\hat{\phi}$ was calculated as the number of discordant pairs for which the first value is zero divided by the number m of all discordant pairs. We then calculate $\hat{\theta}_C = \log\{\hat{\phi}/(1 - \hat{\phi})\}$. Thus $r = 1000$ values of $\hat{\theta}_C$ were obtained, using which the empirical variance was calculated. The mean of the r values of $\hat{\theta}_C$ for each set of parameter values was compared to the value found theoretically to assess bias.

6.5.2 Simulation for unconditional analysis

To obtain simulated values of $\hat{\theta}_U$ and its variance, in each simulation run $\hat{\phi}_1$ was calculated as the number of 'successes' in group 0 divided by n, and $\hat{\phi}_2$ as the number of 'successes' in group 1 divided by n. We then calculate $\hat{\theta}_U = \sqrt{1+k^2\sigma^2}\left\{\log\{\hat{\phi}_2/(1-\hat{\phi}_2)\} - \log\{\hat{\phi}_1/(1-\hat{\phi}_1)\}\right\}$. Thus $r = 1000$ values of $\hat{\theta}_U$ were obtained from which the variance was calculated. The mean of the r estimates of θ_U for each set of parameter values was compared to the value found theoretically to assess bias.

6.5.3 Comparison of theoretical results and results from simulation

The simulation results suggest that the unconditional analysis is in most cases more efficient than the conditional analysis. The ratio of the simulated variance of the conditional analysis to the simulated variance of the unconditional analysis was less than one for 9.26% of the sets of values considered, equal to one for 2.96% and greater than one for 87.8%. Figure 6.5.1 shows the values of the ratio of $\text{var}(\hat{\theta}_C)$, the simulated variance from the conditional analysis, to $\text{var}(\hat{\theta}_U)$, the simulated variance from the unconditional analysis, plotted against θ.

As a check on the accuracy of the approximations to the variance of $\hat{\theta}$ obtained from the conditional and unconditional analyses, as given by equations (6.2.6) and (6.3.5), respectively, the calculated values were compared to the corresponding values obtained by simulation. Figure 6.5.2 shows the ratio of

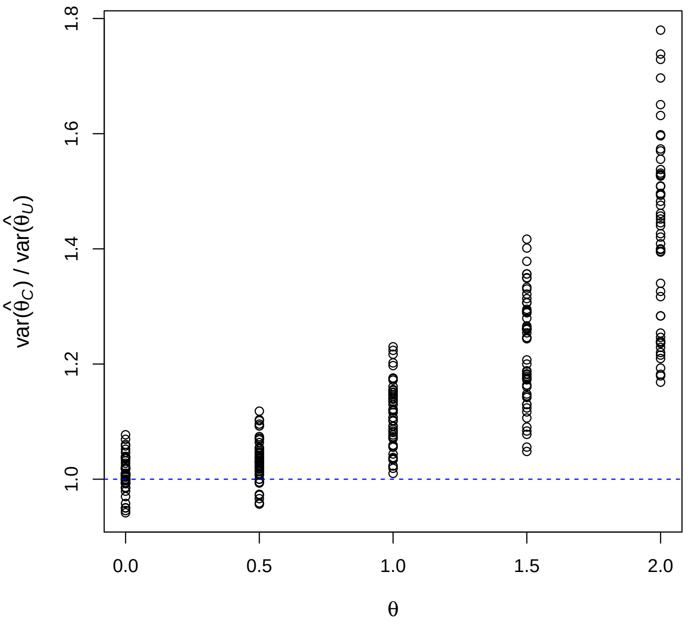

Figure 6.5.1: *Ratio of simulated variance of $\hat\theta_C$ over simulated variance of $\hat\theta_U$ plotted against θ. The dashed line is where the ratio is equal to one.*

the theoretical variance over the simulated variance for the conditional and unconditional analysis, respectively, plotted against σ. For the conditional analysis, the values of the ratio are scattered about one, suggesting that the theoretical approximation to the variance from the conditional analysis is satisfactory. The corresponding plot for the unconditional analysis shows that for smaller values of σ the theoretical variances agree with the simulated, while as σ gets closer to one theoretical approximation tends to yield larger values for the variance than those obtained from the simulation. This is expected given that the expression obtained for $\mathrm{var}(\hat\theta_U)$ relies on an approximation based on the assumption that σ is small. Also, the estimate of θ from the unconditional analysis, $\hat\theta_U$, is slightly biased, especially for large values of μ, σ and θ.

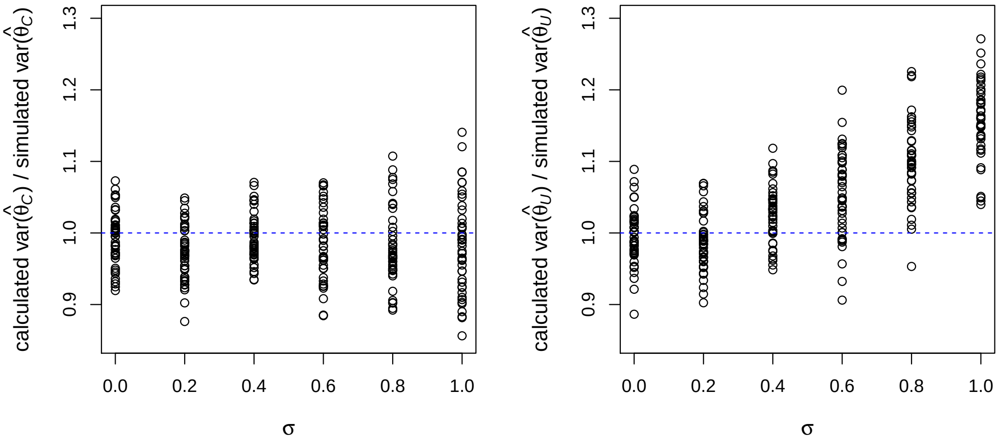

Figure 6.5.2: *Scatterplots of the ratio of the theoretical over the simulated variances for the conditional (left) and unconditional (right) analysis, plotted against σ.*

6.6 Discussion

There are two broad considerations involved in interpreting these results. First, the parameter θ describing the contrast of log odds between treatment and controls in the conditional analysis is defined conditionally on the features M implied by the matching variables. In fact

$$\theta_C = \log \frac{\phi}{1-\phi}, \tag{6.6.1}$$

where $\phi = \mathbb{P}\{(0,1) \mid \text{discordant}\}$. By contrast in the unconditional analysis the log odds contrast is

$$\log \frac{\phi_2}{1-\phi_2} - \log \frac{\phi_1}{1-\phi_1} = \log \frac{\phi_2(1-\phi_1)}{\phi_1(1-\phi_2)}, \tag{6.6.2}$$

where $\phi_1 = \mathbb{P}\{Y_0 = 1\}$ and $\phi_2 = \mathbb{P}\{Y_1 = 1\}$. This is related to θ by

$$\theta = \sqrt{1 + k^2\sigma^2} \left\{ \log \frac{\phi_2}{1-\phi_2} - \log \frac{\phi_1}{1-\phi_1} \right\}. \tag{6.6.3}$$

Equations (6.6.2) and (6.6.3) show that these two parameters are the same if and only if either Y is independent of treatment and control, in which case both parameters are zero, or if the matching is ineffective, $\sigma^2 = 0$. Except in these cases θ is further from zero than the log odds contrast from the unconditional analysis.

There are two implications. First estimates of the contrast of log odds from conditional and unconditional analyses of the same data are not estimating the same parameter and are likely to be different, perhaps seriously so. Secondly comparison of the conclusions from two different studies, one matched and one unmatched, requires care.

The adjustment for the unconditional log odds contrast is a new result which relates the unconditional log odds contrast to the parameter of the conditional analysis. For small σ^2 the two quantities do not differ much, but for large values of σ^2 the difference is substantial. For example, for σ^2 close to 2, the adjusted log odds contrast is about 1.5 times the unadjusted one.

The unconditional analysis seems to be more efficient than the conditional analysis in many cases, in particular when the treatment effect is large. When the treatment effect is close to zero, the conditional analysis is more efficient. Also for testing the significance of the treatment effect the conditional analysis is slightly more efficient. However, matching plus randomization controls for unobserved confounders, a different aspect from variance comparison.

References

Aalen, O. O., Borgan, Ø. and Gjessing, H. K. (2008). *Survival and Event History Analysis: a Process Point of View.* New York: Springer.

Aalen, O. O., Røysland, K., Gran, J. M. and Ledergerber, B. (2012). Causality, mediation and time: a dynamic viewpoint. *Journal of the Royal Statistical Society* A, **175** (4), 831–861.

Abramowitz, M. and Stegun, I. A. (Eds.) (1964). *Handbook of Mathematical Functions: with formulas, graphs, and mathematical tables* (Vol. 55). `DoverPublications.com`.

Agresti, A. (2002). *Categorical Data Analysis*, Second Edition. Hoboken: John Wiley and Sons Ltd.

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, **51** (6), 1173–1182.

Beaumont, M. A. (2010). Approximate Bayesian computation in evolution and ecology. *Annual Review of Ecology, Evolution and Systematics*, **41**, 379–406.

Beaumont, M. A., Zhang, W. and Balding, D.J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Begg, M. D. and Lagakos, S. W. (1993). Loss in efficiency caused by omitting covariates and misspecifying exposure in logistic regression models. *Journal of the American Statistical Association*, **88**, 166–170.

Blum, M. G. B., Nunes, M. A., Prangle, D. and Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, **28** (2), 189–208.

Breen, R., Karlson, K. B. and Holm, A. (2013). Total, direct, and indirect effects in logit and probit models. *Sociological Methods and Research*, **42** (2), 164–191.

Bretagnolle, J. and Huber-Carol, C. (1988). Effects of omitting covariates in Cox's model for survival data. *Scandinavian Journal of Statistics*, **15** (2), 125–138.

Buzbas, E. O. and Rosenberg, N. A. (2013). AABC: approximate approximate Bayesian computation when simulating a large number of data sets is computationally infeasible. *arXiv:1301.6282*.

Cochran, W. G. (1938). The omission or addition of an independent variable in multiple linear regression. *Supplement to the Journal of the Royal Statistical Society*, **5** (2), 171–176.

Cox, D. R. (1958). Two further applications of a model for binary regression. *Biometrika*, **45**, 562–565.

Cox, D. R. (2007). On a generalization of a result of W. G. Cochran. *Biometrika*, **94** (3), 755–759.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, London: Chapman and Hall / CRC.

Cox, D. R. and Kartsonaki, C. (2012). The fitting of complex parametric models. *Biometrika*, **99** (3), 741–747.

Cox, D. R. and Lewis, P. A. W. (1966). *The Statistical Analysis of Series of Events.* London: Chapman and Hall.

Cox, D. R. and Reid, N. (2000). *The Theory of the Design of Experiments.* Boca Raton: Chapman and Hall / CRC.

Cox, D. R. and Smith, W. L. (1954). On the superposition of renewal processes. *Biometrika*, **41**, 91–99.

Cox, D. R. and Snell, E. J. (1989). *The Analysis of Binary Data*, Second Edition. London: Chapman and Hall.

Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. and François, O. (2010). Approximate Bayesian computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–418.

Daniel, R., De Stavola, B. L. and Leon, D. (2013). Causal mediation analysis with multiple causally-ordered mediators. *Symposium on Causal Mediation Analysis*, 28–29 January 2013, Ghent, Belgium.

Dean, T. A., Singh, S. S., Jasra, A. and Peters, G. W. (2011). Parameter estimation for hidden Markov models with intractable likelihoods. *arXiv:1103.5399*.

Diggle, P. J. and Gratton, R. J. (1984). Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society* B, **46** (2), 193–227.

Đoković, D. Ž. (2010). Hadamard matrices of small order and Yang conjecture. *Journal of Combinatorial Designs*, **18** (4), 254–259.

Đoković, D. Ž., Golubitsky, O. and Kotsireas, I. S. (2013). Some new orders of Hadamard and skew-Hadamard matrices. *Journal of Combinatorial Designs*, doi: 10.1002/jcd.21358.

Drovandi, C. C., Pettitt, A. N. and Faddy, M. J. (2011). Approximate Bayesian computation using indirect inference. *Journal of the Royal Statistical Society* C, **60** (3), 317–337.

Fearnhead, P. and Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society* B, **74** (3), 419–474.

Gail, M. H. (1988). The effect of pooling across strata in perfectly balanced studies. *Biometrics*, **44** (1), 151–162.

Gail, M. H., Wieand, S. and Piantadosi, S. (1984). Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, **71** (3), 431–444.

Gillio-Tos, A., Fiano, V., Zugna, D., Vizzini, L., Pearce, N., Delsedime, L., Merletti, F. and Richiardi, L. (2012). DNA methyltransferase 3b (DNMT3b), tumor tissue DNA methylation, Gleason score and prostate cancer mortality: investigating causal relationships. *Cancer Causes and*

Control, **23**, 1549–1555.

Goldberger, A. S. (1972). Structural equation methods in the social sciences. *Econometrica*, **40** (6), 979–1001.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F (Radar and Signal Processing)*, **140** (2), 107–113.

Gouriéroux, C. and Monfort, A. (1993). Simulation-based inference – a survey with special reference to panel-data models. *Journal of Econometrics*, **59**, 5–33.

Gouriéroux, C., Monfort, A. and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, **85**, 85–118.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81** (3), 515–526.

Grelaud, A., Robert, C. P., Marin, J. M., Rodolphe, F. and Taly, J. F. (2009). ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Analysis*, **4** (2), 317–336.

Jiang, W. and Turnbull, B. (2004). The indirect method: inference based on intermediate statistics – a synthesis and examples. *Statistical Science*, **19**, 239–263.

Joyce, P. and Marjoram, P. (2008). Approximately sufficient statistics and Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **7** (1), article 26.

Kartsonaki, C. (2013). ssfit: Fitting of parametric models using summary statistics. R package version 1.0. `http://CRAN.R-project.org/package=ssfit`.

Kharaghani, H. and Tayfeh-Rezaie, B. (2005). A Hadamard matrix of order 428. *Journal of Combinatorial Designs*, **13** (6), 435–440.

Kuha, J. and Goldthorpe, J. H. (2010). Path analysis for discrete variables: the role of education in social mobility. *Journal of the Royal Statistical Society* A, **173** (2), 351–369.

Lager, A. C. J., Modin, B. E., De Stavola, B. L. and Vågerö, D. H. (2012). Social origin, schooling and individual change in intelligence during childhood influence long-term mortality: a 68-year follow-up study. *International Journal of Epidemiology*, **41**, 398–404.

Lange, T. and Hansen, J. V. (2011). Direct and indirect effects in a survival context. *Epidemiology*, **22** (4), 575–581.

MacKinnon, D. P. and Dwyer, J. H. (1993). Estimating mediated effects in prevention studies. *Evaluation Review*, **17**, 144–158.

MacKinnon, D. P., Fairchild, A. J. and Fritz, M. S. (2007). Mediation analysis. *Annual Review of Psychology*, **58**, 593–614.

Marjoram, P., Molitor, J., Plagnol, V. and Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **100**, 15324–15328.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, **57** (5), 995–1026.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, **12** (2), 153–157.

Muthén, B. (2011). Applications of causally defined direct and indirect effects in mediation analysis using SEM in Mplus. `www.statmodel.com/download/causalmediation.pdf`

Neuhaus, J. M., Kalbfleisch, J. D., and Hauck, W. W. (1994). Conditions for consistent estimation in mixed-effects models for binary matched-pairs data. *The Canadian Journal of Statistics*, **22** (1), 139–148.

Nunes M. A. and Balding, D. J. (2010). On optimal selection of summary statistics for approximate Bayesian computation. *Statistical Applications in Genetics and Molecular Biology*, **9** (1), article 34.

Pearl, J. (2001). Direct and indirect effects. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, 411–420, Morgan Kaufmann Publishers Inc.

Pearl, J. (2012). The Mediation Formula: A Guide to the Assessment of Causal Pathways in Nonlinear Models, in Berzuini, C., Dawid, P. and Bernardinelli, L. (eds.) *Causality: Statistical Perspectives and Applications*. Chichester: John Wiley and Sons Ltd.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multi-factorial experiments. *Biometrika*, **33**, 305–325.

Ploubidis, G. B., Mathenge, W., De Stavola, B., Grundy, E., Foster, A. and Kuper, H. (2013). Socioeconomic position and later life prevalence of hypertension, diabetes and visual impairment in Nakuru, Kenya. *International Journal of Public Health*, **58**, 133–141.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. and Feldman, M. W. (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular Biology and Evolution*, **16**, 1791–1798.

R Development Core Team (2008). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. ISBN 3-900051-07-0, URL `http://www.R-project.org`.

Ratmann, O., Andrieu, C., Wiuf, C. and Richardson, S. (2009). Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proceedings of the National Academy of Sciences*, **106** (26), 10576–10581.

Richiardi, L., Fiano, V., Vizzini, L., De Marco, L., Delsedime, L., Akre, O., Gillio Tos, A. and Merletti, F. (2009). Promoter methylation in APC, RUNX3, and GSTP1 and mortality in prostate cancer patients. *Journal of Clinical Oncology*, **27** (19), 3161–3168.

Robert, C. P., Cornuet, J. M., Marin, J. M. and Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences*, **108** (37), 15112–15117.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, **3** (2), 143–155.

Robinson, L. D. and Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review*, **58** (2), 227–240.

Rodriguez-Iturbe, I., Cox, D. R. and Isham, V. (1987). Some models for rainfall based on stochastic point processes. *Proceedings of the Royal Society* A, **410**, 269–288.

Ross, G. J. S. (1972). Stochastic model-fitting by evolutionary operation. *Mathematical models in ecology*, 297–308.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, **12** (4), 1151–1172.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, **69** (1), 239–241.

Schumacher, M., Olscheski, M. and Schmoor, C. (1987). The impact of heterogeneity on the comparison of survival times. *Statistics in Medicine*, **6**, 773–784.

Sisson, S. A., Fan, Y. and Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, **104** (6), 1760–1765.

Struthers, C. A. and Kalbfleisch, J. D. (1986). Misspecified proportional hazard models. *Biometrika*, **73** (2), 363–369.

Student (1908a). The probable error of a mean. *Biometrika*, **6** (1), 1–25.

Student (1908b). Probable error of a correlation coefficient. *Biometrika*, **6** (2/3), 302–310.

Sunnåker, M., Busetto, A. G., Numminen, E., Corander, J., Foll, M. and Dessimoz, C. (2013). Approximate Bayesian Computation. *PLOS Computational Biology*, **9** (1), e1002803.

Tavaré, S., Balding, D. J., Griffiths, R. C. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, **145**, 505–518.

Tein, J.-Y. and MacKinnon, D. P. (2003). Estimating Mediated Effects with Survival Data, in Yanai, H., Rikkyo, A. O., Shigemasu, K., Kano, Y. and Meulman J. J. (Eds.) *New Developments on Psychometrics*, 405–412. Tokyo, Japan: Springer-Verlag Tokyo Inc.

Toni, T., Welch, D., Stelkowa, N., Ipsen, A. and Stumpf, M. P. H. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal*

Society Interface, **6**, 187–202.

VanderWeele, T. J. (2011). Causal mediation analysis with survival data. *Epidemiology*, **22** (4), 582–585.

VanderWeele, T. J. and Vansteelandt, S. (2010). Odds ratios for mediation analysis for a dichotomous outcome. *American Journal of Epidemiology*, **172** (12), 1339–1348.

Vansteelandt, S. (2012). Estimation of Direct and Indirect Effects, in Berzuini, C., Dawid, P. and Bernardinelli, L. (eds.) *Causality: Statistical Perspectives and Applications*, Chichester: John Wiley and Sons Ltd.

Wegmann, D., Leuenberger, C. and Excoffier, L. (2009). Efficient approximate Bayesian computation coupled with Markov Chain Monte Carlo without likelihood. *Genetics*, **182**, 1207–1218.

Wood, S. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, **466**, 1102–1104.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**, 162–177.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: an Introduction using R.* Boca Raton: Chapman and

Hall / CRC.