Conditional Likelihood for Risk Estimation in Genome Scans and Coefficient Shrinkage

by
Arpita Ghosh

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics, School of Public Health.

Chapel Hill
2009

Approved by:

Fred A. Wright, Advisor

Fei Zou, Co-advisor

Donglin Zeng, Committee Member

Andrew B. Nobel, Committee Member

Kari E. North, Committee Member

Abstract

ARPITA GHOSH: Conditional Likelihood for Risk Estimation in Genome Scans and Coefficient Shrinkage.
(Under the direction of Dr. Fred A. Wright and Dr. Fei Zou)

It is widely recognized that genome-wide association studies suffer from inflation of the risk estimates (commonly known as the "winner's curse" or "significance bias") for genetic variants, usually single nucleotide polymorphisms (SNP)s, identified as significant in the genome scan. To handle such significance bias, a number of investigators have proposed using likelihoods that condition on the declared significance of the outcome. We describe an approximate conditional likelihood approach that can be applied using estimates of odds ratios and their standard errors provided by standard statistical software. We also discuss extensions to the situation where, to supplement the primary analysis, risk estimation is performed for multiple correlated phenotypes or gene-environment interactions in the genome scan. The results have considerable importance for the proper design of follow-up studies and risk characterization. Our conditional likelihood approach also lends itself naturally to regression settings, in which shrinkage of multiple coefficients is performed. We use our conditional likelihood to propose a new regression penalty function, and demonstrate that it is competitive with other penalized regression procedures in both low-dimensional and high-dimensional settings.

Acknowledgments

I would like to thank my advisor Dr. Fred Wright for his constant support and excellent mentorship. He has always been patient and encouraging, and I greatly value his advice on various aspects of research. I am indebted to my co-advisor Dr. Fei Zou for her guidance and insightful suggestions. She has provided me with excellent resources and advice whenever I needed it.

I would like to thank Dr. Andrew Nobel for his guidance. I have always enjoyed his availability and encouragement in addition to his expertise on the subject. I am grateful to Dr. Donglin Zeng for his assistance throughout my dissertation. This is a great opportunity to thank him for his generous support and advice. I would like to thank Dr. Kari North for her helpful comments and suggestions.

I sincerely appreciate the support that I have received from the professors and the staff in the Biostatistics department in different stages of my dissertation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past few years a large number of genotype-phenotype associations have been reported in the genome-wide association study (GWAS) literature. However, many of these reported associations have not been replicated (Hirschhorn et al. 2002; Lohmueller et al. 2003). Among those associations that have been replicated, the estimated genetic effect size in the replication sample has often been smaller than that observed in the original GWAS (Todd et al. 2007; Yu et al. 2007; Levy et al. 2009). We investigate one of the possible reasons behind these phenomena, namely that estimates of effect sizes are upwardly-biased simply due to the fact that the genetic variants were *selected* for having achieved statistical significance.

In a GWAS the objective is to identify genetic variants (SNPs) conferring disease susceptibility. Once such a variant is detected, interest lies in quantifying the genetic effect of that variant on the phenotype, based on the same data. For modern whole genome scans, 100,000 to 1 million SNPs may be genotyped. To control family-wise error or false discovery rates, point-wise significance thresholds must be very conservative (Zondervan and Cardon 2007; Todd et al. 2007; Scott et al. 2007), typically in the range $10^{-7} - 10^{-8}$. Using the same dataset for testing and estimation purposes, together with the application of stringent thresholds, distorts the estimation process and produces

inflated estimates of effect sizes for significant SNPs. This phenomenon is commonly known as the "winner's curse" (Lohmueller et al. 2003; Zöllner and Pritchard 2007) or "significance bias" (Ghosh et al. 2008). For example, in a GWAS of blood pressure and hypertension for the CHARGE Consortium reported by Levy et al. (2009), effect size estimates were reported for all thirty SNPS representing the ten most significant loci for each of the three phenotypes systolic blood pressure, diastolic blood pressure, and hypertension. Each of these estimates was higher in magnitude than the estimates in a replication study reported by the Global BPgen Consortium, representing strong empirical evidence for the significance bias phenomenon.

One implication of significance bias is that if the biased estimates are used for the design of replication studies, these replication studies are likely to be underpowered. In addition, the true standard errors of risk estimates can be greatly inflated by the selection procedure. Also, standard confidence intervals for risk estimates can have very poor coverage properties. Although significance bias has been investigated and documented in detail, very few methods have been proposed for reducing or eliminating it.

We have developed a conditional likelihood approach to reduce this bias, which applies in a variety of testing settings (Ghosh et al. 2008). Among the most attractive settings is the analysis of case-control association studies, for which the number of tests (SNPs) is very large, and for which the genetic risk effects are important, interpretable quantities. Our approach is based on the estimate of the genetic effect and its standard error as reported by standard statistical software. Thus it does not require access to the original data and can be applied to published studies, and our method is far easier to implement than competing approaches. We also provide a principled method to construct confidence intervals for the genetic effect while acknowledging the conditioning on statistical significance. We have evaluated the performance of the proposed method

via extensive simulations for a range of genetic models, minor allele frequencies and genetic effect sizes. Finally, we have applied it to published datasets to demonstrate the relevance of our approach to modern whole genome scans.

As an extension to the problem of reducing significance bias for disease risk effect, we have considered the situation where selection of a significant SNP is performed on the basis of one trait, but we wish to perform inference on the risk effect of the significant SNP for another trait (e.g. Type II diabetes and obesity). An immediate problem arises in performing valid inference for the secondary phenotype for a (retrospective) case-control design. If the secondary phenotype is associated with the disease status that forms the basis for case-control comparison, then standard logistic or linear regression applied to the secondary trait can produce severely biased estimates of the secondary risk effects (Nagelkerke et al. 1995; Jiang et al. 2006; Lin and Zeng 2008).

While the secondary analysis for case-control association studies has been addressed by many researchers in the GWAS literature (Jiang et al. 2006; Scott and Wild 2002; Richardson et al. 2007; Scott and Wild 1991; Lee et al. 1997; Lin and Zeng 2008), we describe a novel retrospective likelihood method to analyze binary or continuous secondary phenotypes. The approach models the joint distribution of the disease and secondary phenotypes such that the marginals for each phenotype respect the conventional models. Specifically, we specify the joint distribution such that the marginal distribution for the disease phenotype is logistic and that for the secondary phenotype is logistic or linear for binary or continuous secondary phenotypes, respectively. The approach has considerable appeal as an alternative analysis procedure for secondary phenotypes.

Using our approach for secondary analysis, we next describe a general approach to bias-correction that includes correction for a variety of secondary effects, including risk effects for secondary phenotypes, as well as the estimation of gene-environment

interaction effects. Here bias correction is necessary if the the SNPs of interest have undergone initial significance selection for the primary phenotype, and it is of interest to perform inference for the secondary effects. We demonstrate that the significance bias problem can be substantial for the secondary effects, and is closely related to the correlation between the primary and secondary effect estimates. The problem of significance bias in the estimation of secondary effects has received relatively little attention in the GWAS literature. To address this problem, we have developed an extension of our conditional likelihood approach to a multivariate setting, where multiple effect coefficients are estimated simultaneously . For implementation of this method we require the estimates of the effect sizes, their standard errors, and an estimate of the covariance between them. For secondary phenotypes, we provide formulas to estimate the relevant covariances, and the method can be implemented very easily. We have also developed the method to handle the situation where gene effects, as well as environmental effects and gene-environment interactions are all estimated simultaneously. In addition, we have shown analytically that if we first fit a reduced model for disease risk on gene only, and follow up with a full model only if the reduced model is declared significant, then the effect size estimate for the gene-environment interaction obtained from the full model is not asymptotically biased.

Finally, it is worth noting that the bias-correction procedure is intended to reduce the mean-squared error of effect estimates, with significance thresholds that are especially useful when the proportion of true alternatives is low. Moreover, the effect estimates that are not significant may be thought of as having been thresholded to zero. In this manner, our conditional likelihood approach may be compared to the shrinkage of coefficient estimates and thresholding that is applied in existing penalized regression procedures. We demonstrate that our conditional likelihood can be used to formulate a new penalty that can be used in a regression framework in situations

where the sample size is larger than the number of predictors. By implementing a significance threshold as a tuning parameter for individual predictors, the procedure can create a sparse set of predictors with non-zero coefficient estimates. In addition, we describe a conditional regression procedure that can be used to obtain estimates for our method when the number of predictors is larger than the sample size. When combined with cross-validation, our procedure is an automatic variable selection and coefficient shrinkage approach that is a competitive approach for prediction in high-dimensional regression settings. We demonstrate via simulation that the procedure has good prediction error properties in comparison to competing approaches, especially when the proportion of nonzero coefficients is small.

Chapter 2

Estimating Odds Ratios for Disease Risk in Genome Scans

2.1 Introduction

In genetic studies, it is widely recognized that the control of genome-wide error requires the use of stringent thresholds for significance testing. For genome-wide linkage scans, standard LOD significance thresholds in the range 3.0 to 4.0 correspond to point-wise p-values in the range 10^{-4}–10^{-5}, depending on the model and study design (Lander and Kruglyak 1995). For modern genome-wide association studies (GWAS), 100,000 to 1 million SNP markers may be genotyped, and control of family-wise error or false discovery rates typically requires point-wise significance thresholds in the range 10^{-7}–10^{-8} (Zondervan and Cardon 2007; Todd et al. 2007; Scott et al. 2007). The use of such stringent thresholds is offset somewhat by the belief that GWAS offer greater power than linkage studies for detecting complex disease genes (Risch and Merikangas 1996). Nonetheless, the application of stringent thresholds distorts the inferential process, producing estimates of disease risk effect sizes that may be, on average, far greater in magnitude than the true effect (Lander and Kruglyak 1995; Zondervan and

Cardon 2007; Allison et al. 2002; Chanock et al. 2007; Garner 2007; Göring et al. 2001; Hirschhorn et al. 2002; Ioannidis et al. 2001; Lohmueller et al. 2003; Siegmund 2002; Sun and Bull 2005; Yu et al. 2007; Zöllner and Pritchard 2007). This phenomenon has been described as a form of "winner's curse" by Zöllner and Pritchard (2007) and others, or as a form of regression to the mean (Yu et al. 2007), and has profound importance for genome scans. Although the problem has been described as primarily an issue of bias, we demonstrate below that the variance of risk estimates can also be greatly inflated by the selection procedure. Moreover, standard confidence intervals for risk estimates will have very poor coverage properties, although this issue seems to have received less attention.

Consider a genome association scan for a complex disease in which 10 genomic regions contain disease genes, and each region has a 20% chance of meeting genome-wide significance. Assuming independence of regions, the genome scan has respectable power $1 - (1 - 0.2)^{10} = 0.89$ to achieve significance in at least one region. However, a repeated genome scan of equal size will have power of only 0.2 for any one region, and thus likely not result in "replication" of the first study. A follow-up study might focus on a single significant region, using fewer markers and paying a lower penalty for multiple comparisons. But if the results of the initial genome scan are used as a guide, the follow-up study is likely to be underpowered, relying on an inflated estimate of locus disease risk.

As a statistical phenomenon, the winner's curse should not be confused with additional sources of bias, including variations due to genotyping technologies, or heterogeneity of patient populations from which samples are drawn (Lohmueller et al. 2003; Balding 2006; Wang et al. 2005). The winner's curse is investigated in detailed simulations elsewhere (Garner 2007; Göring et al. 2001; Siegmund 2002; Sun and Bull 2005; Yu et al. 2007; Zöllner and Pritchard 2007), including a recent paper by Garner (2007),

who clarified that the bias can be understood largely through the behavior of Wald statistics for log odds ratios.

Although the bias is simple to understand and to document, reducing or eliminating it may be nontrivial. Zöllner and Pritchard (2007), have described a likelihood approach which requires maximization over numerous parameters, including genotype frequencies and penetrance parameters, while conditioning on declared statistical significance. Their procedure reduces the bias in risk estimation, but cannot be performed using standard statistical software. Yu et al. (2007) have recently applied bootstrapping to correct for significance bias Both of these bias correction approaches are technically feasible for genome scans, but would be highly computationally intensive in that setting.

We describe our alternative approach for estimating genetic effects in terms of odds ratios, which have numerous advantages that have made them standard for analysis of case-control designs (Aschengrau and Seage 2003). A crucial advantage for case-control studies is that the odds ratio (OR) may be estimated consistently, whether the study design is prospective or retrospective (McCullagh and Nelder 1989), and the OR has an interpretation distinct from nuisance parameters such as genotype frequencies. Moreover, in logistic models the OR retains interpretability in the presence of covariates, which is increasingly important for complex disease investigations.

In this paper we introduce a method to correct for significance bias in disease association studies, using an approximate conditional likelihood. The approach is directly based on the OR estimate and its standard error as reported by standard statistical software, and applies to dominant, recessive, or additive genetic models. No modification is necessary when covariates such as population stratification variables have also been fit in the model. The approach may even be applied to published results without access to the original data. In addition, we develop a method to construct accurate confidence intervals for the OR.

We illustrate the performance of our approach via extensive simulations of a disease SNP analyzed by logistic regression. The simulations cover a range of models, disease allele frequencies, and OR values. Compared to nave OR estimation, our approach provides greatly reduced bias and mean-squared error, particularly for the modest effect sizes likely to be encountered in complex diseases. In addition, our confidence interval procedure provides coverage that is accurate or slightly conservative. Performing simulations for OR values near the null presents a challenge, because significant results are very rare when applying genome-wide thresholds. We thus employ a screening approach in which a deterministic trend statistic is used to identify datasets potentially significant in logistic regression.

2.2 Methods

We assume a genetic model with one parameter for the effect of disease genotype, which includes recessive, dominant, and additive models. We use $\beta = \log(OR)$ to denote the true loge odds ratio for disease risk conferred by a referent genotype, or for the contribution of each allele in an additive model. A single locus test statistic for disease association can be expressed as an estimate for β divided by an estimate for its standard error,

$$z = \frac{\hat{\beta}}{\hat{SE}(\hat{\beta})} \tag{2.1}$$

which is compared to the asymptotic null distribution $N(0, 1)$. We will refer to $\hat{\beta}$ and $\hat{SE}(\hat{\beta})$ as naïve estimators, as they are obtained from standard statistical procedures without acknowledging selection based on significance. For our problem, we wish to estimate β only when the SNP is significant in two-sided testing, i.e., $|z| > c$ for a value c corresponding to genome-wide significance. By explicitly considering this selection, below we obtain three new estimators and a confidence interval procedure.

Our approach offers marked improvements over $\hat{\beta}$ and standard confidence intervals. Our exposition includes mathematical and motivational details that we believe will considerably demystify the problem, which has until now appeared more obscure and complex than necessary. The performance of our new estimators is described in the Simulations subsection 2.2.8.

2.2.1 Significance bias (the winner's curse)

When logistic regression is used to test for genetic association, the Wald statistic for genetic effect assumes the specific form of (2.1), with numerator and denominator obtained from maximum likelihood and the information matrix (McCullagh and Nelder 1989; Agresti 2007; Cox and Snell 1989). However, the essence of our approach applies to a wide variety of testing procedures, for which the key requirements typically hold: (i) asymptotic normality of $\hat{\beta}$, and (ii) consistency of the standard error estimate, so that $\hat{SE}(\hat{\beta})/SE(\hat{\beta}) \to 1$. Expressing the test statistic in the form (2.1) provides a straightforward illustration of significance bias, and points the way toward corrected estimation procedures. Related test statistics based on maximum likelihood ratios, efficient scores, or directly based on contingency tables, are all asymptotically equivalent to (2.1) for local departures from the null hypothesis $H_0 : \beta = 0$ (Rao 1973), although this asymptotic equivalence is not necessary to apply our approach. The remainder of this subsection is similar to Garner (2007), but our explicit and expanded treatment provides the grounds for later development.

For large samples, $\hat{SE}(\hat{\beta})$ does not vary markedly in repeated data realizations. Thus the estimate $\hat{\beta}$ and its statistical significance are highly correlated (Garner 2007) and the problem can be restated as single-parameter estimation for a truncated normal distribution. To see this, we define $\mu = \beta/\hat{SE}(\hat{\beta})$, with $Z \dot{\sim} N(\mu, 1)$. Our use of this approximation follows from the standard result $Z - \mu = \frac{\hat{\beta}-\beta}{\hat{SE}(\hat{\beta})} \xrightarrow{D} N(0, 1)$ for increasing

sample size (Wald 1943). The statistical procedures to follow are developed entirely this "μ version" of the problem, which has been greatly simplified by the variance standardization.

Our naïve estimate of μ is $\hat{\mu} = z$, and the expectation can be shown analytically to be

$$E_\mu(Z \big| |Z| > c) = \mu + \frac{\phi(c-\mu) - \phi(c+\mu)}{\Phi(-c+\mu) + \Phi(-c-\mu)}, \tag{2.2}$$

where ϕ and Φ are the density and cumulative distribution function of a standard normal, respectively (see Appendix A). This is the two-sided rejection version of a result given by Garner (2007). As we detail in Results (Section 2.3), the bias can be substantial in realistic settings. In the special case of the null hypothesis $\mu = 0$, it is clear from (2.2) that the naïve estimate z is unbiased, because the two-sided testing procedure is equally likely to falsely declare positive or negative risk (*i.e.*, a protective effect of the referent genotype). It is not clear that the lack of bias for naïve estimation under the null has been fully appreciated (e.g., Figure 2 in Zöllner and Pritchard (2007) does not display the exact null value). However, this lack of bias requires averaging over rejections for both positive and negative z. In any significant dataset, $\hat{\mu}$ must be less than $-c$ or greater than c, and so will be far from the truth under the null. In other words, the lack of bias under the null is offset by very large variance.

2.2.2 An approximate conditional likelihood

The approximating distribution of Z suggests a correspondingly approximate likelihood for μ,

$$L(\mu) = p_\mu(z) = \phi(z - \mu) \tag{2.3}$$

The likelihood applies generally to a wide variety of testing procedures, eliminating any nuisance parameters that have been included in the modeling, including stratification

variables, clinical covariates, or the effects of other SNP genotypes. It is easy to show that the maximum likelihood estimate (MLE) is $\hat{\mu} = z$. A standard approach to likelihood testing for $H_0 : \mu = 0$ (Wald 1943) involves comparing the maximum log-likelihood ratio $LLR = -2\ln(L(\hat{\mu})/L(0))$ to a χ^2_1 density. It is also simple to show that here $LLR = z^2$, so in terms of both estimation and testing, the likelihood simply recapitulates the initial equation (2.1). The advantage to (2.3), however, is that it provides a simple and transparent approach to handle the conditioning. Acknowledging the event that the SNP is declared statistically significant, we have the *conditional* likelihood

$$L_c(\mu) = p_\mu(z\big||Z| > c) = \frac{p_\mu(z)}{P_\mu(|Z| > c)} = \frac{\phi(z-\mu)}{\Phi(-c+\mu) + \Phi(-c-\mu)}, \tag{2.4}$$

Under (2.4), the relationship between numerator and denominator is such that, for a given z, it is quite possible that the most likely value for μ is in the interval $[-c, c]$, even though z itself is conditioned to be outside that range.

Using this conditional approximate likelihood we now derive improved estimators of μ. For any proposed value of μ, we can convert back to the desired log odds ratio using $\beta = \mu\ \hat{SE}(\hat{\beta})$, where $\hat{SE}(\hat{\beta})$ is obtained from standard approaches (i.e., does not consider the significance selection). One remarkable feature of our approach is that we can apply it to published summary results. To do so, we require only the significance threshold c, $\hat{\beta}$, and $\hat{SE}(\hat{\beta})$. The standard error, if not provided directly, can be inferred from c, $\hat{\beta}$, and any one of the following: z, the p-value, or an unconditional OR confidence interval.

2.2.3 The conditional MLE

Using the conditional likelihood, the maximum likelihood principle suggests the MLE estimator,

$$\tilde{\mu_1} = \arg\max_{\mu} L_c(\mu),$$

which can be obtained using numerical maximization for any z and c (hereafter "$\sim$" will signify estimates based on the conditional likelihood). Note that in this setting the conditional maximum likelihood estimate provides no guarantee of unbiasedness or efficiency, a fact that does not appear to have been considered by other investigators. We have already applied large-sample assumptions in constructing the conditional likelihood (2.4), but as we show below, other estimators can provide reduced bias or mean-squared error for certain ranges of μ, and therefore β.

Motivated by bias-reduction, one might attempt to directly correct the bias in $\hat{\mu}$ by solving for μ in the equation $E_\mu(Z \big| |Z| > c) = z$. Such an estimator has intuitive appeal, representing the value of μ for which, after conditioning on significance, we would have *expected* to observe z. Perhaps surprisingly, this "bias-correction" estimator in fact turns out to be $\tilde{\mu_1}$. To see this, we take the derivative of the conditional likelihood with respect to μ, for which the identity $L_c'(\tilde{\mu_1}) = 0$ implies

$$z = \tilde{\mu_1} + \frac{\phi(c - \tilde{\mu_1}) - \phi(c + \tilde{\mu_1})}{\Phi(-c + \tilde{\mu_1}) + \Phi(-c - \tilde{\mu_1})}, \tag{2.5}$$

Comparing equation (2.3) to (2.5) implies that the bias-correction estimator and $\tilde{\mu_1}$ are the same. Similar estimators have been examined in the context of sequential clinical trials, in which effect parameters are estimated only after a stopping boundary has been reached (Liu et al. 2004). Despite its secondary motivation as a bias-correction estimator, the conditional MLE $\tilde{\mu_1}$ is not in fact unbiased, due to nonlinearity in the bias of the naïve estimator $\hat{\mu}$. Moreover, in this setting the conditional MLE has no

special optimality properties, and other estimators may be reasonable. Nonetheless, we will show that $\tilde{\mu_1}$ is markedly improved over the naïve estimator, both in terms of bias and mean squared error.

2.2.4 The mean of the normalized conditional likelihood

The motivation to reduce mean-squared error (MSE) suggests another, perhaps less-obvious estimator,

$$\tilde{\mu_2} = \frac{\int_{-\infty}^{\infty} \mu L_c(\mu) d\mu}{\int_{-\infty}^{\infty} L_c(\mu) d\mu}, \tag{2.6}$$

which is easily calculated numerically. $\tilde{\mu_2}$ is the mean of the random variable following the distribution $L_c(\mu)$, normalized to be a proper density. $\tilde{\mu_2}$ has favorable MSE properties when averaged across a wide range of μ. This fact follows from an interpretation of $\tilde{\mu_2}$ as a posterior mean in a Bayesian treatment of the problem with a flat prior on μ (Leonard and Hsu 1999). However, $\tilde{\mu_2}$ is considered here as an entirely frequentist estimate, with bias and error examined at each value of μ and judged accordingly. For $|z|$ near the boundary c, $\tilde{\mu_2}$ typically represents a less aggressive shrinkage toward 0 compared to $\tilde{\mu_1}$.

2.2.5 A compromise estimator

In the treatment below, we will see that the conditional likelihood is typically skewed, and so $\tilde{\mu_1}$ and $\tilde{\mu_2}$ can differ appreciably for certain values of z. $\tilde{\mu_2}$ can show higher MSE than $\tilde{\mu_1}$ for μ near zero, but is more favorable for μ away from zero, while the bias of $\tilde{\mu_1}$ and $\tilde{\mu_2}$ can be of opposite signs for μ near the significance threshold c. Thus as a practical compromise we also examine the estimator

$$\tilde{\mu_3} = (\tilde{\mu_1} + \tilde{\mu_2})/2,$$

which balances the strengths of $\tilde{\mu_1}$ vs. $\tilde{\mu_2}$.

2.2.6 Illustrations of the conditional likelihood

Figure 2.1 illustrates the conditional and unconditional likelihoods assuming an illustrative constant threshold $c = 5.0$. Panels (a)-(c) correspond to $z = 5.2$, 5.33, and 6.0, respectively. For each panel, the unconditional likelihood is centered and maximized at z (indicated by a dot on each plot). For panel (a), when z is only slightly above the threshold, the conditional likelihood is in contrast shifted aggressively towards zero ($\tilde{\mu_1} = 0.66$, $\tilde{\mu_2} = 2.53$, $\tilde{\mu_3} = 1.60$). When z is well above the threshold (z=6.0, panel (c)) this shift is much smaller ($\tilde{\mu_1} = 5.48$, $\tilde{\mu_1} = 4.94$, $\tilde{\mu_1} = 5.21$). For an intermediate z (panel (b)), the shift is intermediate. Note that our estimates are obtained here for the μ version of the problem, and the conversion $\beta = \mu\hat{SE}(\hat{\beta})$ must be performed before the results are interpreted on the log-odds scale.

As desired, the conditional likelihood shows a clear shift toward zero. But why is the shift so extreme, e.g., when $z = 5.2$? Such a z-value (which is equivalent to $\hat{\mu}$) has already met genome-wide multiple-testing correction for statistical significance, but a shrinkage from $\hat{\mu} = 5.2$ to $\tilde{\mu_1} = 0.66$ (for example) will effect a corresponding proportional reduction in the log odds ratio. Thus it seems our proposed estimation procedures can often adjust the estimated effect size to be *practically* insignificant. To see why the result is reasonable, consider that the conditional likelihood, as a frequentist construction, makes no judgment about the prior plausibility of various values of μ. When presented with a value z for each μ, it considers only the chance that z would have arisen, given that $|z| > c$.

Figure 2.2(d) presents the (truncated normal) conditional densities for z under $\mu = 0.66$ and $\mu = 5.2$. These μ values were chosen because they represent the conditional and unconditional MLEs when $z = 5.2$. Note that these curves are conditional densities

for z, not likelihoods. However, for a fixed value of z, the relative heights of the two curves reflect the conditional likelihoods for the two competing values of μ. From the curves we can see the value $z = 5.2$ is 2.77 times more likely to arise when $\mu = 0.66$ than when $\mu = 5.2$. Expressed in another way, when μ values are truly of large magnitude, then z tends to overshoot the threshold c by a greater amount than was observed here for $z = 5.2$. Thus in this instance we would conclude that μ is not likely to be of large magnitude.

Our three proposed estimators can be easily computed numerically, and simple `R` and Excel programs to do so are available at our website `www.bios.unc.edu/~fwright/genomebias`. Using the threshold $c = 5$ for illustration, we have calculated the conditional expectations and MSEs for the three estimators, shown in Figure 2.2[(a)-(b)]. The three corrected estimators provide dramatically reduced bias compared to the naïve estimator for much of the range of μ. For $\mu = 0$, by symmetry all estimators are unbiased. For $|\mu|$ considerably larger than c, all methods will give estimators near z and will be nearly unbiased. The corrected estimators tend to under-correct for small μ and over-correct for large μ. The conditional MLE $\tilde{\mu}_1$ can be viewed as a first-order attempt to correct the bias, while the data z occupies the same range whether μ is small or large. In a sense, the corrected estimate splits the difference between the two extremes, leading to the observed pattern.

The MSE for $\hat{\mu} = z$ is extremely large for μ near zero, as predicted. MSEs for the corrected estimators are considerably smaller in the range of small to moderate μ. As described above, these estimators are easily converted to the corresponding improved log(OR) estimators $\tilde{\beta}_1$, $\tilde{\beta}_2$, $\tilde{\beta}_3$. Moreover, for large samples the bias and MSE properties for μ will largely carry over to real data, essentially with a rescaling of the axes to convert μ to β.

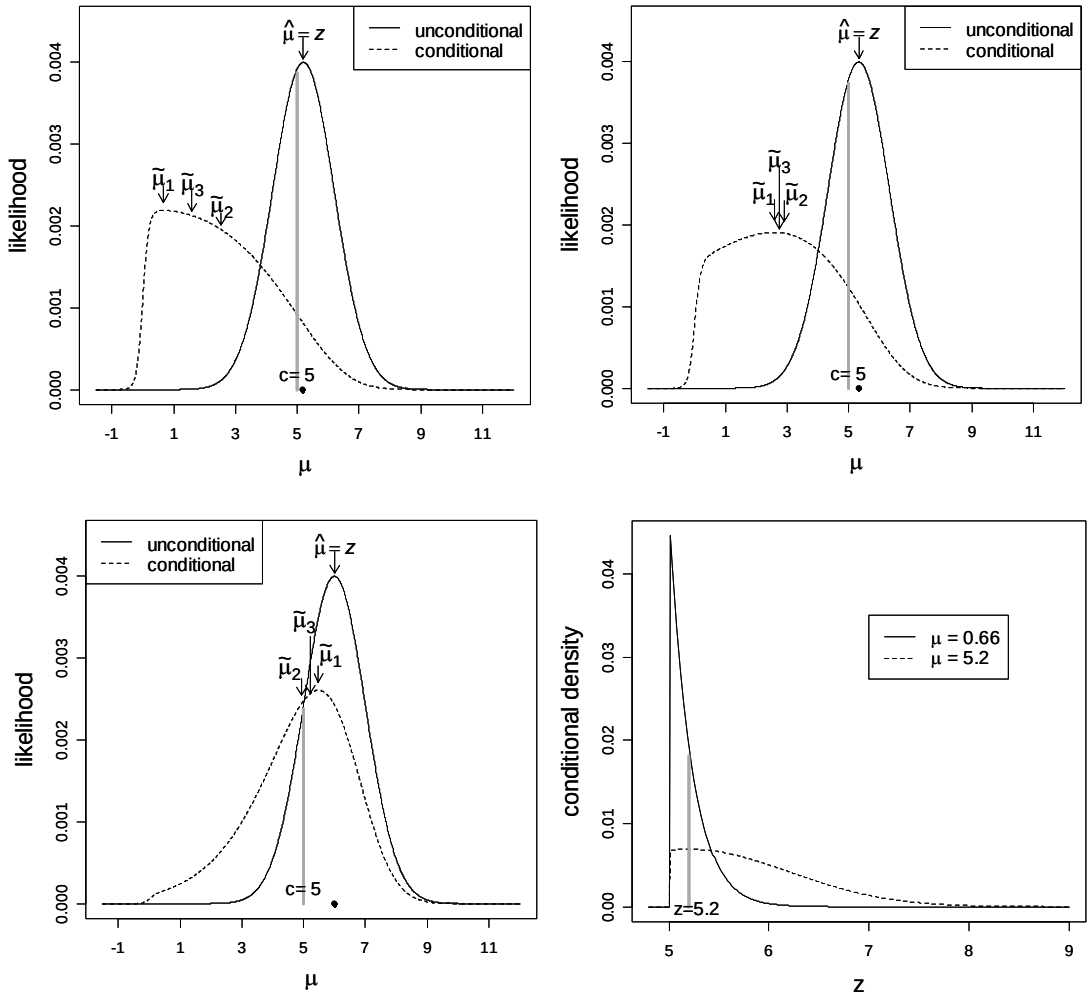


Figure 2.1: **Behavior of the unconditional and conditional likelihoods for μ.** Unconditional and conditional likelihoods of μ are presented for (a) $z = 5.2$, (b) $z = 5.33$ and (c) $z = 6$. The location of the observed z is indicated by a black dot on each plot. The conditional likelihood changes considerably for small changes in z near c. For larger z, the conditional likelihood approaches the unconditional likelihood. Likelihoods for $\mu < -c$ are negligible and not shown. (d): Conditional densities of z for $\mu = 0.66$ and $\mu = 5.2$, with the relative likelihoods highlighted for a fixed value $z = 5.2$.

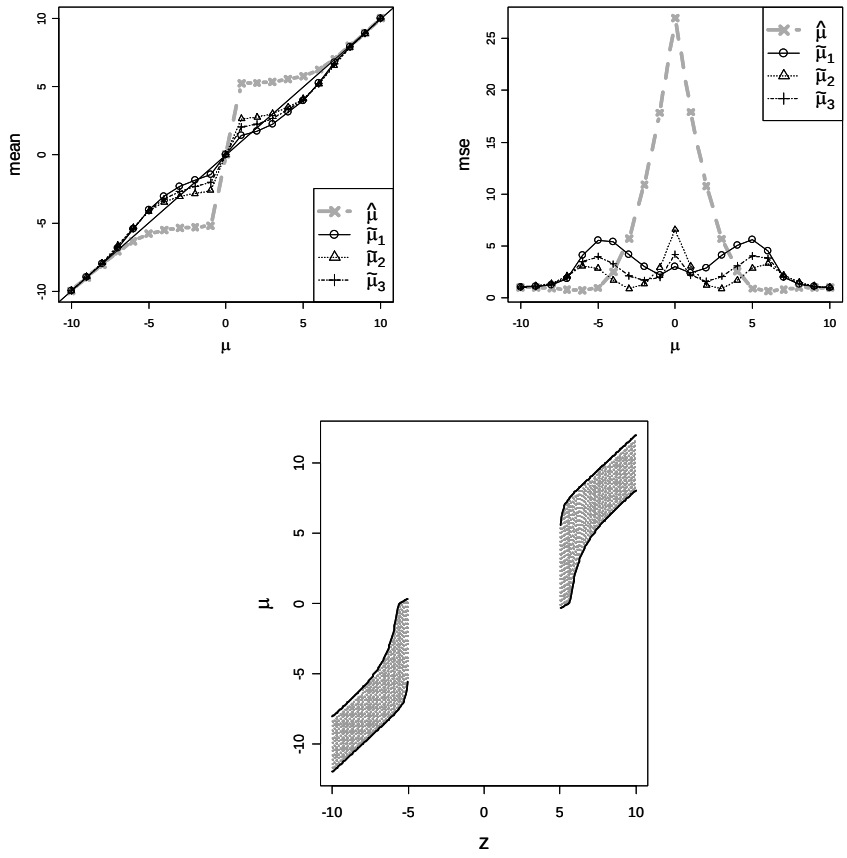


Figure 2.2: **Estimators and confidence intervals for μ with significance threshold $c = 5$.**
(a) The expectation of naïve estimator $\hat{\mu}$ shows substantial bias and (b) very large mean squared error for much of the range of μ, while the corrected estimators have lower bias and MSE (c) Upper and lower confidence bounds for μ as a function of the observed statistic z.

2.2.7 Conditional confidence intervals

Proper interpretation of the corrected μ estimates requires an understanding of estimation error, conditioned on statistical significance. Standard confidence interval (CI) procedures fail in this setting. For example, after conditioning on significance, a standard 95% CI for μ cannot contain 0, for otherwise it would not have been significant. Thus, when $\mu = 0$ the standard CI procedure has zero conditional coverage probability. Zöllner and Pritchard (2007) addressed this issue by using a standard maximum likelihood ratio approach applied to the conditional likelihood. In our setting, a $1-\eta$ CI created in this manner would consist of all μ values such that $2\ln(L_c(\tilde{\mu_1})/L_c(\mu)) \leq q_{1-\eta}$, where $q_{1-\eta}$ is the $1-\eta$ quantile of a χ^2_1 density. However, we have shown via numerical integration that in the μ version of the problem, the true coverage probability of this CI procedure can exhibit markedly conservative or anticonservative departures from $1-\eta$, depending on the true μ. Approaches using the second derivative at $\ln L_c(\tilde{\mu_1})$ to estimate the error variance also fail. The difficulty arises because the conditional m.l.e is not normally distributed, nor is the shape of $L_c(\mu)$ approximately normal for a realized dataset.

To create confidence intervals with correct conditional coverage, we return to the original Neymanian concept of a confidence region (Rao 1973; Lehmann and Casella 1983), which can always be applied when the distribution of a test statistic is known for each value of the unknown parameter. Let $A(\mu, 1-\eta)$ be an acceptance region depending on μ such that

$$P_\mu\left(Z \in A(\mu, 1-\eta) \big| |Z| > c\right) = 1-\eta\ .$$

Given an observed z, the confidence region consists of all values μ such that $z \in A(\mu, 1-\eta)$. It is straightforward to show that this approach gives *exact* coverage probability

$1-\eta$ for any μ. Among possible acceptance regions, we choose $A(\mu, 1-\eta)$ as the interval between the $\eta/2$ and $1-\eta/2$ quantiles of the conditional density $p_\mu(z||Z|>c)$. Note that, although we have presented three competing point estimates for μ, our procedure yields only a single CI. Figure 2.2 (c) shows the upper and lower confidence limits for our CI procedure for each z. Note that the limits are wider when $|z|$ is near c, reflecting less certainty about μ, and can even contain $\mu = 0$. This does not contradict the statistical significance - the intent of the procedure is to obtain correct coverage for any μ (including $\mu = 0$) after conditioning on significance. The conversion of the confidence limits to the β scale is $\left(\mu_{lower}\hat{SE}(\hat{\beta}),\ \mu_{upper}\hat{SE}(\hat{\beta})\right)$. Although our procedure is guaranteed correct conditional coverage in the idealized μ setting, our CI for β relies on large-sample normality assumptions for $\hat{\beta}$. Thus we investigate empirical coverage of our procedure in the Results Section.

2.2.8 Simulations

To describe our simulations, we begin with basic notation for disease association studies. We let y denote the disease status (0=control, 1=case) for an individual, and x denote the SNP genotype predictor value. For a bi-allelic SNP with major allele A and minor allele a, x is defined as follows for genetic models with respect to a:

Recessive

$$g = \begin{cases} 0, & AA \\ 0, & Aa \\ 1, & aa \end{cases}$$

Additive

$$g = \begin{cases} 0, & AA \\ 1, & Aa \\ 2, & aa \end{cases}$$

Dominant

$$g = \begin{cases} 0, & AA \\ 1, & Aa \\ 1, & aa\ . \end{cases}$$

We assume the logistic model for a randomly sampled individual in the population

$$\log\left(P(Y=1|x)/(1-P(Y=1|x))\right) = \alpha + \beta x,$$

for some α, and β is the log odds ratio for a unit increase in x. Rather than specifying α directly, it is more interpretable to solve for α for a specified allele frequency and disease prevalence π. The marginal frequency of x is denoted $p(x)$, and is easily calculated from Hardy-Weinberg assumptions. With fixed disease prevalence, the identity $\pi = \sum_x \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)} p(x)$ was used to calculate α. Finally, solving for the genotype probabilities conditioned on case/control status yields

$$P(X=x|Y=1) = \frac{p(x)}{\pi}\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

$$\text{and} \quad P(X=x|Y=0) = \frac{p(x)}{1-\pi}\frac{1}{1+\exp(\alpha+\beta x)}\,.$$

A standard result is that logistic modeling for β applies even when the data are sampled retrospectively (McCullagh and Nelder 1989).

Each dataset was simulated and analyzed in `R v.2.5.1`. We will denote the total sample size $n = n_{cases} + n_{controls}$, and $n_{cases} = n_{controls}$ throughout. Most simulations consisted of $n = 1000$. This sample size is relatively small for a genome scan, and was intentionally chosen to emphasize any departures from normality or difficulties in estimating $SE(\hat{\beta})$. Larger sample sizes were also examined for several of the setups to examine the effect of sample size on bias, MSE, and confidence coverage. We assumed a disease prevalence of 0.01 throughout - the retrospective sampling is not very sensitive to this specification. We examined β ranging from -0.7 ($OR \approx 0.5$) to 0.7 ($OR \approx 2$). This range corresponds to biological plausibility for complex disease (Ioannidis et al.

2001), and ensures that simulations span the range from low power to high power. For simplicity, we used $c = 5.0$, corresponding to a single p-value of 5.7×10^{-7}, near the genome-wide threshold considered by others (Zondervan and Cardon 2007; Todd et al. 2007; Scott et al. 2007).

For recessive models we considered MAF values of 0.25 and 0.5 - lower values created small expected cell counts that were problematic for sample sizes of 500 in each group. For the additive and dominant models we considered minor allele frequency (MAF) values of 0.05, 0.1, 0.25, and 0.5. A single setup consisted of the genetic model, MAF, and β, and sufficient simulations were performed for each setup to obtain 1000 significant datasets. Setups with $\beta = 0$ required on the order of $10^9 - 10^{11}$ simulations for this rarefied threshold. We sped up the analysis by first applying a chi-square test (Cochran-Armitage trend test for the additive model) to the datasets, which can be obtained without iterative maximization. The chi-square statistic was determined to have a close correspondence to z^2 obtained from the more computationally intensive logistic regression, and a chi-square statistic ≥ 24 was determined to capture essentially all datasets with $z^2 \geq c^2 = 25$. Datasets meeting the chi-square criterion were analyzed via logistic regression in `R glm`. For datasets achieving final significance as determined by logistic regression, $\hat{\beta}$ and $\hat{SE}(\hat{\beta})$ were used to obtain $\tilde{\beta}_1$, $\tilde{\beta}_2$, $\tilde{\beta}_3$, and conditional confidence intervals.

2.3 Results

In all scenarios described here, expectations and mean-squared errors are calculated conditional on significance, i.e., $|z| > c$.

2.3.1 Bias

The top row of Figure 2.3 plots the means for each of the naïve and corrected estimators vs. β (with corresponding OR values) for all models, with MAF=0.25. The naïve estimator shows very large bias, especially for moderate β. All of the corrected estimators show dramatically reduced bias across most of the range examined. For each model, the corrected estimates tend to under-correct for small (magnitude) β while overcorrecting for large β. All of the methods become nearly unbiased for large β, as they must, for the conditional and unconditional likelihoods are nearly identical when $|z|$ is well beyond c. In terms of bias, $\tilde{\beta}_1$ performs best among the corrected estimates for small . However, the over-correction of the conditional MLE can be substantial for moderate to large β, especially for the recessive model. $\tilde{\beta}_2$ shrinks the estimates toward zero less dramatically, resulting in under-correction for a larger part of the range of β. $\tilde{\beta}_3$ strikes a balance between the other two corrected estimates, and has much improved bias for moderate β under the recessive model. All estimators are effectively unbiased for $\beta = 0$. A subtle asymmetry in the plots for positive and negative $\log(OR)$, most evident in the recessive model, occurs because MAF< 0.5 and, for a fixed prevalence, the logistic intercept α depends on β.

2.3.2 Mean squared error

The corresponding MSE values for the estimators are shown in the bottom row of Figure 2.3. The naïve estimator $\hat{\beta}$ exhibits extremely large MSE for most β values examined. For β this is due to high variance, while for moderate β the naïve estimator has low variance but high bias. The corrected estimators show dramatically improved MSE for β in the interval [-0.3, 0.3] (OR ranging from 0.74 to 1.35) that encompasses the bulk of significant associations thus far for complex diseases (Todd et al. 2007; Scott et al. 2007). The MSEs of $\tilde{\beta}_1$ and $\tilde{\beta}_2$ are largely complementary. At $\beta = 0$,

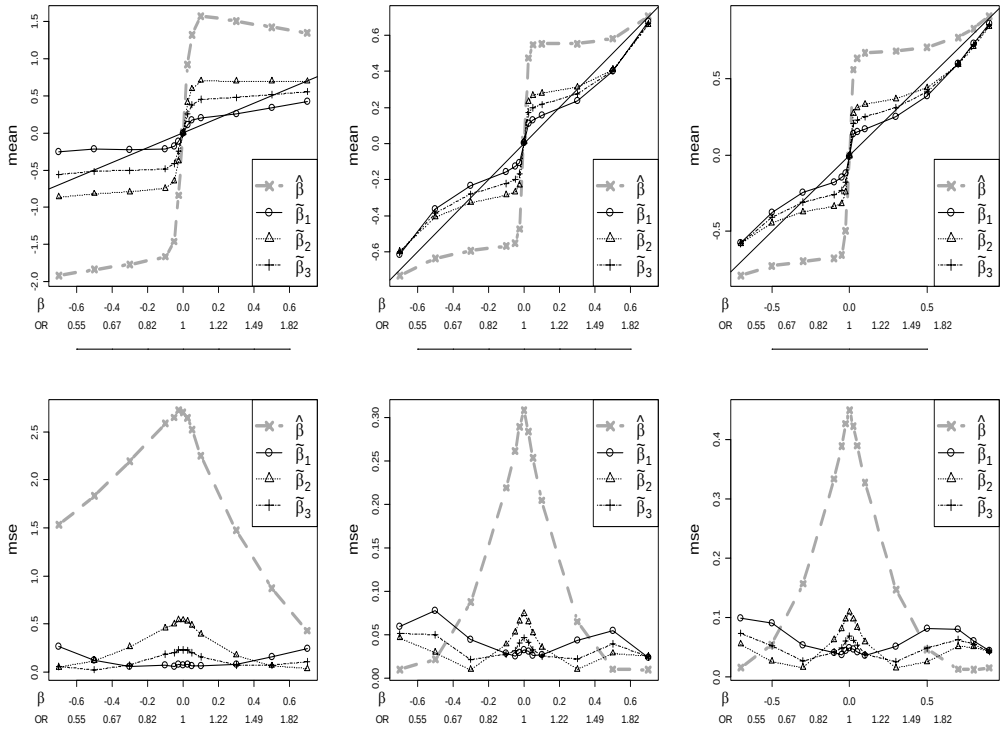


Figure 2.3: **Expectations and mean squared errors for the three genetic models under MAF=0.25.**
For the three models and MAF=0.25, the corrected estimators show greatly improved performance for much of the range of β. *Top row*: expected values for the naïve and conditional likelihood estimators vs. β. *Bottom row*: mean squared errors for the estimators. The y-axes for the MSE plots are rescaled to highlight details - the MSE is considerably larger for the recessive model due to scarcity of the risk homozygotes.

the MSE($\tilde{\beta}_1$) is fairly low, while MSE($\tilde{\beta}_2$) peaks. For larger magnitude β, the roles reverse. As expected, $\tilde{\beta}_3$ exhibits a more even MSE across the range, and represents a reasonable choice for stable error characteristics. For the additive and dominant models, $\hat{\beta}$ exhibits very low MSE for large β. This phenomenon is not as attractive as it appears, essentially resulting from a boundary effect in which $\hat{\beta}$ is nearly constant because z is just barely significant. In particular, for β outside of the plotted range, m.s.e($\hat{\beta}$) rises again to the var($\hat{\beta}$) value encountered in the unconditional setting.

The empirical bias and MSE observed in our simulations essentially follow the results from the version of the estimation problem, with a rescaling of the axes to convert μ

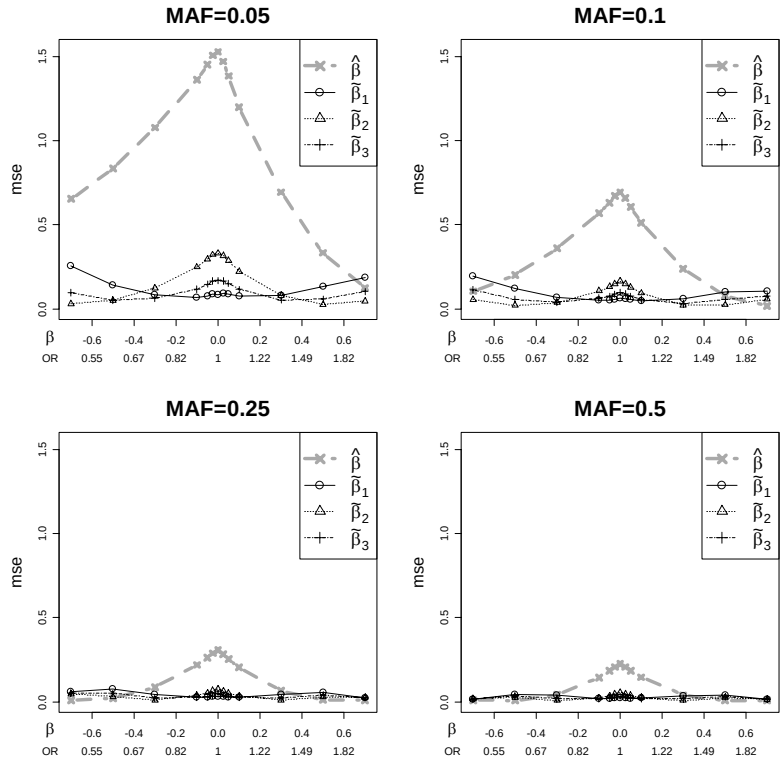


Figure 2.4: **Mean squared errors of the estimators vs. β for MAF values ranging from 0.05 to 0.5.**
The additive model is assumed, with $n = 1000$. The MSEs drop for larger MAF, but the relative performance of the estimators is maintained.

to β. Our empirical results for the remaining MAF values are plotted in Supplemental Figure 2.6, and largely follow the results described for MAF=0.25. Figure 2.4 shows a portion of these results for the additive model, in which the MSE is shown to drop for all estimators as the MAF increases. This occurs because for small MAF the MSE is largely driven by the heterozygote genotype counts, which increase with the MAF. The key point of Figure 2.4 is that the relative advantages of the corrected estimators are preserved across a wide range of MAF values.

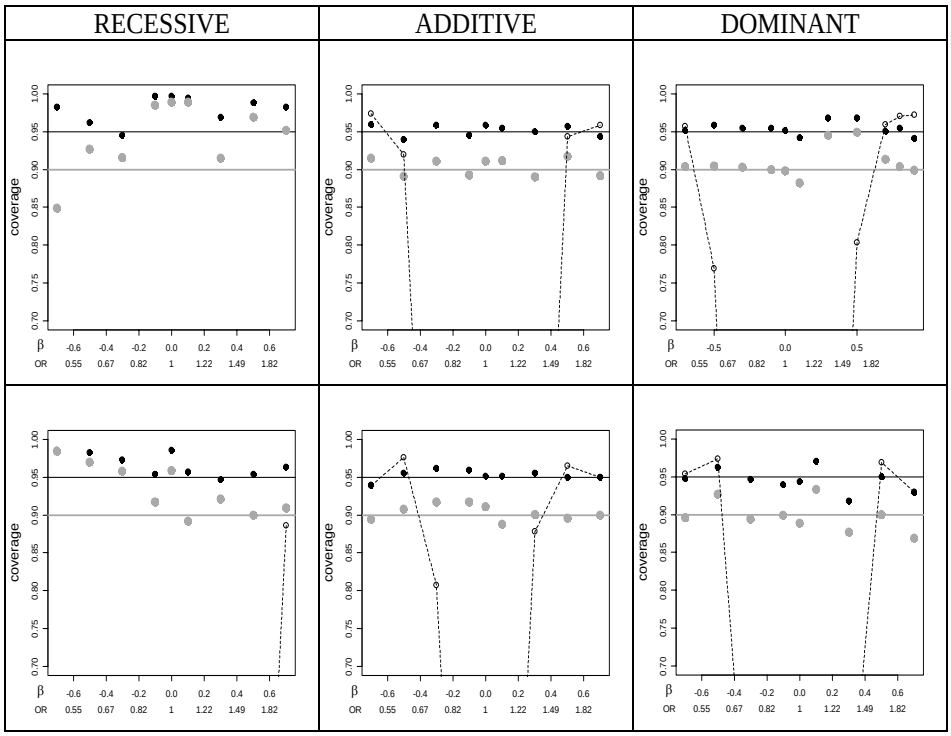


Figure 2.5: **Estimates of the CI coverage probability plotted against β for the three genetic models, MAF=0.25.**
Black dots correspond to 95% CIs, grey dots to 90% CIs. The dashed curves represent coverage of standard 95% CIs which do not acknowledge the significance selection. *Top row*: $n = 1000$ (500 cases and 500 controls). *Bottom row*: $n = 2000$ (1000 cases and 1000 controls). Coverage is close to nominal, except for regions of over-coverage in the recessive model due to small cell counts (note that the y-axis range begins at 0.7). For all models, the coverage will approach the nominal value as the sample size increases further.

2.3.3 Confidence coverage

Figure 2.5 presents the estimated coverage probabilities of 95% and 90% CIs with MAF=0.25 for the three models. The top row shows the results for $n = 1000$. The coverage is close to the nominal level for almost all the setups, except for conservativeness near $\beta = 0$ for the recessive model. The coverage of the naïve confidence intervals is also depicted in the Figure, dropping dramatically out of the axis range to zero coverage for β of small magnitude. For $n = 2000$, the coverage of the proposed procedure

improves further, with a region of modest over-coverage for recessive models. Results for other MAF values are similar, and are presented in Supplemental Figure 2.7.

2.3.4 Sample sizes, thresholds, and covariates

Our setup conditions represent a wide range of realistic scenarios, but cannot represent all situations and complicating factors. Fortunately, the large-sample behavior of the constructed approximate likelihood provides considerable robustness for our conclusions. Supplemental Figure 2.8 shows the results of increasing sample size for several realistic β values for the additive model when MAF=0.25. The bias and MSE for all the estimators are reduced as the sample size increases. For each sample size, the corrected estimators show superior bias and MSE compared to the naïve estimator.

In maximum likelihood settings, the distribution of the Wald test statistic is largely driven by $\beta/SE(\hat{\beta})$. This is also true for our conditional likelihood, because $\beta/SE(\hat{\beta})$ determines the non-centrality of the z-statistic. For a fixed ratio $n_{cases} : n_{controls}$, the standard error is proportional to $1/\sqrt{n}$. Thus, for the setups in Figure 2.3 and Supplemental Figure 2.6, a doubling of the sample size to $n = 2000$ (for example, and assuming cases and controls remain in the same ratio) would produce qualitatively similar results, with perhaps a slight improvement for the corrected estimates as the normality approximation improves. Moreover, we can make the results quantitatively comparable by appropriate rescaling. For example, for any value β for $n = 1000$, the comparable results for $n = 2000$ should correspond to $\beta' = \beta\sqrt{2}$. Supplemental Figure 2.9(a) demonstrates an empirical example of this effective rescaling equivalence for the additive model, MAF=0.25. Thus the conclusions from our simulations extend to larger sample sizes.

Similarly, variations on the threshold c do not have much impact. A value of $c = 5.5$ would be considered quite conservative for genome scans, corresponding to Bonferroni

control of family-wise error at 0.05 for 1.3 million SNPs. Empirical investigation requires many more simulations to achieve significance, but we find that the qualitative behavior of the estimators is unchanged (Supplemental Figure 2.9(b))

Finally, we simulated an example in which the additive model is fit (MAF=0.25), and the logistic regression includes an additional continuous covariate (distributed $N(0,1)$, one fitted regression coefficient) and a discrete covariate (distributed Binomial(2,0.05), two fitted coefficients). The covariates were independent of case-control status and the test-locus genotype. The Wald statistic is relatively insensitive to inclusion of these extra parameters, and the relative change in degrees of freedom quite minimal. Accordingly, the results for our corrected estimators are virtually unchanged compared to the model without covariates (Supplemental Figure 2.9(c) - only $\tilde{\beta}_1$ is shown). Covariate considerations are increasingly important in genome scans, for example to control for confounding population stratification.

2.3.5 Analyses of published datasets

Tables 2.1, 2.2, and 2.3 illustrate our re-analysis of three published genetic association studies: an association study with a modest number of SNPs (Table 2.1), as well as two GWAS (Tables 2.2 and 2.3), all of which had been analyzed using additive models. We begin with a brief description of the three studies, followed by our re-analysis. In Table 2.1, we present the analysis results of Yu et al. (2007) who re-examined the lymphoma dataset described in Wang et al. (2005). They used 48 SNPs and a p-value threshold of $0.1/48 \approx 0.002$. We report the standard OR results and the bootstrap bias-corrected estimates produced by Yu et al. (2007), as well as the estimates from a larger pooled analysis involving seven studies (Rothman et al. 2006). The SNPs rs1800629 and rs909253 were found to be significant, with ORs 1.54 and 1.40, respectively. In Table 2.2 we list four significant SNPs reported by Todd et al. (2007) resulting from

two Type 1 diabetes (T1D [MIM 222100]) GWAS studies, declaring SNPs as significant if they have p-values less than 5×10^{-7}. We also display the results from a larger case-control followup study conducted by Todd et al. (2007) to confirm their results. In Table 2.3 we report the results of a GWAS by Scott Scott et al. (2007), who performed numerous analyses of several Type 2 diabetes (T2D [MIM 125853]) datasets (FUSION, DGI, and WTCCC/UKT2D). We consider here only the SNPs reported by the T2D authors using the declared genome-wide significance threshold ($p < 5 \times 10^{-8}$) for the combined analysis of all studies.

Using only the published odds ratios, p-values and stated significance thresholds, we produced bias-corrected odds ratios for all of these studies. Our corrected β estimates are exponentiated to obtain odds ratios: for example, $\tilde{OR}_1 = \exp(\tilde{\beta}_1)$. For the two lymphoma SNPs (Table 2.1), the p-values are slightly above the threshold, and our bias-corrected estimates shrink the naïve OR estimates markedly. Our estimated values match well with the bootstrap-corrected values obtained by Yu et al. (2007), as well as the pooled analysis results from Rothman et al. (2006).

For the four T1D SNPs (Table 2.2), our analysis results in noticeably less extreme OR estimates than that reported by Todd et al. (2007). The corrected ORs and CIs for the most extreme SNP, rs17696736, are only slightly changed from the published estimated of 1.37 because the result is so extreme ($p = 7.27 \times 10^{-14}$). However, the followup study obtained a considerably lower value (OR=1.16), with the 95% CI not overlapping the earlier estimates, suggesting possible heterogeneity in population sampling. For the two least significant T1D SNPs among those considered, the corrected ORs show a more substantial change. It is worth noting that the OR estimate corresponding to the SNP rs12708716 was shrunk from 0.77 to about 0.82 by our, methods while the estimated OR from the follow-up was 0.83. We also note that for the four significant T1D SNPs, as well as an additional three SNPs approaching significance

(Table 1 of Todd et al. (2007)), the followup study always gave a less extreme OR estimate than the initial studies. This result is strong empirical evidence for significance bias, and that corrected OR approaches are needed.

Table 2.3 gives the results for the combined T2D studies. All of the p-values are considerably beyond the significance threshold, and so the corrected estimates are nearly unchanged from the original estimates. This phenomenon is hopeful, in the sense that with very large studies OR estimates can be attained that will not be shrunk to irrelevance by corrected OR estimates.

Table 2.1: Original vs. corrected odds ratio estimates for published genetic association study I: Association study of lymphoma, Wang et al. (2005) (318 cases and 766 controls)

[a]Standard OR values as reported
[b]Bootstrap correction reported in Ref. 15
[c]Correction method proposed in this manuscript
[d]Replication or other follow-up result for the SNP

SNP	P-value	Reported OR[a], (95% CI)	Bootstrap[b] estimates	Bias-corrected estimates			Bias[b]-corrected (95% CI)	Follow-up[c] OR,
				$\tilde{OR}_1$	$\tilde{OR}_2$	$\tilde{OR}_3$		
rs1800629	$5.7x10^{-4}$	1.54	1.29	1.14	1.28	1.21	(0.96,1.91)	1.29
rs909253	$7.4x10^{-4}$	1.4	1.18	1.08	1.21	1.14	(0.96,1.65)	1.16

Table 2.2: Original vs. corrected odds ratio estimates for published genetic association study II: GWAS of T1D, Todd et al. (2007) (2000 cases and 3000 controls)

[a]Standard OR values as reported
[b]Correction method proposed in this manuscript
[c]Replication or other follow-up result for the SNP

SNP	P-value	Reposted OR[a], (95% CI)	Bias-corrected estimates			Bias[b]-corrected (95% CI)	Follow-up[c] OR, (95% CI)
			$\tilde{OR}_1$	$\tilde{OR}_2$	$\tilde{OR}_3$		
rs17696736	7.27x10-14	1.37 (1.27,1.49)	1.37	1.36	1.37	(1.25,1.49)	1.16 (1.09,1.23)
rs2292239	1.49x10-9	1.3 (1.20,1.42)	1.27	1.24	1.26	(1.10,1.41)	1.28 (1.20,1.36)
rs12708716	1.28x10-8	0.77 (0.70,0.84)	0.81	0.83	0.82	(0.71,0.99)	0.83 (0.78,0.89)
rs2542151	8.4x10-8	1.33 (1.20,1.49)	1.15	1.17	1.16	(0.99,1.45)	1.29 (1.19.1.40)

Table 2.3: Original vs. corrected odds ratio estimates for published genetic association study III: GWAS of T2D, Scott et al. (2007) (9521 cases and 12183 controls)

[a]Standard OR values as reported
[b]Correction method proposed in this manuscript

SNP	P-value	Reported OR[a], (95% CI)	Bias-corrected estimates			Bias[b]-corrected (95% CI)
			$\tilde{OR}_1$	$\tilde{OR}_2$	$\tilde{OR}_3$	
rs7903146	$1.0x10^{-48}$	1.37 (1.31,1.43)	1.37	1.37	1.37	(1.31,1.43)
rs4402960	$8.9x10^{-16}$	1.14 (1.11,1.18)	1.14	1.14	1.14	(1.10,1.18)
rs10811661	$7.8x10^{-15}$	1.2 (1.14,1.25)	1.2	1.2	1.2	(1.14,1.26)
rs8050136	$1.3x10^{-12}$	1.17 (1.12,1.22)	1.17	1.16	1.16	(1.10,1.22)
rs7754840	$4.1x10^{-11}$	1.12 (1.08,1.16)	1.11	1.1	1.11	(1.05,1.16)
rs5219	$6.7x10^{-11}$	1.14 (1.10,1.19)	1.13	1.12	1.12	(1.06,1.19)
rs1111875	$5.7x10^{-10}$	1.13 (1.09,1.17)	1.11	1.1	1.1	(1.02,1.17)

2.4 Discussion

We have presented an approach that greatly reduces significance bias for odds ratios in genome association scans, and is much simpler than competing approaches. We favor the use of $\tilde{\beta}_3$ as a general-purpose estimator with fairly uniform MSE as a function of β. However, all of the three corrected estimators have greatly superior performance to the naïve estimator. Although developed for case-control applications, our methodology is an effective blueprint to perform inference whenever a Wald-like statistic has been used to declare significance. Thus the general approach can be used in numerous other settings, including regression-based quantitative trait association analyses. Our results are qualitatively similar to those of other investigators (Yu et al. 2007; Zöllner and Pritchard 2007) (e.g., see bias curves similar to ours in Figure 2 of Zöllner and Pritchard (2007)). Additional comparisons to these approaches should be performed in future work, although comparison is complicated by differing genetic models. To our knowledge, our approach is the only method that can perform bias correction based only on published summary tables.

The widespread application of conditional likelihood estimators in genome scans will no doubt be discouraging to genetic investigators, who may expend considerable time and expense only to find that a significant SNP is estimated to have a very weak effect. Nonetheless, we view this process as healthy and necessary for the genetics community, and in particular to tamp down expectations that significant findings will be easily replicated. The use of our estimators may also have an additional benefit of

discouraging excessive massaging of data and trying various test procedures to achieve genome-wide significance. If a SNP suddenly becomes significant after numerous data manipulation procedures have been applied, its z-statistic is likely to be only slightly above the threshold c. Thus, as we observed in the μ version of the problem, the conditional likelihood estimator will be dramatically shrunk towards the null. Thus the estimated SNP effect size will be very modest, as is appropriate here for a likely spurious finding.

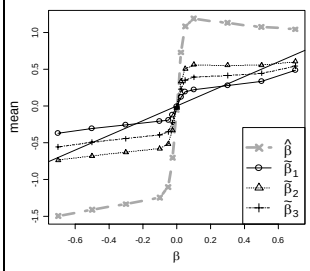
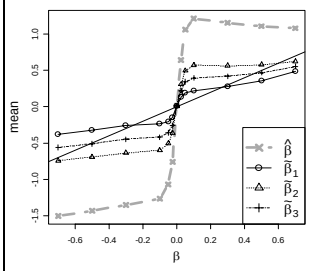
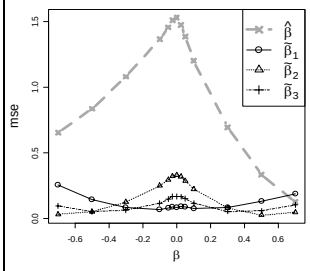
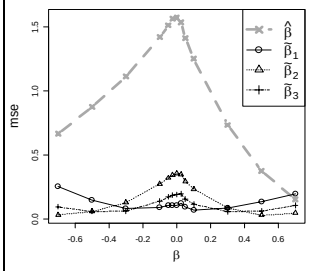
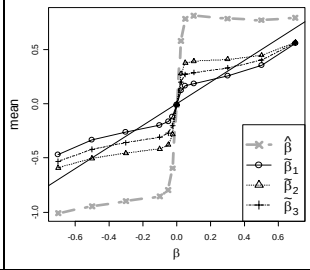
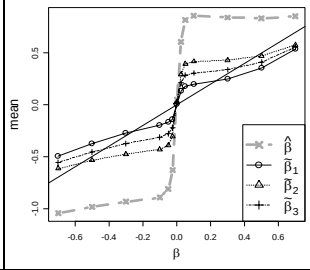
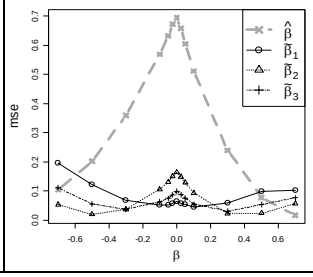
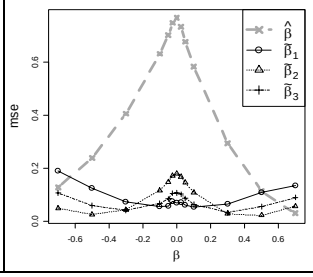
Our current approach does not explicitly consider multi-stage or other sequential designs, in which SNPs meeting a loose standard of significance are used for further testing in a follow-up sample. However, for multistage designs in which almost all SNPs that will eventually be declared significant are carried forward to later stages, the approach may be used directly. Also, our results technically hold for a SNP randomly selected from those achieving the significance threshold, and thus an additional bias may be anticipated for the most highly significant SNPs among a collection of significant SNPs. Although we believe this second source of bias is much less than that produced by significance selection, it is the subject of continuing investigation.

Our rejection-sampling scheme was feasible, but required a massive number of simulations to provide accurate results. Future work in this area may benefit from the practical development of importance sampling or related computational techniques to provide flexible and accurate simulations conditioned on significance.

2.5 Web Resources

The URLs for data presented herein are as follows: Online Mendelian Inheritance in Man (OMIM), `http://www.ncbi.nlm.nih.gov/Omim/`. R code and a simple Excel calculator to perform our method are available at `www.bios.unc.edu/~fwright/genomebias`.

2.6 Supplemental Figures

Model / MAF		RECESSIVE	ADDITIVE	DOMINANT
0.05	MEAN			
	MSE			
0.1	MEAN			
	MSE			

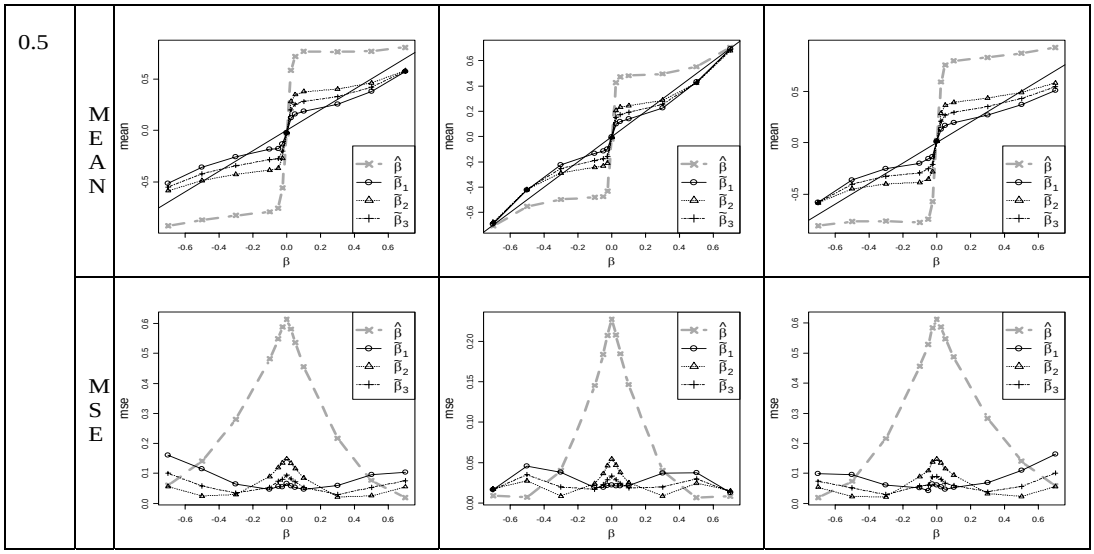
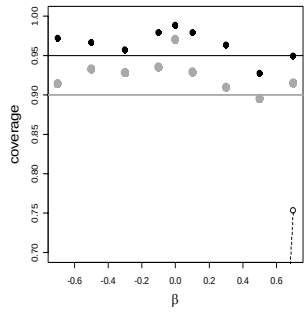
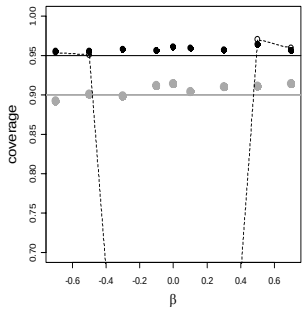
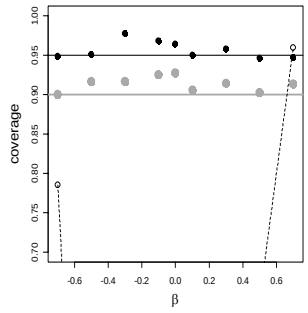
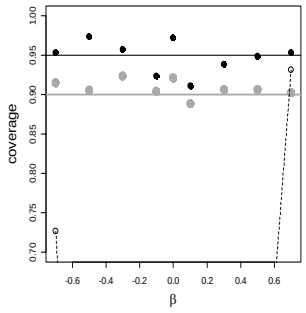
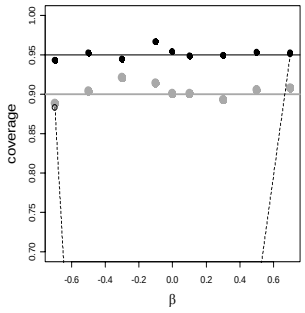
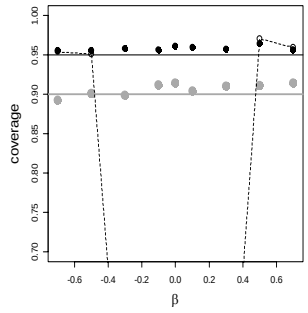
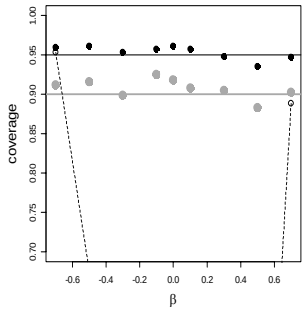


Figure 2.6: **Expected values and mean squared errors for the estimators for the three models.**
MAF values are 0.05, 0.1, and 0.5 for additive and dominant models, and MAF=0.5 for recessive models.

Model / MAF	RECESSIVE	ADDITIVE	DOMINANT
0.05			
0.1			
0.5			

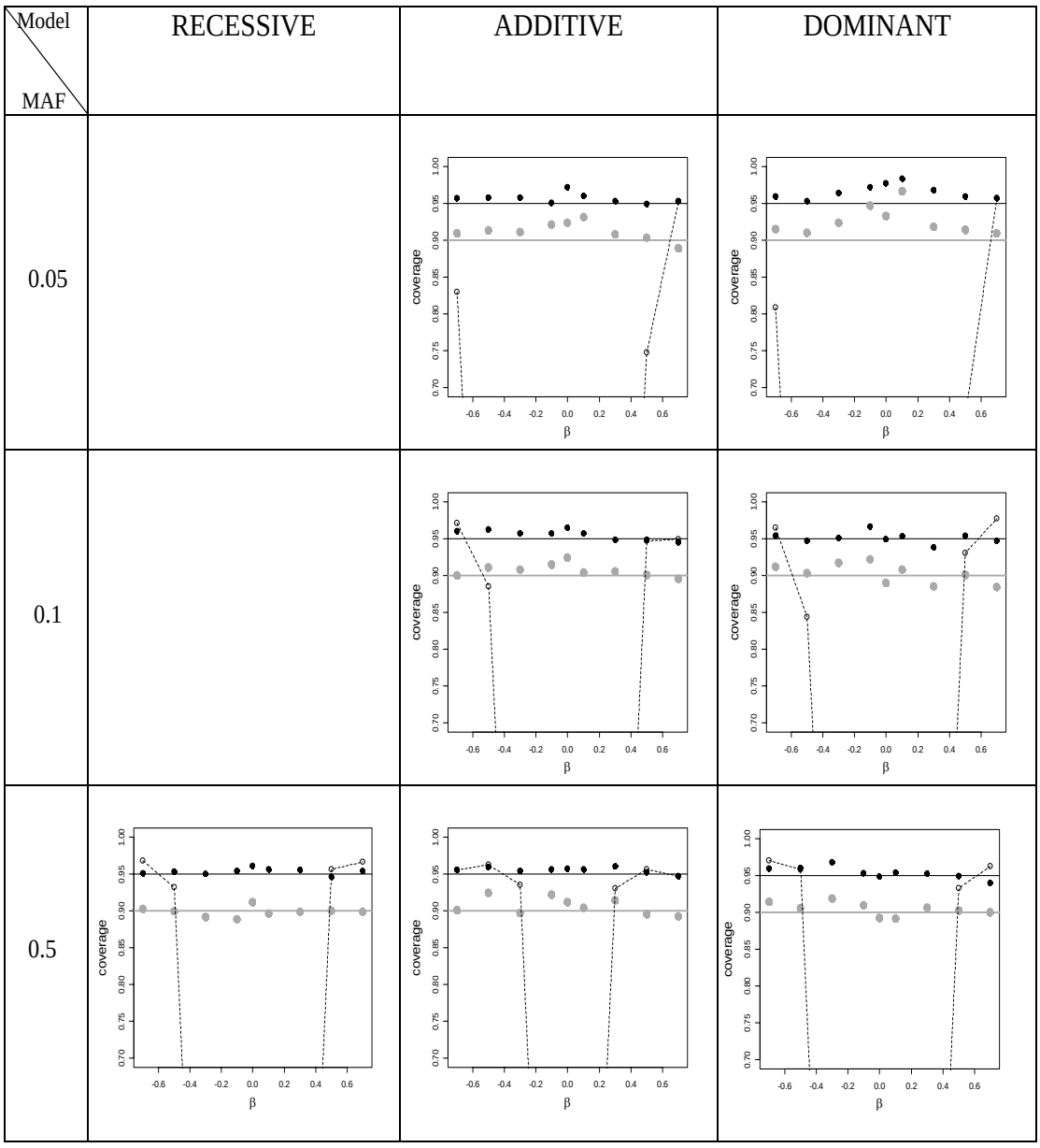


Figure 2.7: **Estimates of the CI coverage probability plotted against β for the three genetic models.**
Various MAF values are shown (for recessive models, only MAF=0.5 is depicted). Black dots correspond to 95% CIs, grey dots to 90% CIs. The dashed curves represent coverage of standard 95% CIs which do not acknowledge the significance selection. First page, $n = 1000$. Second page, $n = 2000$.

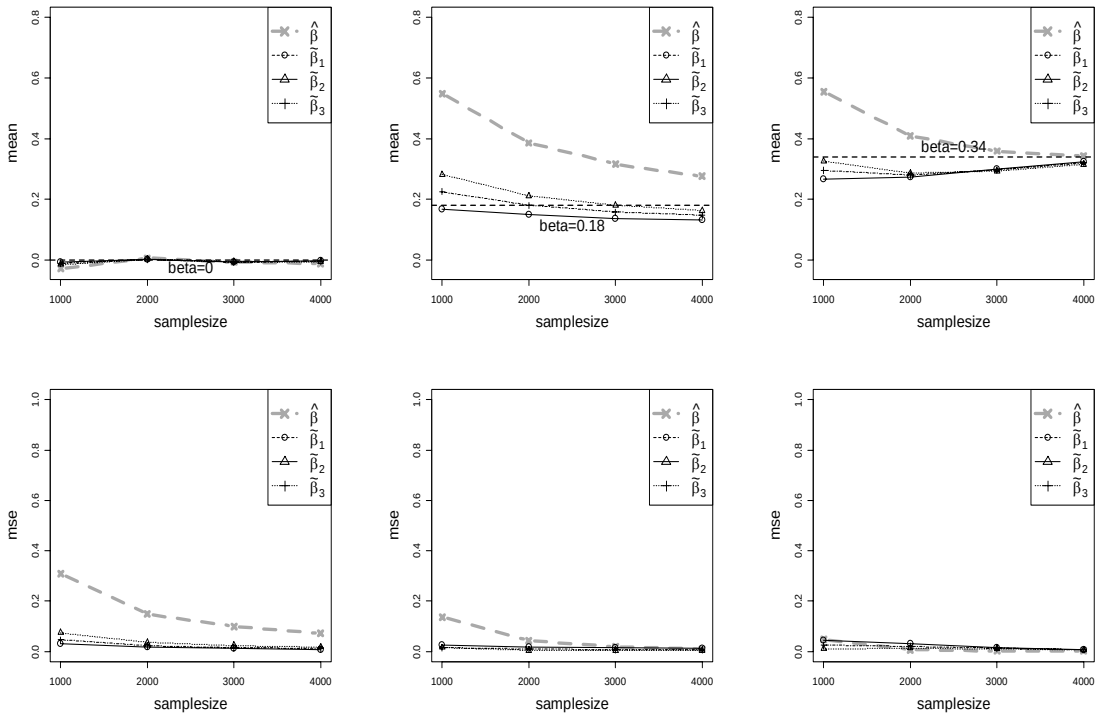


Figure 2.8: **Expected values and the mean squared errors of the estimators for the additive model with MAF=0.25.**
The results are plotted against sample size, for $\beta = 0$, 0.18, and 0.34, corresponding to OR values 1.0,1.2, and 1.4.

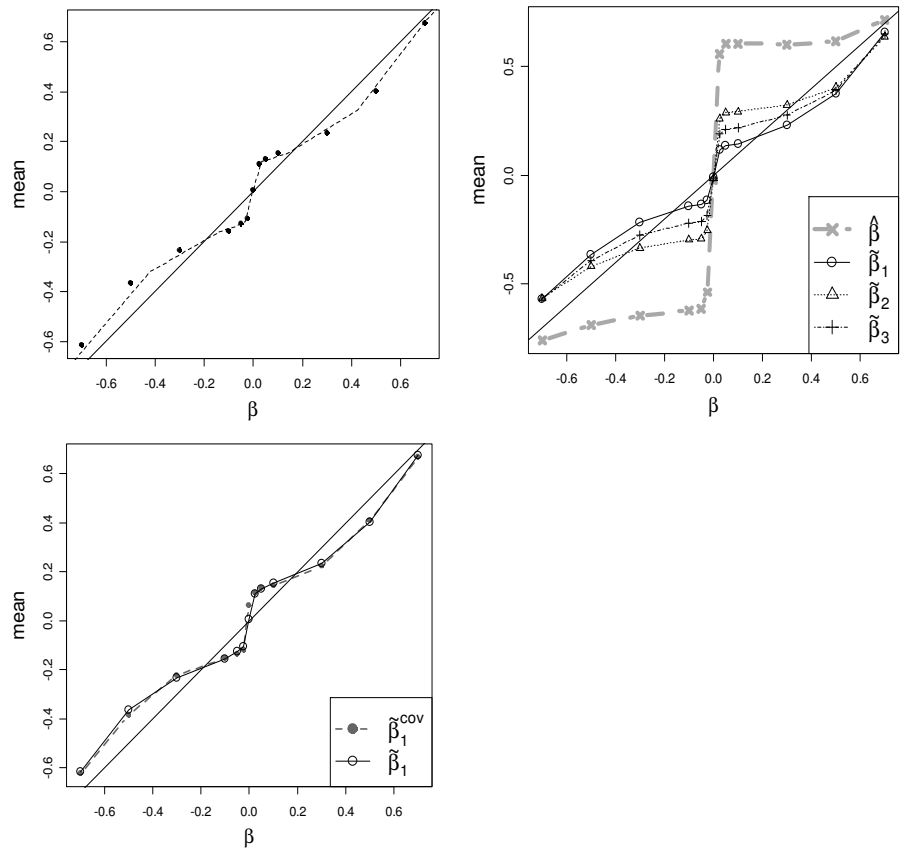


Figure 2.9: **Properties of the corrected estimators extended to additional settings.**
Throughout this figure we use the additive model, MAF=0.25, and $n = 1000$ except where noted. (a) Expectations of $\tilde{\beta}_1$ vs. β (plotted points) for $n = 1000$, overlaid with results for the same estimator vs. $\beta\sqrt{2}$ (dashed line) for $n = 2000$. The close correspondence is a consequence of the unifying treatment in terms of μ. (b) Expectations of the estimators for $c = 5.5$ show that the qualitative behavior is similar to the behavior for $c = 5.0$. (c) Inclusion of both a discrete (2 degree-of-freedom) and a continuous (1 d.f.) covariate in the logistic regression modeling has essentially no effect on the behavior of our estimators.

Chapter 3

Analysis of Secondary Phenotypes in Case-control Studies

3.1 Introduction

Genome-wide association studies (GWAS) require a lot of investment in terms of effort, time, and money. So it is only natural for investigators to want to make maximum use of the data collected in the process of conducting a GWAS. Usually in GWAS the primary research question is identification of SNPs influencing susceptibility to a trait that is of interest. In the process, information on a host of other phenotypes, typically correlated with the primary phenotype, is collected. Subsequent to the initial detection of SNPs significant for the disease phenotype, researchers often want to analyze the secondary phenotypes for efficient use of data and also to supplement the primary analysis (Frayling et al. 2007; Weedon et al. 2008; Gudbjartsson et al. 2008; Lettre et al. 2008; Weedon et al. 2007; Sanna et al. 2008; Loos et al. 2008). In fact, data for the

analysis of a phenotype may come from several GWAS, for which data were originally collected with different research objectives in mind. Because of its efficient design, data for GWAS are usually case-control sampled. In such cases the data cannot be considered as a random sample from the population and ignoring the the ascertainment on the basis of the primary phenotype can produce biased estimate of the association between a SNP and a secondary phenotype (Nagelkerke et al. 1995). For the appropriate analysis of secondary phenotype we have to take into consideration the biased sampling.

There are several options for handling the secondary analysis for case-control data (Jiang et al. 2006), such as a) ignore the sampling mechanism, b) analyze controls only, c) analyze cases only, d) include the disease status as an explanatory variable, e) apply weighted approach, or f) implement maximum likelihood methods. The first method of analyzing the combined sample of cases and controls without accounting for the ascertainment can lead to a severely biased estimate of the risk effect for the secondary phenotype. For case-control data, prospective analysis ignoring the sampling scheme yields valid estimate of the odds ratio for disease risk, but the same reasoning does not hold for the secondary phenotype generally (Nagelkerke et al. 1995). Only under very restrictive conditions, prospective analysis will give unbiased estimate of the population measure of association between the SNP and the secondary phenotype. Nagelkerke et al. (1995) suggest using only the controls for the secondary analysis. This method is approximately valid if the disease is rare. If we analyze only the cases or the controls, then we loose a lot of information. For methods b), c), and d), adjusting for the disease status eliminates the possibility of bias induced by the case-control

sampling mechanism, but the estimates that we derive from each of these methods may be estimating quantities very different from the one that we are interested in. For example, including the disease phenotype as a covariate is a convenient way of incorporating the sampling mechanism in the analysis but not necessarily the true model that we believe in. Monsees et al. (2009) have recently discussed in details the situations under which the naïve analysis that ignores the ascertainment or the analysis that includes disease status as a covariate are valid.

The standard survey approach or Horvitz-Thompson approach uses weights inversely proportional to the selection probabilities (Jiang et al. 2006; Scott and Wild 2002). Richardson et al. (2007) describe it as a stratum-weighted logistic regression for a binary secondary phenotype and compare its merits with the usual practice of adjusting for the disease status by including it in the regression as a covariate. It is, of course, necessary to have knowledge of the sampling fractions for cases and controls, which we do for nested case-control studies. But for population-based case-control studies, this information is not readily available. Rather than focus on the absolute sampling fractions, we can try to estimate from external information, such as the prevalence, the ratio of the the sampling fractions which would be sufficient for the purpose of weighted regression.

An alternative approach to secondary analysis is to use the retrospective likelihood (Jiang et al. 2006; Scott and Wild 1997, 2001a, 1991; Lee et al. 1997; Lin and Zeng 2008) which conditions on the disease status. To work with the retrospective likelihood

one needs to model the joint distribution of the primary and the secondary phenotypes given the genotype and other covariates. The joint distribution can be factorized further into the marginal distribution of the secondary phenotype and the conditional distribution of the primary phenotype given the secondary phenotype. Our interest lies in estimating the parameters of the marginal distribution of the secondary phenotype given the genotype and the covariates. As Jiang et al. (2006) have described, we can either treat the conditional distribution of the primary phenotype given the secondary phenotype non-parametrically or we can model it as a logistic regression. Lin and Zeng (2008) have developed likelihood methods for analysis of both binary and continuous secondary traits where they have modeled the the conditional distribution of the primary phenotype given the secondary phenotype as a logistic regression. An alternative way to specify the joint model is to parametrically model the marginals of the primary and the secondary phenotypes and also build a parametric model for their association given the genotype and the covariates. The distribution of the genotype and the covariates is a nuisance parameter in the retrospective likelihood. It is difficult to parameterize the covariate distribution and is usually treated non-parametrically.

We describe a method to analyze secondary phenotypes, both binary and continuous, where we model the joint distribution of the phenotypes such that the marginals for each phenotype respect the commonly used models for analyzing them separately. Or in other words, we specify the joint distribution such that the marginal distribution for the disease phenotype is always logistic and that for the secondary phenotype is logistic or linear depending on whether it is binary or continuous respectively. We have

allowed for inclusion of covariates in our models and have performed extensive simulations to compare the performance of our proposed approach with the performances of the naïve method of prospectively analyzing the combined sample of cases and controls ignoring the biased sampling, case-only analysis, controls-only analysis, and the weighted method.

3.2 Methods

Let D denote the disease phenotype (0=control, 1=case), Y the secondary phenotype, and G the SNP genotype. Let $\mathbf{Z}$ denote the vector of covariates in the model, such as gender, age, or environmental factors. The secondary phenotype, as well as the covariates, may be either dichotomous or continuous. The data were sampled from the population with respect to the disease status variable D. We are interested in the association between the secondary phenotype Y and the SNP genotype G, adjusting for the covariates $\mathbf{Z}$. The appropriate likelihood that takes into account the case-control sampling mechanism is the retrospective likelihood $P(Y, G, Z|d)$. For case-control sampled data $(d_i, y_i, g_i, \mathbf{z}_i), i = 1, 2, \ldots, n$, the retrospective log-likelihood is

$$
\begin{aligned}
l &= \log L \\
&= \sum_{i=1}^{n} \log P(Y = y_i, G = g_i, \mathbf{Z} = \mathbf{z}_i | d_i) \\
&= \sum_{i=1}^{n} \log P(D = d_i, Y = y_i | g_i, \mathbf{z}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \log P(G = g_i, \mathbf{Z} = \mathbf{z}_i) \\
&\qquad - \sum_{i=1}^{n} \log P(D = d_i) \; .
\end{aligned}
$$

For prospectively collected data, we can make inference about $\boldsymbol{\theta}$ from $\sum_{i=1}^{n} \log P(D = d_i, Y = y_i | g_i, \mathbf{z}_i; \boldsymbol{\theta})$ and ignore $P(G = g, \mathbf{Z} = \mathbf{z})$. But, for case-control data, we cannot ignore $P(G = g, \mathbf{Z} = \mathbf{z})$ since it is intertwined with $\boldsymbol{\theta}$ in

$$
P(D = d) = \sum_{y} \sum_{g} \sum_{\mathbf{z}} P(D = d, Y = y | g, \mathbf{z}; \boldsymbol{\theta}) P(G = g, \mathbf{Z} = \mathbf{z}) \; .
$$

The retrospective likelihood, therefore, is a function of $\boldsymbol{\theta}$, the parameter of interest, and $P(G = g, \mathbf{Z} = \mathbf{z})$, the nuisance parameter. We assume that disease prevalence is known approximately and incorporate that information in the likelihood. Under known prevalence, say $P(D = 1) = \Pi$, we maximize

$$
l = \sum_{i=1}^{n} \log P(D = d_i, Y = y_i | g_i, \mathbf{z}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \log P(G = g_i, \mathbf{Z} = \mathbf{z}_i)
$$

with respect to $(\boldsymbol{\theta}, P(G = g, \mathbf{Z} = \mathbf{z}))'$ subject to the constraint

$$\sum_{y} \sum_{g} \sum_{\mathbf{z}} P(D = 1, Y = y|g, \mathbf{z}; \boldsymbol{\theta}) P(G = g, \mathbf{Z} = \mathbf{z}) = \Pi \, . \tag{3.1}$$

We assume that G and $\mathbf{Z}$ are independent in the population. Since G is discrete and can take at most three values, we treat the probability distribution of G, $p(g)$, as a nuisance parameter and maximize the likelihood with respect to it, subject to the constraint $\sum_g p(g) = 1$. It is generally difficult and unreasonable to parameterize the covariate distribution $P(\mathbf{Z} = \mathbf{z})$. If all the covariates are categorical and there are tractable number of combinations of the levels of the covariates, then we can treat them as nuisance parameters and maximize the likelihood with respect to them. For illustration purpose, let us consider the situation where the genotype is coded as 0 or 1 with $P(G = 1) = \delta$ and we have a single binary covariate, Z, with probability of success ψ. Then we maximize

$$\begin{aligned} l(\boldsymbol{\theta}, \delta, \psi) &= \sum_{i=1}^{n} \log P(D = d_i, Y = y_i|g_i, \mathbf{z}_i; \boldsymbol{\theta}) + n_{G1} \log \delta \\ &\quad + (1 - n_{G1}) \log (1 - \delta) + n_{Z1} \log \psi + (1 - n_{Z1}) \log (1 - \psi) \, , \end{aligned} \tag{3.2}$$

$$\text{where} \quad n_{G1} = \sum_{i=1}^{n} g_i \quad \text{and} \quad n_{Z1} = \sum_{i=1}^{n} z_i \, ,$$

with respect to $(\boldsymbol{\theta}, \delta, \psi)'$ subject to the constraint (3.1) to get the maximum likelihood estimate of $\boldsymbol{\theta}$. The MLE is consistent and asymptotically normal and the covariance

matrix of the MLE can be consistently estimated by the inverse of the observed information matrix. This method becomes infeasible very quickly as the number of covariates increases and it does not allow for continuous covariates. For continuous covariates we can assume the profile likelihood approach (Lee et al. 1997; Wild 1991; Lin and Zeng 2008; Scott and Wild 2001b). Suppose Z is now a continuous covariate. We have n parameters $P(Z = z_i) = p_i$, $i = 1, \ldots, n$ describing the distribution of Z. To get the maximum likelihood estimate of $\boldsymbol{\theta}$ we need to maximize

$$\begin{aligned} l(\boldsymbol{\theta}, \delta, p_1, \ldots, p_n) &= \sum_{i=1}^{n} \log P(D = d_i, Y = y_i | g_i, \mathbf{z}_i; \boldsymbol{\theta}) \\ &\quad + \sum_{i=1}^{n} \log P(G = g_i) + \sum_{i=1}^{n} \log p_i \end{aligned}$$

with respect to $(\boldsymbol{\theta}, \delta, p_1, \ldots, p_n)'$ subject to the constraints $\sum_{i=1}^{n} p_i = 1$ and

$$\sum_{y} \sum_{g=0}^{1} \sum_{i=1}^{n} P(D = 1, Y = y | g, z_i; \boldsymbol{\theta}) P(G = g) p_i = \Pi \, .$$

Using Lagrange multipliers, we maximize

$$\begin{aligned} & l(\boldsymbol{\theta}, \delta, p_i, \ldots, p_n) \\ = \; & \sum_{i=1}^{n} \log P(D = d_i, Y = y_i | g_i, \mathbf{z}_i; \boldsymbol{\theta}) \\ & + \sum_{i=1}^{n} \log P(G = g_i) + \sum_{i=1}^{n} \log p_i + \lambda_1 (\sum_{i=1}^{n} p_i - 1) \\ & + \lambda_2 \left(\sum_{y} \sum_{g=0}^{1} \sum_{i=1}^{n} P(D = 1, Y = y | g, z_i; \boldsymbol{\theta}) P(G = g) p_i - \Pi \right) \end{aligned} \tag{3.3}$$

with respect to $(\boldsymbol{\theta}, \delta, p_i, i = 1, \ldots, n)'$. λ_1 and λ_2 are determined using the constraints. Maximizing(3.3) with respect to p_i we get,

$$\frac{1}{p_i} - \lambda_1 - \lambda_2 \sum_y \sum_{g=0}^{1} P(D = 1, Y = y|g, z_i; \boldsymbol{\theta}) P(G = g) = 0 \; . \tag{3.4}$$

Multiplying the above equation by p_i on both sides and then taking a sum over i we get,

$$\lambda_1 = n - \lambda_2 \Pi \; .$$

Substituting λ_1 in (3.4), we have

$$p_i = \left(n - \lambda_2 \Pi + \lambda_2 \sum_y \sum_{g=0}^{1} P(D = 1, Y = y|g, z_i; \boldsymbol{\theta}) P(G = g) \right)^{-1} \; .$$

Thus, the profile log-likelihood for $(\boldsymbol{\theta}, \delta)'$ is

$$\begin{aligned} l_{profile}(\boldsymbol{\theta}, \delta) \quad &= \sum_{i=1}^{n} \log P(D = d_i, Y = y_i|g_i, \mathbf{z}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \log P(G = g_i) \\ &\quad - \sum_{i=1}^{n} \log \left(n - \lambda_2 \Pi + \lambda_2 \sum_y \sum_{g=0}^{1} P(D = 1, Y = y|g, z_i; \boldsymbol{\theta}) P(G = g) \right) \; , \end{aligned}$$

where λ_2 is determined by

$$\sum_{i=1}^{n} \left(n - \lambda_2 \Pi + \lambda_2 \sum_y \sum_{g=0}^{1} P(D = 1, Y = y|g, z_i; \boldsymbol{\theta}) P(G = g) \right)^{-1} = 1 \; .$$

The estimates from the profile likelihood are consistent and asymptotically normal and the covariance matrix can be consistently estimated by the inverse of the observed information matrix obtained from the profile likelihood. We propose an alternative approach that has been used in various contexts: the pseudo likelihood idea put forward by Gong and Samaniego (1981) for parametric inference and later extended by Hu and Lawless (1997) to a semiparametric setting. The idea involves maximizing the pseudo likelihood $L_p\left(\boldsymbol{\theta}, p(g), \hat{P}\left(\mathbf{Z}=\mathbf{z}\right)\right)$, where $\hat{P}(\mathbf{Z}=\mathbf{z})$ is a nonparametric estimate of $P(\mathbf{Z}=\mathbf{z})$. Under known prevalence the pseudo log-likelihood l_p is

$$\begin{aligned} l_p &= \log L_p\left(\boldsymbol{\theta}, p(g), \hat{P}\left(\mathbf{Z}=\mathbf{z}\right)\right) \\ &= \sum_{i=1}^{n} \log P(D=d_i, Y=y_i|g_i, \mathbf{z}_i; \boldsymbol{\theta}) + \sum_{i=1}^{n} \log P(G=g_i) \end{aligned}$$

along with the constraint

$$\sum_{y}\sum_{g}\sum_{\mathbf{z}} P(D=1, Y=y|g, \mathbf{z}; \boldsymbol{\theta}) P(G=g) \hat{P}(\mathbf{Z}=\mathbf{z}) = \Pi \, .$$

Noting that

$$P\left(\mathbf{Z}=\mathbf{z}\right) = P\left(\mathbf{Z}=\mathbf{z}|D=1\right) P(D=1) + P\left(\mathbf{Z}=\mathbf{z}|D=0\right) P(D=0) \, ,$$

we estimate $P\left(\mathbf{Z}=\mathbf{z}\right)$ by

$$\hat{P}\left(\mathbf{Z}=\mathbf{z}\right) = \hat{P}\left(\mathbf{Z}=\mathbf{z}|D=1\right) \hat{P}(D=1) + \hat{P}\left(\mathbf{Z}=\mathbf{z}|D=0\right) \left(1-\hat{P}(D=1)\right) \, ,$$

where $\hat{P}(\mathbf{Z} = \mathbf{z}|D = i)$ $, i = 0, 1$ are valid estimates being the empirical cumulative distribution functions based on the controls and the cases respectively. For $\hat{P}(D = 1)$ we have to depend on external information. The pseudo maximum likelihood estimate(MLE) is then obtained by maximizing the pseudo log-likelihood, l_p,with respect to the parameter of interest, $\boldsymbol{\theta}$, and the nuisance parameter, $p(g)$. A rigorous development of the the asymptotic properties of the pseudo MLE is complicated. Hu and Lawless (1997) discuss the asymptotics for pseudo likelihood methods in the context of response-related missing covariates. Following the same lines we plan to lay down the details of the asymptotic theory for pseudo likelihood estimation in our situation. We now discuss how to parameterize the joint distribution $(D, Y|g, \mathbf{z})$ for binary and continuous secondary phenotypes.

3.2.1 Binary secondary phenotype

There are several ways to parameterize the bivariate distribution $(D, Y|g, \mathbf{z})$. For dichotomous Y we are interested in models for which the marginal distributions of D and Y given g and $\mathbf{z}$ are both logistic. The bivariate logistic model, considered by Palmgren (1989), is one such model that has been used before in this context (Jiang et al. 2006; Lee et al. 1997) and is conceptually very simple. It is based on the fact that the joint distribution of two binary variables can be specified in terms of their marginal probabilities and their odds ratio. Thus, for a randomly sampled individual in the population we specify the joint distribution of D and Y given g and $\mathbf{z}$ as:

$$logit\ P(D = 1 \mid g, \mathbf{z}) = \alpha_1 + \beta_1 g + \boldsymbol{\gamma_1}'\mathbf{z}$$

$$logit\ P(Y = 1 \mid g, \mathbf{z}) = \alpha_2 + \beta_2 g + \boldsymbol{\gamma_2}'\mathbf{z}$$

$$\log OR(D, Y \mid g) = \frac{P(D = 1, Y = 1 \mid g, \mathbf{z})P(D = 0, Y = 0 \mid g, \mathbf{z})}{P(D = 1, Y = 0 \mid g, \mathbf{z})P(D = 0, Y = 1 \mid g, \mathbf{z})} = \alpha_3 + \beta_3 g\ .$$

The pseudo log-likelihood is, therefore, a function of $\boldsymbol{\theta} = (\alpha_1, \beta_1, \boldsymbol{\gamma_1}, \alpha_2, \beta_2, \boldsymbol{\gamma_2}, \alpha_3, \beta_3)'$, the parameter of interest, and $p(g)$, the nuisance parameter. With fixed disease prevalence, the identity

$$P(D = 1) = \sum_{\mathbf{z}} \sum_{g} \frac{\exp(\alpha_1 + \beta_1 g + \boldsymbol{\gamma_1}'\mathbf{z})}{1 + \exp(\alpha_1 + \beta_1 g + \boldsymbol{\gamma_1}'\mathbf{z})} p(g)\hat{P}(\mathbf{Z} = \mathbf{z}) \tag{3.5}$$

is used to compute α_1 given $(\beta_1, \boldsymbol{\gamma_1}, \alpha_2, \beta_2, \boldsymbol{\gamma_2}, \alpha_3, \beta_3)'$ and $p(g)$. Thus, the retrospective pseudo log-likelihood, under known prevalence, is

$$l_p = \sum_{i=1}^{n} \log P(D = d_i, Y = y_i | g_i, \mathbf{z}_i) + \sum_{i=1}^{n} \log P(G = g_i)\ ,$$

which is a function of $\boldsymbol{\theta} = (\beta_1, \boldsymbol{\gamma_1}, \alpha_2, \beta_2, \boldsymbol{\gamma_2}, \alpha_3, \beta_3, \boldsymbol{\gamma_3})'$, and $p(g)$. We obtain the pseudo MLE of $\boldsymbol{\theta}$ by maximizing l_p with respect to $(\boldsymbol{\theta}, p(g))'$. We can write down the joint probabilities $\pi_{ij} = P(D = i, Y = j | g, \mathbf{z})\ , i, j = 0, 1$ in terms of the marginal

probabilities $\pi_{i.} = P(D = i|g, \mathbf{z})$ and $\pi_{.j} = P(Y = j|g, \mathbf{z})$, and the odds ratio $\psi = \frac{\pi_{11}\pi_{00}}{\pi_{10}\pi_{01}}$,

$$\pi_{11} = \begin{cases} \frac{1}{2}(\psi - 1)^{-1}\left\{a - \sqrt{(a^2 + b)}\right\} & , \quad \psi \neq 1 \\ \pi_{1.}\pi_{.1} & , \quad \psi = 1 \ , \end{cases}$$

where $a = 1 + (\pi_{1.} + \pi_{.1})(\psi - 1)$ and $b = -4\psi(\psi - 1)\pi_{1.}\pi_{.1}$. The rest of the $\pi_{ij}s$ can be derived from π_{11} and the marginals.

3.2.2 Continuous secondary phenotype

For continuous Y, we consider joint models for $(D, Y|g, \mathbf{z})$ such that the marginal distribution of D given g and $\mathbf{z}$ follows a logistic regression model and the marginal distribution of Y given g and $\mathbf{z}$ is normal, that is,

$$logit\ P(D = 1 \mid g, \mathbf{z}) = \alpha_1 + \beta_1 g + \boldsymbol{\gamma_1}'\mathbf{z}$$

$$\text{and} \quad Y \mid g, \mathbf{z} \sim N\left(\alpha_2 + \beta_2 g + \boldsymbol{\gamma_2}'\mathbf{z}, \sigma_2^2\right) \ .$$

We introduce two levels of latency to come up with a joint model which satisfies the above conditions. We assume that the disease status variable, D, is derived from thresholding a latent continuous variable, U, whose marginal density given g and $\mathbf{z}$ is logistic with location parameter $\mu_1 = \alpha_1 + \beta_1 g + \boldsymbol{\gamma_1}'\mathbf{z}$ and scale parameter 1, that is,

$$p_{U|G,\mathbf{z}}(u \mid g, \mathbf{z}) = \frac{\exp\left(-(u - \mu_1)\right)}{(1 + \exp\left(-(u - \mu_1)\right))^2} \ ; -\infty < u < \infty$$

$$\text{and} \quad D = \begin{cases} 1, & U \geq 0 \\ \\ 0, & U < 0 \end{cases} .$$

$$\Rightarrow P(D = 1 \mid g, \mathbf{z}) = \frac{\exp(\mu_1)}{1 + \exp(\mu_1)} .$$

In order to connect U with Y we introduce another latent variable V. We assume that U is derived by transforming a continuous variable, V, whose marginal distribution given g and $\mathbf{z}$ is normal with mean μ_1 and variance 1, and that $(V, Y|g, \mathbf{z})$ follows bivariate normal. The required transformation is $U = \mu_1 + \log \frac{\Phi(V-\mu_1)}{1-\Phi(V-\mu_1)}$. Thus, we specify the following population model for the bivariate response $(D, Y|g, \mathbf{z})$,

$$\begin{pmatrix} V \\ \\ Y \end{pmatrix} \Bigg| \, g, \mathbf{z} \sim N \left(\begin{pmatrix} \mu_1 \\ \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_2 \\ \\ \rho\sigma_2 & \sigma_2^2 \end{pmatrix} \right) \text{, where}$$

$$\mu_1 = \alpha_1 + \beta_1 g + \boldsymbol{\gamma_1}'\mathbf{z} \ , \ \mu_2 = \alpha_2 + \beta_2 g + \boldsymbol{\gamma_2}'\mathbf{z} \ , \text{ and } \rho = \alpha_3 + \beta_3 g \ .$$

Let $U = \mu_1 + \log \frac{\Phi(V-\mu_1)}{1-\Phi(V-\mu_1)}$ and $D = \begin{cases} 1, & U \geq 0 \\ \\ 0, & U < 0 \end{cases}$.

Assuming disease prevalence to be known, the retrospective pseudo log-likelihood is a function of $\boldsymbol{\theta} = (\beta_1, \boldsymbol{\gamma_1}, \alpha_2, \beta_2, \boldsymbol{\gamma_2}, \sigma_2, \alpha_3, \beta_3)'$, and $p(g)$,

$$l_p = \sum_{i=1}^{n} \log P(D = d_i, Y = y_i | g_i, \mathbf{z}_i) + \sum_{i=1}^{n} \log P(G = g_i) \ ,$$

where the identity

$$P(D=1)=\sum_{\mathbf{z}}\sum_{g}\frac{\exp(\alpha_1+\beta_1 g+\boldsymbol{\gamma_1}'\mathbf{z})}{1+\exp(\alpha_1+\beta_1 g+\boldsymbol{\gamma_1}'\mathbf{z})}p(g)\hat{P}(\mathbf{Z}=\mathbf{z}) \tag{3.6}$$

is used to compute α_1 given $\boldsymbol{\theta}$ and $p(g)$. We obtain the pseudo MLE of $\boldsymbol{\theta}$ by maximizing l_p with respect to $(\boldsymbol{\theta}, p(g))'$. The joint probability of the primary and secondary phenotypes given g and $\mathbf{z}$ can be expressed as

$$\begin{aligned}
& P(D=d, Y=y|g,\mathbf{z}) \\
= \;& P(Y=y|g,\mathbf{z})P(D=d|y,g,\mathbf{z}) \\
= \;& \frac{1}{\sigma_2}\phi\left(\frac{y-\mu_2}{\sigma_2}\right)\left\{dP(D=1|y,g,\mathbf{z})+(1-d)\left(1-P(D=1|y,g,\mathbf{z})\right)\right\} \;,
\end{aligned}$$

$$P(D=1|y,g,\mathbf{z})=\Phi\left(\frac{\frac{\rho}{\sigma_2}(y-\mu_2)-\Phi^{-1}\left(\frac{1}{1+\exp(\mu_1)}\right)}{\sqrt{1-\rho^2}}\right) \;.$$

3.2.3 Simulation

We perform extensive simulations to compare the performance of our likelihood method with 1) the naïve method that involves including the disease status variable as a covariate in the regression for Y on g and $\mathbf{z}$, 2) cases-only analysis, 3) controls-only analysis, and 4) weighted regression. We are interested in estimating the effect of the genotype on the secondary phenotype, adjusting for the covariates. So β_2 is our main parameter of interest and we judge the different estimators of β_2 on the basis of bias and mean squared error (MSE). We fix the sample size at 3000, 1500 cases and 1500 controls, for

each dataset that we analyze. We consider a SNP with dominant mode of inheritance for both primary and secondary phenotypes and a minor allele frequency of 0.25. The population is assumed to be in Hardy-Weinberg equilibrium. We perform simulations for both binary and continuous secondary phenotypes. We examine β_1 and β_2 across the range -0.6 (OR=0.55) to 0.6 (OR=1.82). This grid corresponds to a biologically plausible range of values for complex diseases and helps us understand how these parameters affect bias and variance of the estimates. The γ coefficients are drawn at random from $N(0,1)$. α_1 is derived such that the disease prevalence is 0.05. For the binary secondary phenotype, a prevalence of 0.2 is used. For parameterizing the association between the primary and the secondary phenotypes given the genotype and the covariates, we used $\alpha_3 = \log(9)$ and $\beta_3 = 0$. For the continuous secondary phenotype, we set $\alpha_2 = 1$, $\sigma_2 = 1$, $\alpha_3 = 0.6$, and $\beta_3 = 0$.

We present two sets of simulations. The first set corresponds to a single binary covariate with probability of success 0.45. In the second set of simulations we add another covariate, a continuous one which is normally distributed and independent of the binary covariate. So each simulation setup is indexed by the number of covariates: one or two, the nature of the secondary phenotype: binary or continuous, and the pair (β_1, β_2), We replicated 1000 datasets for each simulation setup. For the first set of simulations we parameterize the covariate distribution by the probability of success of the binary covariate and maximize the likelihood with respect to it, as described in Methods 3.2. In the second set of simulations we apply the pseudo likelihood approach.

3.3 Results

Figures 3.1 and 3.2 present simulation results for a single binary covariate. Figure 3.1 corresponds to a binary secondary phenotype and Figure 3.2 is for a continuous secondary phenotype. Figures 3.3 and 3.4 depict simulation results when we have two covariates, one binary and the other continuous. For a binary secondary phenotype we have Figure 3.3 and for the continuous case we have Figure 3.4.

3.3.1 One binary covariate

We have plotted the means and the MSEs of the estimators of β_2 against β_1 for different values of β_2. The first column of Figure 3.1 shows the means and the second column depicts the MSEs. The different rows correspond to different values of β_2. For each plot the estimators are denoted by different plotting symbols and color: the magenta crosses are for the retrospective likelihood, the black pluses correspond to the naïve method, the red solid circles are for the case-only analysis, the green ones correspond to the controls-only analysis, and the blue triangles are for the weighted method.

The first column of Figure 3.1 shows the means of the estimators of β_2. The naïve estimator shows very large bias. It overestimates for positive values of β_1 and underestimates for negative values of β_1. There is almost no difference in bias for the estimators derived by considering only the case or the control populations. Both these methods have smaller bias than the naïve method but the direction of bias is opposite. The weighted estimator and the MLE are effectively unbiased for all values of (β_1, β_2). All

the methods become nearly unbiased when both β_1 and β_2 are zero, *i.e.*, the SNP is not associated with either phenotype. As we decrease the value of β_2 from zero, the values of β_1 for which the biases in the naïve, the cases-only, and the controls-only estimators are almost reduced to zero also decrease. Similarly, as β_2 slides from zero in the positive direction, the values of β_1 for which the three estimators are unbiased also slides in the positive direction, but it is difficult to predict which particular combination of β_1 and β_2 would give us practically unbiased estimators for these three methods.

The second column of Figure 3.1 shows the MSEs of the estimators of β_2. The naïve estimator exhibits large MSE for most of the grid of (β_1, β_2) values. The MSE drops remarkably when the naïve estimator is almost unbiased but it is difficult to characterize the parameter values for which it happens. The cases-only analysis has lower MSE than the controls-only analysis for the (β_1, β_2) values examined. The weighted estimator and the MLE exhibit a even performance across the range of β_1, the MLE having slightly lower MSE throughout than the weighted estimator. Both of them have smaller MSE than the naïve, the cases-only, and the controls-only estimators for most of the grid.

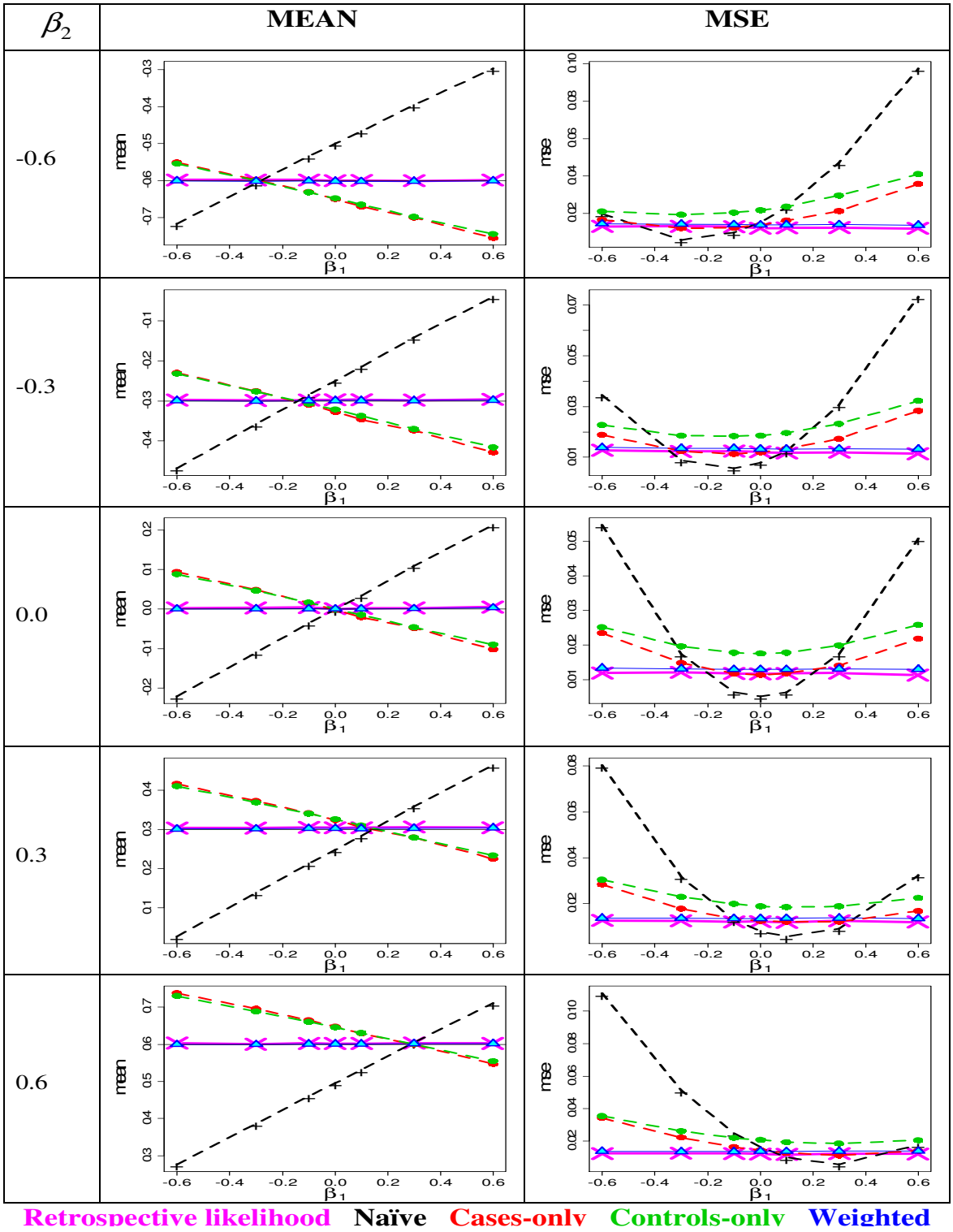


Figure 3.1: Means and MSEs for binary secondary phenotype with one covariate

When the secondary phenotype is continuous the bias and the MSE of the estimators do not depend on the value of β_2. So for Figure 3.2 we have only a single row of plots displaying the means and the MSEs of the estimators plotted against β_1. The left and the right plots show the means and the MSEs respectively. From the left plot we see that the estimator derived from the cases has extremely large bias while the controls-only estimator is remarkably less biased. The bias for the naïve estimator lies between those of the cases-only and the controls-only estimators but is opposite in direction. The weighted estimator and the MLE are virtually unbiased. All the estimators are nearly unbiased when β_1 is zero, i.e., when the SNP is not associated with the disease. The corresponding MSE values for the estimators suggest that the MLE has the lowest MSE for almost all β_1 values examined. The estimator derived from the case population has the highest MSE, followed by the naïve estimator. The weighted estimator has a low MSE throughout the β_1 range, but it is slightly higher than that for the MLE. The MSE of the estimator based on the control population is marginally higher than the MSE for the weighted approach.

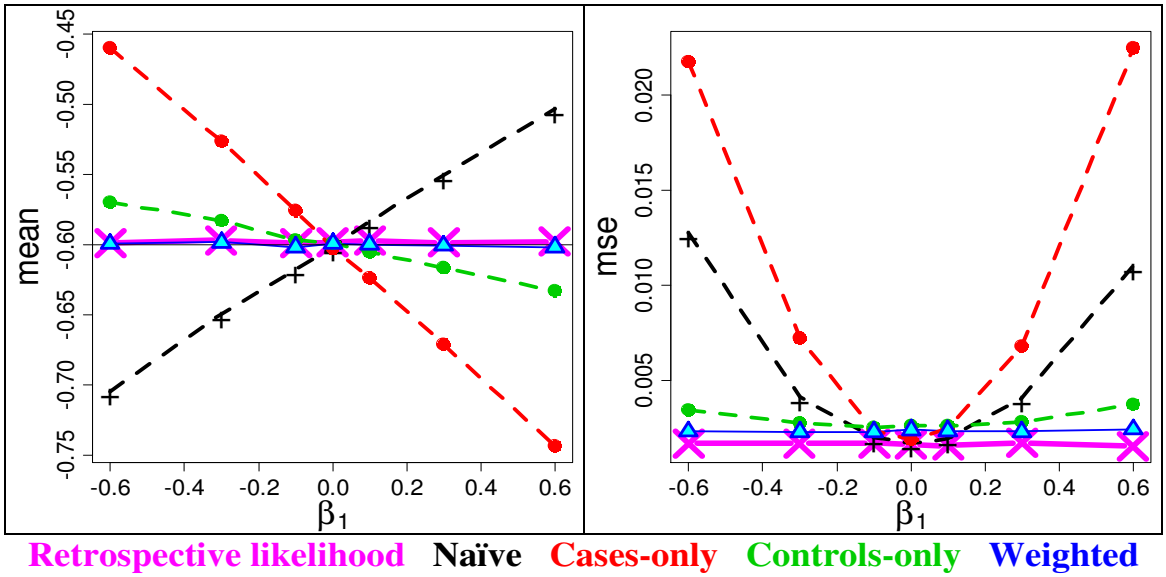


Figure 3.2: Means and MSEs for continuous secondary phenotype with one covariate

3.3.2 One binary and one continuous covariate

Figures 3.3 and 3.4 display the means and the MSEs of the estimators obtained via pseudo likelihood approach in the case where we have two covariates: one binary and one continuous. Figure 3.3 shows how the means and the MSEs of the estimators change according to the values of β_1 and β_2 when we have a binary secondary phenotype. The plots in the first column of Figure 3.3 displaying the means of the estimators are similar in pattern to the plots in the first column of Figure 3.1 where we had only one covariate. However, the difference between the estimators obtained from the cases and the controls separately is pronounced here: the case-only estimator is clearly more biased than the controls-only estimator. The naïve estimator exhibits extremely large bias. The pseudo MLE and the weighted estimator reduce bias significantly compared to the naïve estimator and are nearly unbiased throughout the grid. The corresponding

MSE plots are displayed in the second column of Figure 3.3. The pseudo MLE exhibits a low MSE across the entire grid. Even though the weighted estimator has a even performance across the β_1 range, it has a higher MSE than pseudo MLE throughout. The cases-only estimator has smaller MSE than the controls-only estimator for most the the (β_1, β_2) values examined. The naïve estimator exhibits very large MSE values, especially for large values of β_1 and β_2. MSE of the naïve estimator drops below all the other estimators for a short range of β_1 values for each value of β_2. This is due to the fact that the naïve estimator is very close to the true parameter value in this range, but as mentioned before, it is difficult to characterize these (β_1, β_2) combinations where the naïve estimator performs well.

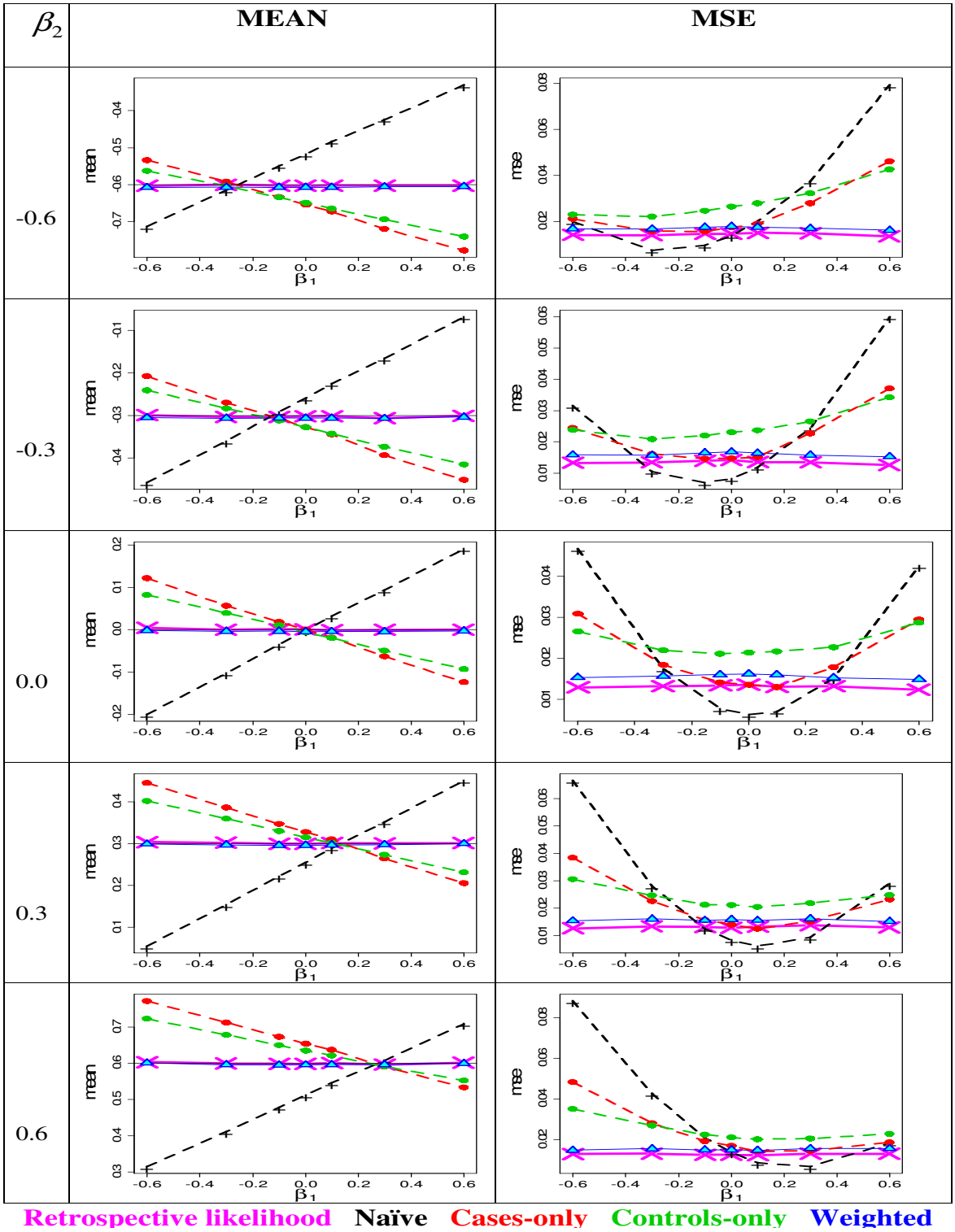


Figure 3.3: Means and MSEs for binary secondary phenotype with covariates

Figure 3.4 displays the means and the MSEs of the estimators, when we have a continuous secondary phenotype, plotted against β_1. We see that adding a continuous covariate and applying pseudo likelihood approach almost affected no change in the means and the MSEs of the estimators and Figure 3.4 is almost the same as Figure 3.2.

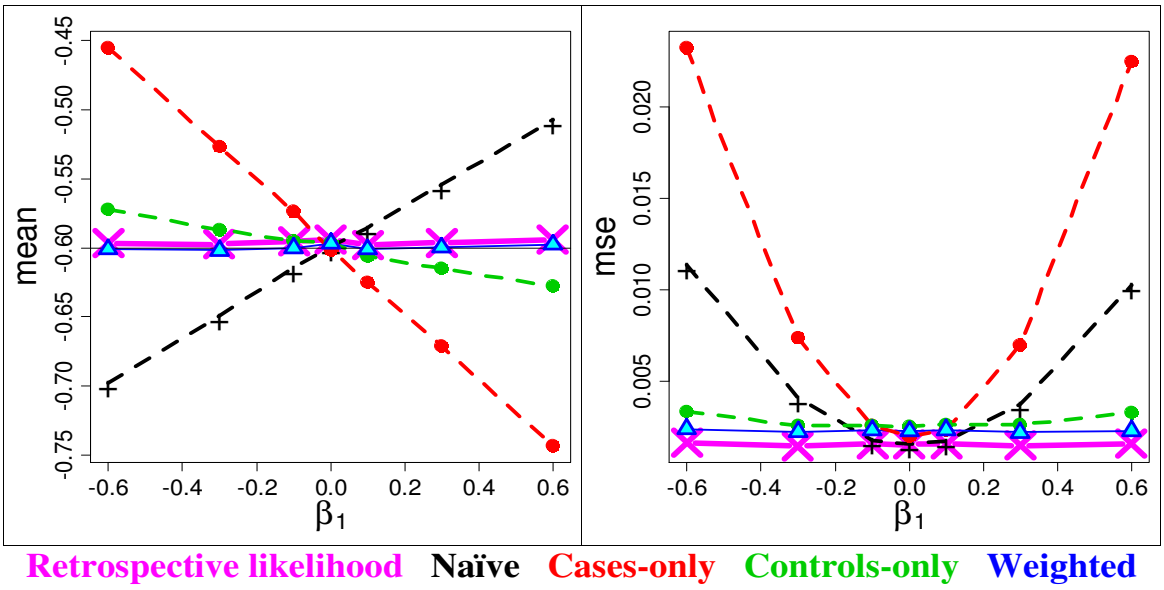


Figure 3.4: Means and MSEs for continuous secondary phenotype with covariates

3.4 Discussion

We have presented an approach for appropriate analysis of secondary phenotypes for case-control data, the disease status being the primary phenotype. We have assumed a retrospective likelihood approach to analyzing the data. Our method provides for both binary and continuous secondary phenotypes. We have used the Palmgren model (Palmgren 1989) to specify the joint distribution of the primary and the secondary phenotypes when the secondary phenotype is binary. For the continuous case, we have

suggested a novel bivariate model involving two latent variables. This joint model is such that the marginals of the primary and the secondary phenotypes follow the distributions that are conventionally used to analyze them separately. We have provision for including covariates, both binary and continuous, in the analysis. We have discussed in detail how to handle the covariate distribution and have introduced the pseudo likelihood approach for easier implementation of our method.

There are several ways of analyzing the secondary phenotype, a few of them are *ad hoc* and the rest more formal, as pointed out by Jiang et al. (2006). The naïve method of ignoring the sampling scheme, methods involving analyzing only the cases or the controls, and the method that conditions on the disease status by including it in the regression are generally considered as *ad hoc* and it is not definitely known under which circumstances they will provide valid estimates of the association between the genotype and the secondary phenotype. Lin and Zeng (2008) provide a detailed critique of the standard methods. However, their theoretical conclusions are applicable for the conditional model that they have assumed to be true. It is difficult to predict how any of these methods would behave if we were to believe in any other population model for the joint distribution of the primary and the secondary phenotypes.

The likelihood approach and the weighted approach have theoretical justifications. The likelihood method involves the retrospective likelihood. We can work with the retrospective likelihood in several ways depending on how we decide to model the population and also how we parametrize it. Also, there are many options for handling the covariate distribution. The weighted approach, on the other hand, is free of any

modeling assumption. It is simple, fast, and easy to implement. But it is usually less efficient than the likelihood method as we have also seen in our simulations. With only two covariates in the model, it had higher MSE than the pseudo MLE. Moreover, with the weighted method there is no way of exploring the relationship between the primary and the secondary phenotypes. The likelihood method provides us with estimates and standard errors of the parameters specifying that relationship.

Pseudo maximum likelihood estimation was first proposed by Gong and Samaniego (1981) in the parametric set up. Later Hu and Lawless (1997) applied it for problems with response-related missing covariates where they replaced an unknown distribution, the nuisance parameter for the problem, in the likelihood with its empirical estimate. For their problem they have presented in details the asymptotic theory and also estimates of the asymptotic variance of the pseudo MLE. The use of pseudo likelihood estimation in our case would allow us a great amount of flexibility and would obviate the need for explicitly dealing with the covariate distribution. It also makes the implementation very fast. But as a cost for this flexibility, the variance of the pseudo MLE is not the inverse of the negative Hessian matrix as is usually the case for likelihood methods or even profile likelihood methods. The variance of the pseudo MLE is complicated because of the nuisance parameter and the uncertainly involved in estimating it. Future work in the area of secondary analysis for case-control data would benefit from the development of the asymptotic theory of pseudo MLE and estimate of the asymptotic variance of the pseudo MLE for our problem.

It is important to understand the behavior of the different estimators to be able to

predict whether in a particular situation they would provide valid estimates or not. We need to develop theoretical properties of the estimators to be able to understand the direction of their bias. It is also required to understand how the different parameters such as prevalence, association between the genotype, and the primary and secondary phenotypes, the association between the two phenotypes, or the MAF affect the bias and the MSE of the estimators and would have important implications for future work in this area.

Chapter 4

Significance Bias for Secondary Phenotypes and GXE Interaction

4.1 Introduction

The primary purpose of genome-wide association studies (GWAS) is identification of SNPs influencing susceptibility to complex traits. Since, in modern whole genome scans, usually hundreds of thousands of SNPs are genotyped, thresholds in the range $10^{-7} - 10^{-8}$ are generally used for point-wise significance (Todd et al. 2007; Zondervan and Cardon 2007; Scott et al. 2007). Using the original data for estimation purpose coupled with the application of stringent thresholds, distorts the estimation process, producing inflated estimates of effect sizes. After detection, genetic effect of the significant SNPs are estimated based on the same data. This phenomenon (commonly referred to as the "winner's curse" (Lohmueller et al. 2003; Zöllner and Pritchard 2007) or "significance bias" (Ghosh et al. 2008)) has profound importance for estimation of genetic effects

and is well documented in the literature (Zöllner and Pritchard 2007; Ghosh et al. 2008; Garner 2007; Göring et al. 2001; Siegmund 2002; Sun and Bull 2005; Yu et al. 2007). A related problem arises for risk estimation of secondary effects, such as secondary phenotypes or gene-environment interactions, when the secondary analysis is restricted to SNPs that are found to be significant in the primary analysis. Such secondary bias can be substantial but has received no attention so far in the GWAS literature. If the biased results are used for the design of follow-up studies, they are likely to be underpowered, relying on an inflated estimate of effect size. The variance of the risk estimates may also be affected.

Most genetic studies gather information on a host of variables besides the disease status, the primary phenotype. Subsequent to the initial detection of the SNPs significant for the primary phenotype, estimation of the impact of these SNPs on other correlated traits is of interest (Frayling et al. 2007). The well-known phenomenon of significance bias which affects the estimation of the disease risk effect also distorts the estimation of the effect of the significant SNP on correlated phenotypes. Also, the sampling design followed is of critical importance in deciding the analysis procedure for the secondary trait. GWAS commonly employs case-control design to collect data on a range of qualitative and quantitative variables and usually standard logistic or linear regressions, ignoring the case-control sampling mechanism, are applied for the secondary analysis. It has been proved that only under very restrictive conditions, these analyses methods would yield unbiased estimates of the population parameters of interest (Nagelkerke et al. 1995). Hence, for the analysis of secondary traits, sampling

bias (Nagelkerke et al. 1995; Lee et al. 1997; Lin and Zeng 2008) adds to the problem of significance bias in case-control sampled data. We have presented a detailed discussion of existing methods for secondary analysis of case-control data and have developed a retrospective likelihood method in Chapter 3 that we use for analysis in this chapter.

Interplay of genes and environmental factors contribute to the susceptibility to complex traits. After a SNP is identified, researchers are often interested in estimating the gene-environment interaction effect. Since this estimation is performed conditional on the fact that the SNP has been found significant for the disease phenotype, significance bias can be a major concern for the estimation of gene-environment interaction effect. Results exploring the phenomenon of significance bias for estimation of interaction effect have received very little attention in the GWAS literature.

Significance bias for estimation of the primary effect, *i.e.*, the disease risk effect, and ways of reducing it or eliminating it has been investigated in detail in many publications (Garner 2007; Ghosh et al. 2008; Göring et al. 2001; Siegmund 2002; Sun and Bull 2005; Yu et al. 2007; Zöllner and Pritchard 2007). While the problem of significance bias is well appreciated in the context of disease phenotype, it has not yet been explored for analyses of secondary effects, be it effect size for additional phenotypic trait or gene-environment interaction. We have recently proposed a conditional likelihood approach (Ghosh et al. 2008), based on the estimate of genetic effect and its standard error, to correct for the bias in effect size estimation for disease risk in case-control association studies. We propose an extension of the conditional likelihood approach to the multivariate setting where multiple effect coefficients are simultaneously estimated.

For implementing this method we need the naïve estimates of the primary and secondary effect sizes and an estimate of the covariance between the naïve estimates. We provide formulas for estimating the covariances for different sampling scenarios. We prove that, under certain conditions, estimation of the gene-environment interaction effect conditional on the significance of the marginal effect of the SNP is not affected by significance bias. However, after a SNP is found significant in a logistic regression involving gene, environment, and their interaction, if we wish to estimate the interaction effect, the estimation may be distorted by significance bias. In that case, our proposed method can be applied to provide bias-reduced estimates of the interaction effect.

We illustrate the performance of our approach via extensive simulations. The simulations cover a biologically plausible range of disease effect sizes. We show results for both prospective and retrospective sampling schemes. Compared to the naïve estimation ignoring the selection based on significance, our approach provides remarkably reduced bias and mean-squared error. The results have considerable importance for the proper analysis of secondary effects, and in the design of follow-up studies.

4.2 Methods

We assume a genetic model for disease risk that includes among other covariates, a single SNP with either recessive, dominant, or additive mode of inheritance. We use $\beta_1 = \log(OR_1)$ to denote the primary effect: the true log odds ratio for disease risk

conferred by a referent genotype or by each allele as in an additive model, adjusting for other covariates in the model. A single locus test statistic for the primary effect can be expressed as an estimate for β_1 divided by an estimate for its standard error,

$$Z_1 = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)} ,$$

which is compared to its asymptotic null distribution $N(0,1)$. Let $\boldsymbol{\beta}_{-1} = \begin{pmatrix} \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$ denote the vector of true log odds ratios attributable to the secondary effects such as the effects of the SNP on secondary phenotypes or the effect of SNP-environment interaction on disease risk. Our goal is to estimate $\boldsymbol{\beta}_{-1}$ only when the SNP is significant for the disease in two-sided testing, *i.e.*, $|z_1| > c$ for a value c corresponding to genome-wide significance. We will refer to $\hat{\boldsymbol{\beta}}_{-1}$, obtained from standard statistical procedures without acknowledging this selection of the SNP based on significance prior to estimation, as the naïve estimator. We have recently reported an approximate conditional likelihood approach to estimate β_1 conditional on the SNP being significant for disease association (Ghosh et al. 2008). We have shown that our method offers marked improvements over the naïve estimation procedure. We extend the concept of explicitly considering the selection in formulating the likelihood to the multivariate setting and propose three new estimators.

4.2.1 Significance bias

Estimation of $\boldsymbol{\beta}_{-1}$ can be restated as a mean-parameter estimation problem for truncated multivariate normal distribution with known variance-covariance matrix. To see how, we define $\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \\ \boldsymbol{\beta}_{-1} \end{pmatrix}$, $\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \\ \hat{\boldsymbol{\beta}}_{-1} \end{pmatrix}$, and $\mathbf{Z} = \begin{pmatrix} Z_1 \\ \\ \mathbf{Z}_{-1} \end{pmatrix}$ where $\mathbf{Z}_{-1} = \begin{pmatrix} Z_2 \\ \vdots \\ Z_p \end{pmatrix}$ with $Z_i = \frac{\hat{\beta}_i}{\hat{SE}(\hat{\beta}_i)}$. $\hat{\beta}_i$s and $\hat{SE}(\hat{\beta}_i)$s are obtained from maximum likelihood and the information matrix. From the standard result

$$\begin{pmatrix} \frac{\hat{\beta}_1-\beta_1}{\hat{SE}(\hat{\beta}_1)} \\ \frac{\hat{\beta}_2-\beta_2}{\hat{SE}(\hat{\beta}_2)} \\ \vdots \\ \frac{\hat{\beta}_p-\beta_p}{\hat{SE}(\hat{\beta}_p)} \end{pmatrix} \rightarrow_D N(\mathbf{0}, \mathbf{R}) \text{ with increasing sample size,}$$

where $\mathbf{R} = corr(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} 1 & \boldsymbol{\rho}' \\ \\ \boldsymbol{\rho} & \mathbf{R}_{22} \end{pmatrix}$, it follows that $\mathbf{Z} \dot{\sim} N(\boldsymbol{\mu}, \mathbf{R})$, $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \\ \boldsymbol{\mu}_{-1} \end{pmatrix}$ where $\boldsymbol{\mu}_{-1} = \begin{pmatrix} \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$ with $\mu_i = \frac{\beta_i}{\hat{SE}(\hat{\beta}_i)}$. A consistent estimate of $\mathbf{R}$ can be obtained from the likelihood theory. For the ease of mathematical discourse, we assume that it

is known. The conditional density of $\mathbf{Z}$ given $|Z_1| > c$ is:

$$\begin{aligned} p_{\boldsymbol{\mu}}(\mathbf{z}|\ |Z_1| > c) &= \frac{p_{\boldsymbol{\mu}}(\mathbf{z})}{P(|Z_1| > c)} \\ &= \frac{N_p(\mathbf{z}; \boldsymbol{\mu}, \mathbf{R})}{\Phi(-c - \mu_1) + \Phi(-c + \mu_1)} \ , \end{aligned}$$

where $|z_1| > c$, $-\infty < z_i < \infty$, $i = 2, \ldots, p$, $-\infty < \mu_i < \infty$, $i = 1, \ldots, p$, and Φ is the cumulative distribution function of a standard normal. The statistical exposition to follow is based entirely on this "μ-version" of the problem. It is evident that, in the "μ-version", the problem boils down to estimation of $\boldsymbol{\mu}$ from $p_{\boldsymbol{\mu}}(\mathbf{z}|\ |Z_1| > c)$: a truncated multivariate normal density with mean $\boldsymbol{\mu}$ and known variance-covariance matrix.

Our naïve estimate of $\boldsymbol{\mu}$ based on $p_{\boldsymbol{\mu}}(\mathbf{z})$ is $\hat{\boldsymbol{\mu}} = \mathbf{z}$, and the expectation can be shown analytically to be (see Appendix B, section B1)

$$E_{\boldsymbol{\mu}}\left(\mathbf{Z}|\ |Z_1| > c\right) = \boldsymbol{\mu} + \begin{pmatrix} 1 \\ \\ \boldsymbol{\rho} \end{pmatrix} \frac{\phi(c - \mu_1) - \phi(c + \mu_1)}{\Phi(-c - \mu_1) + \Phi(-c + \mu_1)} \ , \tag{4.1}$$

where ϕ is the density function of a standard normal. It is clear from (4.1) that the bias incurred in naïve estimation of $\boldsymbol{\mu}_{-1}$ is $\boldsymbol{\rho}$ times the bias in $\hat{\mu}_1$, thus being the same for different values of $\boldsymbol{\mu}_{-1}$. Equation (4.1) illustrates the phenomenon of significance bias and implies that in the special case of the null hypothesis $\mu_1 = 0$ being true, the naïve estimates are unbiased.

4.2.2 An approximate conditional likelihood

The approximate asymptotic distribution of $\mathbf{Z}$, approximate since $\boldsymbol{\mu}$ is not truly a parameter, suggests the following approximate likelihood for $\boldsymbol{\mu}$,

$$L(\boldsymbol{\mu}) = p_{\boldsymbol{\mu}}(\mathbf{z}) = N_p(\mathbf{z}; \boldsymbol{\mu}, \mathbf{R}) \ .$$

The above likelihood applies to a wide variety of settings, being free from any nuisance parameters that may have been included in the model, such as other clinical covariates, stratification variables, or the effects of other SNP genotypes. The maximum likelihood estimate (MLE) from the above likelihood is $\hat{\boldsymbol{\mu}} = \mathbf{z}$. By explicitly considering the fact that the SNP has been found significant, we have the approximate *conditional* likelihood

$$L_c(\boldsymbol{\mu}) = p_{\boldsymbol{\mu}}(\mathbf{z}| \ |Z_1| > c) = \frac{N_p(\mathbf{z}; \boldsymbol{\mu}, \mathbf{R})}{\Phi(-c - \mu_1) + \Phi(-c + \mu_1)} \ .$$

We develop three improved estimators for $\boldsymbol{\mu}$ based on this approximate conditional likelihood. For any proposed $\hat{\boldsymbol{\mu}}$ we can easily convert back to $\hat{\boldsymbol{\beta}}$ using $\hat{\beta}_i = \hat{\mu}_i \hat{SE}(\hat{\beta}_i)$. Hence, for estimation of $\boldsymbol{\mu}$ we require only the significance threshold c, $\hat{\boldsymbol{\beta}}$, and $\hat{Var}(\hat{\boldsymbol{\beta}})$.

4.2.3 The conditional MLE

The conditional likelihood suggests, as one possible solution, the MLE, given by

$$\widetilde{\boldsymbol{\mu}}^{(1)} = arg\ max_{\boldsymbol{\mu}}\ L_c(\boldsymbol{\mu}) \ ,$$

which can be obtained by maximizing the conditional likelihood with respect to μ_1 and $\boldsymbol{\mu}_{-1}$ separately, and then solving for them simultaneously. Hereafter "$\sim$" will signify estimates based on the conditional likelihood.

$$\begin{aligned}\frac{\partial}{\partial \mu_1} L_c(\boldsymbol{\mu}) = 0 \quad &\Rightarrow \quad \frac{\partial}{\partial \mu_1}\left\{\frac{\phi(\frac{z_1-\mu_c}{\sigma_c})}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\right\} = 0 \\ &\Rightarrow \quad \frac{z_1-\mu_c}{\sigma_c^2} = \frac{\phi(-c+\mu_1)-\phi(c+\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\,, \end{aligned} \tag{4.2}$$

where $\mu_c = \mu_1 + \boldsymbol{\rho}'\mathbf{R}_{22}^{-1}(\mathbf{z}_{-1} - \boldsymbol{\mu}_{-1})$ and $\sigma_c^2 = 1 - \boldsymbol{\rho}'\mathbf{R}_{22}^{-1}\boldsymbol{\rho}$.

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu_{-1}}} L_c(\boldsymbol{\mu}) = 0 \quad &\Rightarrow \quad \frac{\partial}{\partial \boldsymbol{\mu_{-1}}} N_{p-1}\left(\mathbf{z}_{-1};\ \boldsymbol{\mu}_{-1} + \boldsymbol{\rho}(z_1-\mu_1), \mathbf{R}_{22} - \boldsymbol{\rho}\boldsymbol{\rho}'\right) = 0 \\ &\Rightarrow \quad \mathbf{z}_{-1} = \boldsymbol{\mu}_{-1} + \boldsymbol{\rho}(z_1-\mu_1)\,. \end{aligned} \tag{4.3}$$

Substituting (4.3) in (4.2) we have,

$$z_1 - \mu_1 = \frac{\phi(c-\mu_1)-\phi(c+\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}. \tag{4.4}$$

We solve for μ_1 from (4.4) and plug it in (4.3) to get $\widetilde{\mu}_1^{(1)}$ and $\widetilde{\boldsymbol{\mu}}_{-1}^{(1)}$, respectively. $L_c(\boldsymbol{\mu})$ lends itself to the above mathematical treatment because it can be expanded as

$$L_c(\boldsymbol{\mu}) = \frac{N_{p-1}(\mathbf{z}_{-1};\boldsymbol{\mu}_{-1},\mathbf{R}_{22})\phi(\frac{z_1-\mu_c}{\sigma_c})}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\,, \tag{4.5}$$

or as

$$L_c(\boldsymbol{\mu}) = \frac{\phi(z_1-\mu_1)N_{p-1}\left(\mathbf{z}_{-1};\boldsymbol{\mu}_{-1}+\boldsymbol{\rho}(z_1-\mu_1), \mathbf{R}_{22}-\boldsymbol{\rho}\boldsymbol{\rho}'\right)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\,. \tag{4.6}$$

(4.5) uses the conditional density of Z_1 given $\mathbf{Z_{-1}}$ and (4.6) uses the conditional density of $\mathbf{Z_{-1}}$ given Z_1. The maximum likelihood estimator is not unbiased. This motivates us to explore other probable estimators. The moments estimator obtained by solving $E_{\boldsymbol{\mu}}(\mathbf{Z}|\ |Z_1| > c) = \mathbf{z}$ for $\boldsymbol{\mu}$ is one such possibility. Such a "bias-correction" estimator has intuitive appeal, representing the value $\boldsymbol{\mu}$ for which the truncated multivariate normal distribution would be expected to generate $\mathbf{z}$. Surprisingly, the maximum likelihood estimator and the moments estimator are the same. Comparing and combining $E_{\boldsymbol{\mu}}(\mathbf{Z}|\ |Z_1| > c) = \mathbf{z}$, (4.1), (4.3), and (4.4), we arrive at the above conclusion.

4.2.4 The mean of the normalized conditional likelihood

The motivation to reduce mean squared error (MSE) suggests another estimator,

$$\widetilde{\boldsymbol{\mu}}^{(2)} = \frac{\int \boldsymbol{\mu} L_c(\boldsymbol{\mu}) d\boldsymbol{\mu}}{\int L_c(\boldsymbol{\mu}) d\boldsymbol{\mu}}\ .$$

We can think of $\widetilde{\boldsymbol{\mu}}^{(2)}$ as the mean of the random variable following the distribution $L_c(\boldsymbol{\mu})$, normalized to be a proper density. However, the multivariate integration after some simplification boils down to obtaining $\widetilde{\mu}_1^{(2)}$ as the mean of the random variable following the distribution $L_c(\mu_1) = \frac{\phi(z_1-\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}$, normalized to be a proper density, and then plugging it in $\mathbf{z}_{-1} - \boldsymbol{\rho}(z_1 - \widetilde{\mu}_1^{(2)})$ to get $\widetilde{\boldsymbol{\mu}}_{-1}^{(2)}$ (see Appendix B, section B2). Thus,

$$\begin{cases} \widetilde{\mu}_1^{(2)} = \frac{\int \mu_1 L_c(\mu_1) d\mu_1}{\int L_c(\mu_1) d\mu_1} \\ \widetilde{\boldsymbol{\mu}}_{-1}^{(2)} = \mathbf{z}_{-1} - \boldsymbol{\rho}(z_1 - \widetilde{\mu}_1^{(2)}) \end{cases}\ .$$

4.2.5 A compromise estimator

The conditional likelihood $L_c(\boldsymbol{\mu})$ is typically skewed, as a result $\widetilde{\boldsymbol{\mu}}^{(1)}$ and $\widetilde{\boldsymbol{\mu}}^{(2)}$ may differ appreciably for certain values of $\mathbf{z}$. Thus, as a practical compromise we also examine the estimator

$$\widetilde{\boldsymbol{\mu}}^{(3)} = (\widetilde{\boldsymbol{\mu}}^{(1)} + \widetilde{\boldsymbol{\mu}}^{(2)})/2 \,,$$

which balances the strengths of $\widetilde{\boldsymbol{\mu}}^{(1)}$ and $\widetilde{\boldsymbol{\mu}}^{(2)}$.

4.2.6 Secondary phenotype

We performed several simulations to assess our proposed bias-correction method for secondary phenotypes. To describe our simulations, we begin with the notation. Let d denote the disease status (0=control, 1=case), y the secondary phenotype value, and g the genotype predictor value for an individual. For a biallelic SNP with major allele A and minor allele a, g is defined as follows for different genetic models with respect to a,

Recessive	*Additive*	*Dominant*
$g = \begin{cases} 0, & AA \\ 0, & Aa \\ 1, & aa \end{cases}$	$g = \begin{cases} 0, & AA \\ 1, & Aa \\ 2, & aa \end{cases}$	$g = \begin{cases} 0, & AA \\ 1, & Aa \\ 1, & aa \,. \end{cases}$

Logistic regression of D on g is used to test for disease association, and once the SNP is found to be significant for the disease, effect of the SNP on the secondary phenotype is of interest. For simulation purposes, we consider a binary secondary phenotype and

assume a logistic model for the the conditional distribution of Y given g. The marginal distribution of G is denoted by $p(g)$, which, given the minor allele frequency, is derivable from Hardy-Weinberg equilibrium. We jointly model the bivariate response (D, Y) given g. There are several ways to parametrize the bivariate distribution $(D, Y|g)$. We consider only those models for which the marginal distributions of D and Y given g are both logistic. The bivariate logistic model, considered by Palmgren (1989), is one such model and is conceptually very simple. It is based on the fact that the joint distribution of two binary variables can be specified in terms of their marginal probabilities and their odds ratio. Thus, for a randomly sampled individual in the population we specify the joint distribution of D and Y given g as:

$$logit\ P(D = 1 \mid g, \mathbf{z}) = \alpha_1 + \beta_1 g + \boldsymbol{\gamma_1}'\mathbf{z}$$

$$logit\ P(Y = 1 \mid g, \mathbf{z}) = \alpha_2 + \beta_2 g + \boldsymbol{\gamma_2}'\mathbf{z}$$

$$\log OR(D, Y \mid g) = \frac{P(D = 1, Y = 1 \mid g, \mathbf{z})P(D = 0, Y = 0 \mid g, \mathbf{z})}{P(D = 1, Y = 0 \mid g, \mathbf{z})P(D = 0, Y = 1 \mid g, \mathbf{z})} = \alpha_3 + \beta_3 g\ .$$

For data collected prospectively, use of separate logistic regressions for D and Y is "valid" in the sense that the estimates obtained from the analysis are consistent estimates of the corresponding population parameters. For case-control data, a logistic modeling for β_1 applies. However, for the secondary analysis, ordinary logistic regression of Y on g to infer about β_2 is not necessarily valid. For appropriate analyses of the secondary phenotypes we refer to the retrospective likelihood method described in

Chapter 3.

Each dataset was simulated and analyzed in R v.2.5.1. We considered a sample of size $n = n_{cases} + n_{controls} = 3000$, and for retrospective sampling, we assumed $n_{cases} = n_{controls} = 1500$. We assumed a disease prevalence of 0.18 and a MAF value of 0.25 throughout. We fixed β_2 at 0.3, the bias-correction approach is not sensitive to this specification, as is suggested by equation (4.1). We used $c = 5.0$ corresponding to a p-value of $5.7\text{x}10^{-7}$, near the genome scan threshold considered by others (Scott et al. 2007; Todd et al. 2007; Zondervan and Cardon 2007). We simulated data only under the dominant mode of inheritance. We fixed the prevalence for the secondary trait at 0.2 and examined β_1 ranging from -0.7 (OR $\approx$0.5) to 0.7 (OR $\approx$ 2). This spans the range of biologically plausible values for complex diseases. We used $\alpha_3 = \log 9$ and $\beta_3 = 0$. For each value of β_1 enough simulations were performed to guarantee 1000 significant datasets. The null and near-null scenarios required massive simulation, on the order of 10^{10} for some scenarios, to obtain sufficient number of rejections, *i.e.*, significant datasets. We initially vetted datasets using chi-square statistics before performing logistic regression of D on g. Datasets meeting the criteria: chi-square statistic value ≥ 24.5 were used to capture datasets with $z_1^2 \geq c^2 = 25$. Finally, datasets achieving $z_1^2 \geq c^2 = 25$ were analyzed to obtain the parameter estimates, standard errors, and an estimate of the correlation between the parameter estimates. These estimates are then used to calculate $\widetilde{\beta}_2^{(1)}$, $\widetilde{\beta}_2^{(2)}$, and $\widetilde{\beta}_2^{(3)}$. Analytical derivation for formulae used to obtain covariance estimates are provided in Appendix B, section B3.

4.2.7 Gene-environment interaction

Our bias correction approach can also be applied for estimation of gene-environment interaction effect subsequent to finding the SNP significant. Let e denote the environment predictor value (0=not exposed, 1=exposed). We assume the following logistic model for a randomly sampled individual in the population,

$$logit\ P(D = 1 \mid g, e) \quad = \quad \beta_0^{(b)} + \beta_G^{(b)} g + \beta_E^{(b)} e + \beta_{GE}^{(b)} ge\ .$$

If we wish draw inference on $\beta_{GE}^{(b)}$ conditional on the gene being significant in two-sided testing, *i.e.*, $|z_G| = \frac{\hat{\beta}_G^{(b)})}{\hat{SE}(\hat{\beta}_G^{(b)})} > c$, then parameter estimates of the genetic effect and the gene-environment interaction effect, and an estimate of the correlation between the two estimates, as provided by any standard statistical software, will be sufficient to implement the bias-correction approach. For simulation purposes, we generated data prospectively from the above model with β_g ranging from -0.7 (OR $\approx$0.5) to 0.7 (OR $\approx$ 2). We fixed β_e and β_{ge} at 0.3 and 0.2 respectively. We considered a biallelic SNP with dominant effect on the disease phenotype. We fixed disease prevalence at 0.01, the threshold c at 5, and considered a MAF of 0.25. We compared the naïve estimator of the gene-environment interaction effect, $\hat{\beta}_{ge}$, with the conditional m.l.e. $\widetilde{\beta}_{ge}$. However, the usual practice is to first fit a logistic model of D on g only, and if the gene is found significant, then, a full model involving gene, environment, and their interaction follows. We demonstrate analytically that, $cov(\hat{\beta}_G^{(s)}, \hat{\beta}_{GE}^{(b)}) = 0$, where $\hat{\beta}_G^{(s)}$ is the estimate of the genetic effect obtained from fitting a logistic regression of D on g only. Detailed

derivation is provided in the Appendix B, section B4. Hence, for prospective data, the gene-environment interaction effect estimated from the full model subsequent to the SNP being found significant for the disease in the gene-only model, is not affected by significance bias.

4.3 Results

In all the simulation scenarios described here, expectations and MSEs are calculated conditional on significance. All of them are plotted against β_1 with corresponding odds ratio values.

4.3.1 Secondary phenotype

Figures 4.1 and 4.2 display the simulation results for the secondary phenotype. Figure 4.1 plots the means of the the naïve and corrected estimators. The corresponding MSE plots are shown in Figure 4.1.

Bias

The first column of Figure 4.1 plots the means for the naïve and corrected estimators for β_1. The naïve estimator shows very large bias, especially for small to moderate β_1. The corrected estimators show dramatically reduced bias for most β_1 values examined, although, they tend to undercorrect for small β_1 and overcorrect for large β_1. All the estimators are nearly unbiased for large values of β_1. $\widetilde{\beta}_1^{(1)}$ performs best among

the three corrected estimators for small β_1, but shrinks aggressively for moderate to large β_1 resulting in overcorrection. $\widetilde{\beta}_1^{(2)}$ shrinks much less dramatically resulting in undercorrection. $\widetilde{\beta}_1^{(3)}$ strikes a balance between the two. We notice that the magnitude of bias in all the estimators is less for case-control data than that for data collected prospectively. A possible explanation is that the standard error for data collected prospectively would tend to be bigger than that for data collected retrospectively, resulting in bigger bias. However rescaling of the axes demonstrates their inherent similarity in the "$\mu-$version". The second column of Figure 1 plots the means for the naïve and corrected estimators for β_2. The bias results for β_2 essentially follow the same pattern as β_1. Though less dramatic, the corrected estimators show greatly reduced bias for much of the range of β_1. The corrected estimators preserve their relative advantages over one another. As before, $\widetilde{\beta}_2^{(1)}$ reduces the bias appreciably for small β_1, whereas, $\widetilde{\beta}_2^{(2)}$ would be preferred for moderate to large β_1. $\widetilde{\beta}_2^{(3)}$ exhibits a more even performance across most of the range examined. The plots under the two sampling scenarios would match up with a rescaling of the axes.

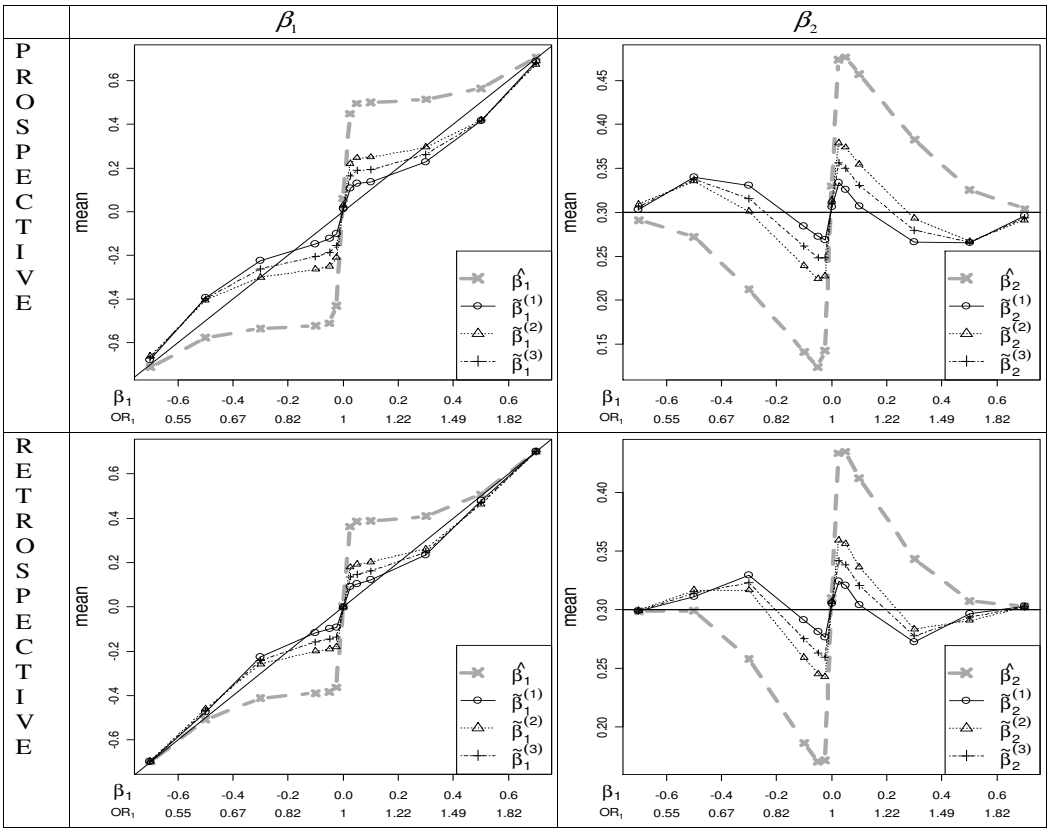


Figure 4.1: First column: expected values for the naïve and the conditional likelihood estimators of β_1. Second column: expected values for the naïve and the conditional likelihood estimators of β_2.

Mean squared error

The first column of Figure 4.2 shows the MSE values for the estimators of β_1. The naïve estimator exhibits extremely large MSE for most of the range considered. The corrected estimators offer marked improvement over the naïve estimator, especially for small β_1. The m.s.es of $\widetilde{\beta}_1^{(1)}$ and $\widetilde{\beta}_1^{(2)}$ are largely complementary, and $\widetilde{\beta}_1^{(3)}$ performs evenly across the range. The second column presents the corresponding plots for β_2. The MSE plots for β_2 preserve the relative merits of the corrected estimators compared to the naïve estimator. For small to moderate β_1, the corrected estimators show major

improvement over the naï estimator. At $\beta_1 = 0$, the MSE of the naïve estimator is more than twice the MSE for any of the corrected estimators, this is due to high variance of the naïve estimator at $\beta_1 = 0$. For moderate β_1 the naïve estimator has low variance but high bias, again resulting in higher MSE than that for corrected estimators. For all the sampling scenarios, $\widetilde{\beta}_1^{(3)}$ represents a reasonable choice for most β_1 values examined here.

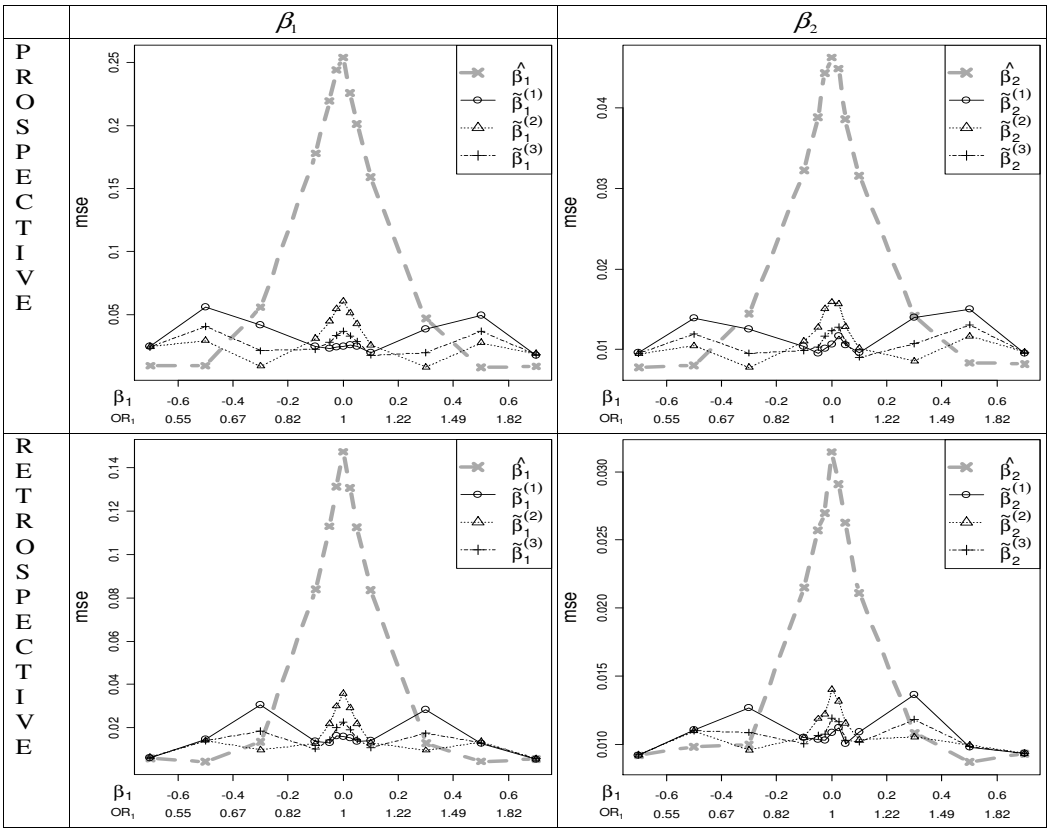


Figure 4.2: First column: mean squared errors for the naïve and the conditional likelihood estimators of β_1. Second column: mean squared errors for the naïve and the conditional likelihood estimators of β_2.

4.3.2 Gene-environment interaction

Figure 4.3 displays the simulation results for the gene-environment interaction effect. The top row of Figure 4.3 plots the means for the the naïve estimator and the bias-reduced estimator, $\widetilde{\beta}_2^{(1)}$, against β_g (with corresponding OR values). The left plot corresponds to the effect of the gene on disease risk, β_g, and the right plot is for the gene-environment interaction effect. The naïve estimators show extremely large bias specially for moderate values of β_g. $\widetilde{\beta}_2^{(1)}$ reduces bias dramatically for most of the β_g values examined. Both the estimators are nearly unbiased for large β_g values. $\widetilde{\beta}_2^{(1)}$ tends to under-correct for small values of β_g while over-correcting for large values of β_g. Both the naïve estimator and the conditional m.l.e. are practically unbiased at $\beta_g = 0$. The bias occurs in opposite direction for the gene effect and the interaction effect. The corresponding MSE plots are shown in the bottom row of Figure 4.3. The MSE plots for the gene effect and the interaction effect are very similar in pattern. The naïve estimator shows extremely large MSE for most values of β_g. $\widetilde{\beta}_2^{(1)}$ has remarkably improved MSE for most of the range considered.

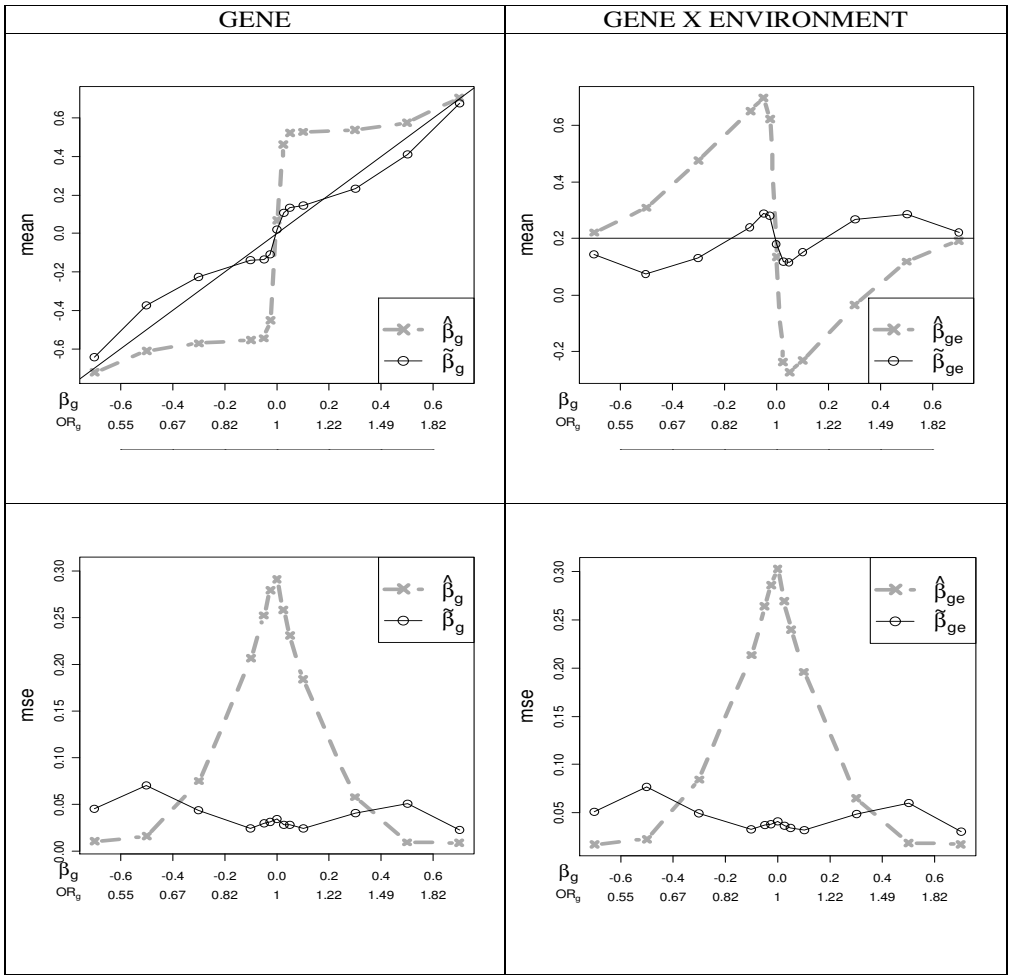


Figure 4.3: Top row: expected values for the naïve and the conditional likelihood estimators of β_g and β_{ge}. Bottom row: mean squared errors for the naïve and the conditional likelihood estimators of β_g and β_{ge}.

4.4 Discussion

We have investigated significance bias for secondary effects and we have proposed an approach to obtain corrected estimates with remarkably less bias and mean-squared error than the naïve estimates. We recommend $\widetilde{\beta}_2^{(3)}$ since it exhibits an even performance for all β_1 values considered. However, all three corrected estimators show substantial improvement over the naïve estimator. We have described a valid secondary analyses method for case-control data. We have explored the nature of significance bias for gene-environment interaction effect and have clarified conditions under which the estimate of the interaction effect would not be affected by significance bias.

Secondary analyses are very common for GWAS where data are collected on a variety of phenotypic traits, both quantitative and qualitative. For estimation of genetic effects of a SNP significant for the disease on additional secondary phenotypes, significance bias would produce biased estimates of effect sizes. The magnitude of bias would depend on the correlation between the estimates of the disease risk effect and the effect sizes of the significant SNP on the secondary phenotypes. For our particular simulation setup the correlation was modest, hence the results were not as dramatic as those for disease risk effect, β_1. But even for such modest correlation, the MSEs for the corrected estimators were a few times less than that for the naïve estimator, specially for small disease risk effect sizes.

For genome scans data are usually collected retrospectively, being cost-effective and an efficient use of resources. But case-control samples are not representative of the

population, resulting in selection bias. The semiparametric approach of Lee et al. (1997) is applicable only if the sampling fractions for the two stratum are known, which in most cases are not. The likelihood approach by Lin and Zeng (2008) can be performed for both binary and continuous secondary phenotype, but the joint modeling involves a slightly restrictive model, referred to as "conditional" model (Lee et al. 1997), where the odds ratio between the disease phenotype and the secondary phenotype given the genotype is independent of the genotype. Also, the marginal distribution of D given g is not logistic. The bivariate logistic model formulation is a fully-parametrized model and lends itself to the common belief that the the marginal distribution of the disease status variable given the genotype is logistic.

Inference on β_2 is not complete without a confidence interval. Standard confidence interval(CI) procedures, carried out without acknowledging the selection of the SNP based on significance, perform very poorly in this setting. To see this, we revert to the μ-version of the problem. After conditioning on significance, a standard 95% CI for μ_1 cannot contain 0. Thus, when the null hypothesis $\mu_1 = 0$ is true, the coverage of standard CI for μ_1 is 0, which in turn throws the standard CI preocedures for $\boldsymbol{\mu}_{-1}$ off balance. Let us consider the bivariate version of the problem,

$$\begin{pmatrix} Z_1 \\ \\ Z_2 \end{pmatrix} \sim N_2 \left(\begin{pmatrix} \mu_1 \\ \\ \mu_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \\ \rho & 1 \end{pmatrix} \right) .$$

We wish to find a $1-\eta$ CI for μ_2 taking into account that $|z_1| > c$ has been already observed. The simplest CI for μ_2 will be $\tilde{\mu}_2 \pm q_{1-\frac{\eta}{2}} \sqrt{1-\rho^2}$ where $\tilde{\mu}_2 = z_2 - \rho(z_1 - \tilde{\mu}_1)$,

and $q_{1-\frac{\eta}{2}}$ is the $1-\frac{\eta}{2}$ quantile of the standard normal distribution. Alternatively, we can construct a profile likelihood for μ_2, normalize it to be a proper density, then take its $\frac{\eta}{2}$ and $1-\frac{\eta}{2}$ quantiles to be the upper and lower confidence limits for μ_2. We can also integrate the joint likelihood over μ_1, normalize it to be a proper density, and then take its $\frac{\eta}{2}$ and $1-\frac{\eta}{2}$ quantiles to be the upper and lower confidence limits for μ_2. We plan to apply these CI procedures to the simulation setup described earlier and judge their performance. Although we believe that these procedures will perform well, a more principled confidence interval construction for μ_2 is worth investigating.

Our current simulations cover only the dominant mode of inheritance. For additive model, the retrospective likelihood would involve two nuisance parameters and stability of the estimates obtained from standard optimization techniques might be an issue. Also, we have always considered models with only one parameter for the genetic effect. Future work extending our approach to a more general setup would be important. It would be interesting to investigate whether joint modeling of the correlated phenotypes is better than the marginal logistic regression even when both yield valid estimates of the population parameters of interest.

Chapter 5

Variable Selection via a Conditional Likelihood-based Penalty

5.1 Introduction

Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{5.1}$$

where the response $\mathbf{y}$ is a nx1 vector and the design matrix $\mathbf{X}$ is of order nxp. So the data consists of $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$ where y_i is the response and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$ is the vector of predictor values for the i^{th} observation in the sample. Let ϵ_i, $i = 1, \ldots, n$ be independently and identically distributed as $N(0, \sigma^2)$. We assume, without loss of generality, that the predictors are standardized and the response is centered so that $\sum_i x_{ij} = 0$, $\sum_i x_{ij}^2 = 1$, $j = 1, \ldots, p$, and $\sum_i y_i = 0$.

In the linear regression setup, our goal is to find a linear model that provides a concise description of how the predictors affect the response. The model selection problem entails selecting variables that might best describe that relationship, and estimating the coefficients corresponding to those variables. With the simultaneous advent of high-speed computing and high-throughput technologies, most of the recent research problems involve datasets with large number of predictors, especially high-dimensional datasets with fewer observations than predictors. For example, a typical gene expression data has tens of thousands of genes (predictors) and only a few hundred arrays (observations). High-dimensional data arise in various fields of scientific research including computational biology, finance, biomedical imaging, satellite imagery, and many others. High-dimensional datasets present a challenge to traditional methods of model selection and underline the importance of model selection techniques.

We usually judge the usefulness of a predictive model on the basis of prediction accuracy and interpretability. Prediction accuracy is a quantitative measure for assessing model fit. We can calculate the expected prediction error, also known as *test* error or *generalization* error (Hastie et al. 2001), of a regression fit $\mathbf{X}\hat{\boldsymbol{\beta}}$ at $\mathbf{X} = \mathbf{x}_0'$,

$$\begin{aligned} EPE(x_0) &= E\left[(\mathbf{x}_0'\boldsymbol{\beta} + \epsilon_0 - \mathbf{x}_0'\hat{\boldsymbol{\beta}})^2\right] \\ &= \sigma_\epsilon^2 + \left[E(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) - \mathbf{x}_0'\boldsymbol{\beta}\right]^2 + E\left[\mathbf{x}_0'\hat{\boldsymbol{\beta}} - E(\mathbf{x}_0'\hat{\boldsymbol{\beta}})\right]^2 \\ &= \sigma_\epsilon^2 + Bias^2 + Variance\ . \end{aligned}$$

The first term is the irreducible error part and it cannot be avoided even if $\boldsymbol{\beta}$ was

known. The second and the third components can be controlled and they make up the *mean squared error* of $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$. The second term is the squared bias: the amount by which the average of the estimate differs from $\mathbf{x}_0'\boldsymbol{\beta}$; the third term is the variance: the expected squared deviation of $\mathbf{x}_0'\hat{\boldsymbol{\beta}}$ around its mean. We generally try to minimize the mean squared error or expected prediction error for a model. Interpretability of a model, on the other hand, is more qualitative in nature, and involves discerning which variables play an important role in predicting the response.

Ordinary least squares (OLS) minimizes the residual sum of squares

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 .$$

It is intuitively appealing, but OLS fitting does not always provide a satisfactory model in terms of prediction accuracy and interpretability. It produces best linear unbiased estimators, but the variance of the predicted values is often high. The interpretability of the model is also seriously hampered since OLS retains all the predictors. With too many variables in the model, it is difficult to understand which variables are really important in predicting the response. Moreover, in the high-dimensional setting, it is not possible get an OLS solution, there being no unique solution to the system of linear equations involving the coefficients.

Traditional approaches to model selection, such as best subset regression or stepwise regression, retain a subset of the candidate predictors, eliminate the rest, and use OLS to estimate the coefficients corresponding to the ones retained. Subset selection

generally achieves better prediction accuracy than the full model by selecting a only a subset of the candidate predictors. The selection of the subset of variables is based on either best subset regression or forward/backward stepwise selection. Best subset regression is generally considered impractical for $p > 30$. Backward selection can only be used when $N > p$, while forward selection can always be used. Hybrid strategies using both forward and backward moves in each step can also be used. The best subset size or the best model among the sequence of models produced by each of the above procedures is the tuning parameter and typically the model that minimizes an estimate of the test error is chosen. Subset selection, though conceptually simple and produces easily interpretable models, has serious drawbacks. It is a discrete process, *i.e.*, either makes a coefficient zero or inflates it. This inherent discreteness makes subset selection extremely variable. It is not stable with respect to small perturbations in the data.

Ridge regression (Hoerl and Kennard 2000) (Hoerl and Kennard, 1970), on the other hand, retains all the predictors in the model and modifies how the coefficients are estimated. The ridge estimate $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to } \sum_j \beta_j^2 \leq t \ ,$$

where t is the tuning parameter. The above optimization problem can be equivalently expressed as,

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j \beta_j^2 \right\} \ .$$

There is one-to-one correspondence between t and λ. The tuning parameter is chosen to minimize an estimate of the expected prediction error. Ridge regression achieves better performance than OLS through a bias-variance trade-off. It is a continuous process and the ridge estimates are stable, *i.e.*, if we delete a single data point, the new ridge estimates, for the same tuning parameter will be close to the old. However, ridge estimates retain all the predictors in the model resulting in less interpretability. Ridge regression can be used in the high-dimensional setting.

Methods which select subsets and shrink estimates simultaneously would be more welcome as they would retain good features of both subset selection and ridge regression. Breiman (1995) proposed the nonnegative garrote for better subset regression. His procedure minimizes

$$\sum_{i=1}^{n}\left(y_i - \sum_j c_j \hat{\beta}_j^0 x_{ij}\right)^2 \qquad \text{subject to } \; c_j >= 0, \; \sum_j c_j \leq t \; ,$$

where $\hat{\beta}_j^0$ are the OLS estimates and $\hat{\beta}_j(t) = c_j(t)\hat{\beta}_j^0$ are the new estimates. The nonnegative garrote starts with the OLS estimates and then as we tighten the garrote, some of the coefficients are set to zero and the remaining ones are shrunk. Breiman showed via simulations that nonnegative garrote outperforms subset selection and is comparable to ridge regression unless the model has a large number of small effects. In terms of stability, nonnegative garrote is intermediate between subset selection and ridge regression. But it depends heavily on the OLS estimates, the nonnegative garrote estimates are expected to suffer in situations where the OLS estimates perform poorly,

and cannot be used when there are more predictors than samples.

Motivated by the idea of nonnegative garrote, Tibshirani (1996) proposed a new technique called LASSO: least absolute shrinkage and selection operator. The LASSO estimate $\hat{\boldsymbol{\beta}}$ is defined by

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to } \sum_j |\beta_j| \leq t \ .$$

LASSO can produce sparse solution. When there are a large number of candidate predictors parsimony is an important issue. LASSO can be implemented in the high-dimensional setting but it cannot select more variables than number of observations. Frank and Friedman (1993) introduced bridge regression and Fu (1998) developed a general approach to solve for bridge estimators. Bridge regression minimizes

$$\left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 \right\} \qquad \text{subject to } \sum_j |\beta_j|^q \leq t \ .$$

It includes subset selection ($q = 0$), LASSO ($q = 1$), and ridge regression ($q = 2$) as special cases. Huang et al. (2008) studied the asymptotic properties of bridge estimators for an increasing number of predictors.

Fan and Li (2001) proposed a non-convex penalty function, the smoothly clipped

absolute deviation (SCAD) penalty. The SCAD penalty function is defined as follows:

$$p_\lambda(|\beta_j|) = \begin{cases} \lambda|\beta_j|, & |\beta_j| \leq \lambda \\ -\frac{|\beta_j|^2 - 2a\lambda|\beta_j| + \lambda^2}{2(a-1)}, & \lambda < |\beta_j| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2}, & |\beta_j| > a\lambda \end{cases},$$

where a and λ are the tuning parameters and are chosen such that they minimize an estimate of the expected prediction error. They have shown via simulations that the choice $a \approx 3.7$ works quite well for various variable selection problems. In the context of SCAD penalty they argued that a "good" penalty function should be unbiased, sparse, and continuous in data and showed that SCAD penalty possesses all three desirable properties. Fan and Peng (2004) discussed asymptotic properties of the non-concave penalized likelihood estimator when the number of covariates increase to infinity with the sample size. Zou and Li (2008) suggested one-step sparse estimates based on local linear approximation (LLA) for maximizing the penalized likelihood for concave penalty functions.

All the methods described above can be viewed as applying different penalty functions to the OLS criterion and can be regarded as penalized least squares procedures. A form of the penalized least squares objective function is

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \, . \tag{5.2}$$

Minimizing (5.2) with respect to $\boldsymbol{\beta}$ gives penalized least squares estimator of $\boldsymbol{\beta}$. The

penalty functions $p_\lambda(|\beta_j|)$ do not have to be the same for all j and are allowed to depend on λ. The L_2 penalty function $p_\lambda(|\beta_j|) = \lambda|\beta_j|^2$ corresponds to ridge regression, while the L_1 penalty function $p_\lambda(|\beta_j|) = \lambda|\beta_j|$ to LASSO. The L_0 or entropy penalty, $p_\lambda(|\beta_j|) = \lambda I(|\beta_j| \neq 0)$, corresponds to variable subset selection. Nonnegative garrote, also, can be expressed as penalized least squares with additional sign constraints,

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j \frac{|\beta_j|}{|\hat{\beta}_j^0|} \right\} \qquad \text{subject to } \beta_j \hat{\beta}_j^0 \geq 0 \ \ \forall j.$$

When the columns of X are orthonormal, the penalized least squares problem boils down to minimizing

$$(\hat{\beta}_j^0 - \beta_j)^2 + p_\lambda(|\beta_j|)$$

for each j separately. This simplification allows us to study the estimator as a function of the data. In this special case, the L_1 penalty function yields

$$\hat{\beta}_j = sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+ \ ,$$

where γ depends on λ. This is called 'soft threshold' estimator by Donoho and JOHNSTONE (1994) and is typically used in the wavelet analysis. The ridge solution for orthonormal X is,

$$\hat{\beta}_j = \frac{1}{1+\gamma} \hat{\beta}_j^0 \ .$$

The nonnegative garrote estimate is

$$\hat{\beta}_j = \left(1 - \frac{\gamma}{\hat{\beta}_j^{0\,2}}\right)^+ \hat{\beta}_j^0 \ .$$

The hard thresholding penalty function $p_\lambda(|\beta_j|) = \lambda^2 - (|\beta_j| - \lambda)^2 I(|\beta_j| < \lambda)$ results in the hard thresholding rule

$$\hat{\beta}_j = \hat{\beta}_j^0 I(|\hat{\beta}_j^0| > \gamma) \ .$$

The hard thresholding rule is also derivable from the entropy penalty function. For orthonormal X the resulting SCAD estimator is given by

$$\hat{\beta}_j = \begin{cases} sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+, & |\beta_j| \leq 2\lambda \\ \left\{(a-1)\hat{\beta}_j^0 - sgn(\hat{\beta}_j^0)a\lambda\right\}/(a-2), & 2\lambda < |\beta_j| \leq a\lambda \\ \hat{\beta}_j^0, & |\beta_j| > a\lambda \end{cases} \ .$$

We have plotted these estimators as functions of the data in Figure (5.1). The hard thresholding penalty, corresponding to plot (a), satisfies sparsity and unbiasedness, but it is not continuous in data. Plot (b) shows that the ridge estimator is continuous, but it is neither a thresholding rule nor is it unbiased for large values of the parameter. The LASSO penalty corresponding to plot (d) suggests that LASSO satisfies sparsity and continuity but not unbiasedness. The SCAD rule, as we had mentioned before, satisfies all three properties and so does the nonnegative garrote.

The penalized least squares estimators are biased but their variance is smaller than the OLS estimator, thus with a little sacrifice of bias we can achieve better performance

on the average in terms of mean squared error or prediction error. Some of these penalty functions restrict the coefficients in such a way that a number of them are reduced to zero, thus effectively performing variable selection. The idea of applying penalty functions to the OLS criterion can be extended to penalized likelihood to encompass likelihood-based models. A form of the penalized log-likelihood is

$$\sum_{i=1}^{n} l_i - \sum_{j=1}^{p} p_\lambda(|\beta_j|) \ ,$$

where l_i is the log-likelihood for y_i. In the linear regression setup the penalized least squares and the penalized likelihood estimators are exactly the same for type 1 penalty functions defined by Zou and Li (2008), such as the bridge penalties or the logarithm penalty.

The penalized likelihood estimators can be interpreted from a Bayesian point of view. The penalty functions can be thought of as log-prior densities for the parameters. Thus LASSO can be viewed as Bayes posterior mode under independent Laplace priors and ridge estimate can be interpreted as mode of the posterior distribution with independent Gaussian priors for the parameters. Since the posterior distribution is Gaussian, the ridge estimator is also the posterior mean. The SCAD penalty corresponds to an improper prior. Thus LASSO, ridge, and SCAD are all Bayes estimates but with different prior distributions for the regression coefficients (Hastie et al. 2001; Tibshirani 1996).

Efron et al. (2004) proposed least angle regression, and its LASSO and forward

stagewise variations (LARS). Their paper describes a new model selection algorithm, simple modifications of which give LASSO and forward stagewise regression. It has revolutionized the way LASSO is implemented, implementation being a natural concern for very large number of predictors. Although LASSO is a very appealing variable selection tool, it has a few drawbacks. In high-dimensional scenario, LASSO does not perform satisfactorily, it cannot choose more variables than number of samples. Also, if there is a group of correlated variables among which the pairwise correlations are very high, then LASSO tends to choose any one variable from the group. In the usual regression setup, if the correlation between the predictors is high, ridge regression usually outperforms LASSO. In an attempt to retain good features of both ridge regression and LASSO, Zou and Hastie (2005) presented a new regularization and variable selection method, the elastic net, which is particularly useful when there are more variables than observations. The naïve elastic net criterion minimizes

$$\left\{\sum_{i=1}^{n}\left(y_i-\sum_j \beta_j x_{ij}\right)^2\right\}+\lambda_1\sum_j|\beta_j|+\lambda_2\sum_j\beta_j^2\ . \tag{5.3}$$

(5.3) is equivalent to the minimizing

$$\left\{\sum_{i=1}^{n}\left(y_i-\sum_j \beta_j x_{ij}\right)^2\right\} \qquad \text{subject to } (1-\gamma)\sum_j|\beta_j|+\gamma\sum_j\beta_j^2\le t\ ,\ \gamma=\frac{\lambda_2}{\lambda_1+\lambda_2}\in[0,1]\ ,$$

where the elastic net penalty $(1-\gamma)\sum_j|\beta_j|+\gamma\sum_j\beta_j^2$ is a convex combination of the lasso and the ridge penalty. The naïve elastic net is a two-step procedure: ridge-type shrinkage followed by lasso-type thresholding, double shrinkage does not help much

with variance while introducing extra bias. The elastic net estimator is a rescaled naïve elastic net estimator, that corrects for the double shrinkage and is defined as

$$\hat{\beta}(\text{elastic net}) = (1 + \lambda_2)\hat{\beta}(\text{naïve elastic net}) \ .$$

Elastic net exhibits the attractive property of "grouping effect": highly correlated predictors have the same regression coefficients. LASSO does not possess this property (Zou and Hastie 2005).

With high-dimensional data, there are two different objectives: ensuring high prediction accuracy and identifying the set of predictors with nonzero regression coefficients. Identifying the true "sparse pattern", referred to as variable selection consistency, is particularly important when the true underlying model is sparse. LASSO variable selection is not necessarily consistent. Hence, LASSO is not an oracle procedure (Fan and Li 2001; Zou 2006). We call a procedure, δ, an oracle procedure (Fan and Li 2001; Zou 2006) if $\hat{\boldsymbol{\beta}}(\delta)$ has the following properties asymptotically:

- identifies the right subset model, $\left\{j, \hat{\beta}_j \neq 0\right\} = \{j, \beta_j \neq 0\}$
- has the optimal estimation rate, $\sqrt{n}\left(\hat{\boldsymbol{\beta}}(\delta) - \boldsymbol{\beta}\right) \rightarrow_D N(\mathbf{0}, \Sigma^*)$, where Σ^* is the covariance matrix knowing the true subset model.

Zou (2006) proposed a variation of lasso, adaptive LASSO, which minimizes

$$\hat{\boldsymbol{\beta}} = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \sum_j \beta_j x_{ij} \right)^2 + \lambda \sum_j \hat{w}_j |\beta_j| \right\} \ ,$$

where $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$, $\gamma > 0$, and $\hat{\boldsymbol{\beta}}$ is a root-n consistent estimator of $\boldsymbol{\beta}$. Adaptive LASSO is fast, effective, and enjoys the oracle property. By incorporating data-dependent weights adaptive LASSO manages to reduce the bias of LASSO. The weights make the resulting estimator nearly unbiased when the true unknown parameter is large. Huang et al. (2007) studied the asymptotic properties of adaptive LASSO for sparse high-dimensional regression models when number of covariates increase with the sample size. When the number of predictors exceeds the sample size they show that under the partial orthogonality condition adaptive LASSO is an oracle procedure if marginal regression is used to obtain the initial estimator.

Yuan and Lin (2006) studied a slightly different problem of selecting grouped variables (factors) for achieving better prediction accuracy in regression problems where interest lies in finding important explanatory factors for the response variable. They extended LASSO, LARS, and nonnegative garrote to group LASSO, group LARS, and group nonnegative garrote for factor selection. Wang et al. (2007) developed group SCAD regression in the same spirit. Candes and Tao (2007) proposed the Dantzig selector. Wasserman et al. (2007) advocated a three-stage procedure: in the first stage a set of candidate models: LASSO, marginal regression, and forward stepwise regression, are fitted to the data, in the second stage one of them is selected by cross-validation, and in the third stage hypothesis testing is used to eliminate some of the variables. The first two stages are referred to as "screening" and the third one is the "cleaning" stage. Radchenko and James (2008) described a modification to LASSO to prevent overshrinkage of LASSO by using two tuning parameters, one for selecting variables and the other to

control the amount of shrinkage. They call it VISA: Variable Inclusion and Shrinkage Algorithms. Fan and Lv (2006) introduced the concept of sure independence screening in ultrahigh-dimensional problems and described a sure screening method based on a correlation learning, called the Sure Independence Screening (SIS). James *et al.* (2009) explored the relationship between LASSO and Dantzig selector and described a new algorithm, DASSO, which uses a LARS-type algorithm to compute the entire solution path for the Dantzig selector. There are many other methods in the model selection literature that have been proposed for simultaneous variable selection and coefficient shrinkage.

We present a method that involves the two principal components of model selection: variable selection and estimation of the coefficients corresponding to the selected variables. To select a variable, we test whether the regression coefficient corresponding to that variable is zero or not, based on the observed test coefficient. Then we estimate the regression coefficient based on a conditional likelihood that takes into account the result of the testing of hypothesis. We incorporate the information of whether the variable was found to be significant or not while constructing the conditional likelihood for estimation. This idea is an application of the conditional likelihood approach (Ghosh et al. 2008)) to overcoming the "winner's curse" (Lohmueller et al. 2003; Zöllner and Pritchard 2007), or "significance bias" (Ghosh et al. 2008)), in genome-wide association studies. The conditional likelihood suggests a non-convex penalty function that can be used in the penalized likelihood framework for coefficient shrinkage. Thus, our proposed penalty function has a natural motivation based on the selection procedure

involving the test coefficient. We call the resulting method Test Coefficient Shrinkage or TCS.

With focus on ensuring high prediction accuracy, we describe a penalization technique based on the TCS penalty in the linear regression framework. We extend our method to high-dimensional regression problems. We illustrate the performance of our approach via extensive simulations. We use a real data example that has been widely used in the model selection literature to compare the performances of different methods. We judge the performance of TCS and other popular methods such as LASSO, ridge, and SCAD using simulations for both the usual scenario where we have more observations than covariates and the high-dimensional setup with more predictors than observations. The simulations cover a wide range of models.

5.2 Methods

Consider the linear regression model (5.1). Our goal is to find the best linear fit to the data in terms of prediction error. With this objective in mind, we propose a shrinkage method based on penalized likelihood. In the linear regression setup the penalized log-likelihood assumes the form

$$l_{penalized}(\boldsymbol{\beta}, \sigma^2) = -n \log \sigma - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) - \sum_j p_\lambda(|\beta_j|) \ . \tag{5.4}$$

We suggest a novel penalty function, TCS penalty,

$$p_\lambda(|\beta_j|) = \log\left[\Phi(-\lambda - \mu_j) + \Phi(-\lambda + \mu_j)\right] \ ,$$

where we define $\mu_j = \frac{\beta_j}{SE(\hat{\beta}_j^0)}$,. The motivation for this penalty function stems from accounting for the testing of a regression coefficient to select it. Let us consider the situation where we have only one predictor x,

$$y_i = \beta x_i + \epsilon_i, \ \ i = 1, \ldots, n \ ,$$

and we assume that the error variance σ^2 is known. Our goal is to build a model for y. We first test whether x has any predictive ability, that is, we test the null hypothesis $H_0 : \beta = 0$ against the alternative $H_1 : \beta \neq 0$ on the basis of the test statistic $Z = \frac{\hat{\beta}^0}{SE(\hat{\beta}^0)} = \frac{\hat{\beta}^0}{\sigma}$. We reject H_0 if the observed Z is greater in magnitude than some prespecified quantity λ. If $H_0 : \beta = 0$ is accepted, then we predict y using $\bar{y}$. If we reject the null, we need an estimate for β to be able to predict y. We construct a conditional likelihood for β which takes into account the fact that the null has been rejected. We note that $Z \sim N(\mu, 1)$, where $\mu = \frac{\beta}{\sigma}$. The conditional likelihood for μ is

$$\begin{aligned} L_c(\mu) &= \frac{p_\mu(z)}{P(|Z| > \lambda)} \\ &= \frac{\phi(z - \mu)}{\Phi(-\lambda - \mu) + \Phi(-\lambda + \mu)} \ . \end{aligned} \tag{5.5}$$

We maximize $L_c(\mu)$ with respect to μ to derive $\tilde{\mu}$, the conditional maximum likelihood estimate of μ. $\tilde{\mu}$ may be regarded as a penalized likelihood estimator with the TCS penalty $\log\left[\Phi(-\lambda-\mu)+\Phi(-\lambda+\mu)\right]$ since

$$\begin{aligned}\tilde{\mu} &= \arg\max_{\mu} L_c(\mu) \\ &= \arg\max_{\mu} \left\{\log\phi(z-\mu) - \log\left[\Phi(-\lambda-\mu)+\Phi(-\lambda+\mu)\right]\right\} .\end{aligned}$$

$\tilde{\mu}$ can be easily converted to an estimate for β using the one-to one correspondence $\tilde{\beta}=\tilde{\mu}\sigma$. We can alternatively think of $\tilde{\beta}$ as

$$\tilde{\beta} = \begin{cases} 0 & , \text{ accept } H_0 \\ \tilde{\mu}\sigma & , \text{ reject } H_0 \end{cases} . \tag{5.6}$$

Thus $\tilde{\beta}$ is a thresholding rule shrinking the estimate $\hat{\beta}^0$ to zero if we accept the null hypothesis and shrinking it to some non-zero value if the null is rejected, the amount of shrinkage being determined by λ. Larger the value of λ, the greater is the shrinkage. The rule $\tilde{\beta}$ is not continuous in the observed data $\hat{\beta}^0$. If we believe continuity of a penalty function to be a desirable property as advocated by Fan and Li (2001), we can make the rule $\tilde{\beta}$ continuous in $\hat{\beta}^0$ by defining $\tilde{\beta}$ as

$$\tilde{\beta} = \tilde{\mu}\sigma . \tag{5.7}$$

Thus, if we accept the null hypothesis, instead of shrinking the estimate $\hat{\beta}^0$ all the way to zero, the new rule shrinks it to a value very close to zero. With this new definition, $\tilde{\beta}$ loses its natural motivation as a test coefficient shrinkage and is no longer a thresholding rule, thus not engaging in variable selection anymore. We can consider $\tilde{\beta}$ as a shrinkage estimator of β. We plot the TCS estimator $\tilde{\beta}$ as a function of $\hat{\beta}^0$ using both the thresholded definition (5.6) and the non-thresholded (5.7) one.

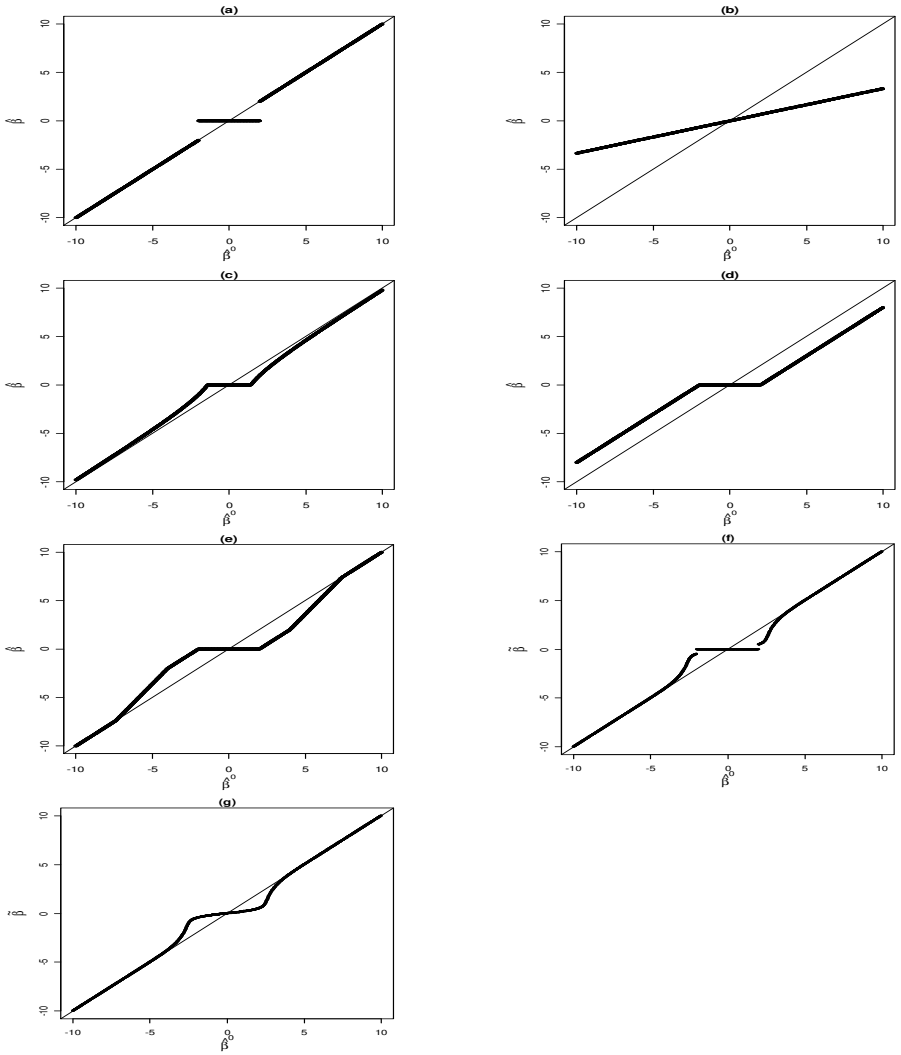


Figure 5.1: Plots of shrinkage estimators as function of data for (a) hard thresholding rule, (b) ridge, (c) nonnegative garrote, (d) LASSO, (e) SCAD, (f) thresholded TCS, and (g) non-thresholded TCS where X is orthonormal, $\lambda = 2$, and a=3.7

We use the TCS penalty function $\log\left[\Phi(-\lambda-\mu)+\Phi(-\lambda+\mu)\right]$ in the penalized likelihood framework to obtain shrinkage estimates of regression coefficients. There is a Bayesian interpretation to our proposed TCS penalty.

$$\log\phi(z-\mu)-\log\left[\Phi(-\lambda-\mu)+\Phi(-\lambda+\mu)\right]$$

can be thought of as log posterior density for μ, where the prior for μ is $p(\mu)\propto\frac{1}{[\Phi(-\lambda-\mu)+\Phi(-\lambda+\mu)]}$. This results in considering $\tilde{\mu}$ as Bayes posterior mode. Substituting the TCS penalty in (5.4) we get

$$\begin{aligned} l_{penalized}(\boldsymbol{\beta},\sigma^2) &= -n\log\sigma-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &\quad -\sum_j \log\left[\Phi(-\lambda-\mu_j)+\Phi(-\lambda+\mu_j)\right]\ . \end{aligned} \tag{5.8}$$

To obtain penalized maximum likelihood estimate of $\boldsymbol{\beta}$ for a given λ, we maximize (5.8) with respect to $(\boldsymbol{\beta},\sigma^2)$. We choose λ to minimize an estimate of the expected prediction error.

5.2.1 High-dimensional setup

We would like to apply our method in the high-dimensional case. For $p>n$ setup, the direct application of the penalized likelihood with the TCS penalty is not feasible since the penalty term involves standard error of OLS estimates, and OLS estimates are not defined in the high-dimensional situation. But we can apply our penalty if we consider

a subgroup of variables less than the number of observations where we can find OLS estimates. In particular, if we regress on a single predictor then we can obtain shrinkage estimate for the regression coefficient by applying our penalty. Any standard software would give us $\hat{\beta}^0$ and $\hat{SE}(\hat{\beta}^0)$ from a univariate regression. We define $Z = \frac{\hat{\beta}^0}{\hat{SE}(\hat{\beta}^0)}$. Using the well-known asymptotic result $\frac{\hat{\beta}^0 - \beta}{\hat{SE}(\hat{\beta}^0)} \xrightarrow{d} N(0,1)$, we have the approximate asymptotic result $Z \dot{\sim} N(\mu, 1)$, where $\mu = \frac{\beta}{\hat{SE}(\hat{\beta}^0)}$ is not exactly a parameter. We can implement either the thresholded or the non-thresholded version of TCS penalty. The thresholded estimate would be

$$\tilde{\beta} = \begin{cases} 0 & , \quad |z| \leq \lambda \\ \tilde{\mu}\hat{SE}(\hat{\beta}^0) & , \quad |z| > \lambda \end{cases},$$

where $\tilde{\mu} = \arg\max_{\mu} \frac{\phi(z-\mu)}{\Phi(-\lambda-\mu)+\Phi(-\lambda+\mu)}$. The non-thresholded estimate would be $\tilde{\beta} = \tilde{\mu}\hat{SE}(\hat{\beta}^0)$.

We develop an iterative procedure where we apply this univariate regression idea with residuals as the response variable. This idea is similar to the coordinate-wise descent algorithms (Friedman et al. 2007) for convex optimization problems. At each step of the iterative procedure we start with an initial estimator, update it one regression coefficient at a time, and then repeat this process with the updated estimate as the initial value for the next iteration. Let $\widetilde{\boldsymbol{\beta}}^{(k-1)}$ be the initial estimator of $\boldsymbol{\beta}$ at the k^{th} step of the procedure. We use the subscript $-j$ to signify that the j^{th} column or component is left out. For the j^{th} predictor we regress $\mathbf{y} - \mathbf{X}_{-j}\widetilde{\boldsymbol{\beta}}_{-j}$ on $\mathbf{x}_j$, get $\tilde{\beta}$ by shrinking $\hat{\beta}^0$, and replace $\widetilde{\beta}_j^{(k-1)}$ by $\tilde{\beta}$. We then move on to the next predictor. After we are finished

with all the predictors we have $\widetilde{\boldsymbol{\beta}}^{(k)}$. We then start the $(k+1)^{th}$ step with $\widetilde{\boldsymbol{\beta}}^{(k)}$ as the initial estimator. Getting $\tilde{\beta}$ from $\hat{\beta}^0$ is very fast and the computation of the residuals is also quick because only one component gets updated each time which makes the whole iterative procedure very efficient. For the first step we define $\widetilde{\boldsymbol{\beta}}^{(0)}$ as a vector of $\tilde{\beta}$'s obtained by shrinking the marginal regression coefficients. For a particular step of the iteration we have to loop through all the predictors, one predictor at a time, but we need to decide on a order in which to loop through the variables. We enter the predictors in decreasing order of magnitude of the initial estimator for that iteration. We decided on this order so as to eliminate any random element in the process and end up with the same estimate of $\boldsymbol{\beta}$ every time we run the procedure for a particular dataset. Also, we need a stopping rule for the iterative procedure. Tseng (1988) has established that coordinate-wise algorithms for convex optimization problems converge to their optimal solution under separability of the penalty function. For TCS penalty, the iterative procedure does not enjoy such convergence properties since the penalty function is not convex. So we continue the iteration for 50 steps and then choose the $\widetilde{\boldsymbol{\beta}}$ which gives the minimum training error in the last 10 steps. This strategy is based on the empirical observation that the training and the test errors have similar paths over the iteration steps which led us to believe that training error can serve as a stopping rule criterion.

5.2.2 Numerical examples

We apply our method to a data on prostrate cancer widely used in the variable selection literature. The data comes from a study conducted by Stamey et al. (1989). The dependent variable is *lpsa*: level of prostate specific antigen in blood serum. The relevant covariates are a number of clinical measures in men about to receive a radical prostatectomy: *lcavol* (log cancer volume), *lweight* (log prostate weight), *age*, *lbph* (log of the amount of benign prostatic hyperplasia), *svi* (seminal vesicle invasion), *lcp* (log of capsular penetration), *gleason* (Gleason score), and *pgg45* (percent of Gleason scores 4 or 5). We first standardize the predictors to have 0 mean and unit variance. We randomly split the data into a training set of size 67 and a test set of size 30. We applied OLS, LASSO, ridge, SCAD, and TCS to compare how each method performs in finding the best linear fit to the data. Every method, other than OLS, involves a tuning parameter which is chosen to minimize an estimate of the prediction error based on 10-fold cross-validation. We follow a "one-standard error" rule, in which the least complex model is chosen whose estimated prediction error is one standard deviation above the minimum estimated prediction error. This conservative approach follows from the thought that prediction error is estimated with some error. The final chosen model is then applied to the test set to assess its prediction error.

We use a simulation study to compare TCS with OLS, ridge, LASSO, and SCAD in the usual $n > p$ situation. We simulated 100 datasets consisting of n observations from the model

$$\mathbf{y}^{n\text{x}1} = \mathbf{X}^{nxp}\boldsymbol{\beta}^{p\text{x}1} + \boldsymbol{\epsilon}^{n\text{x}1} \ , \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n) \ ,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)'$. The columns of X and $\boldsymbol{\epsilon}$ are standard normal. The correlation between $\mathbf{x}_i$ and $\mathbf{x}_j$ is $\rho^{|i-j|}$ with $\rho = 0.5$. This numerical example is used in several publicationsFan and Li (2001); Tibshirani (1996); Zou and Hastie (2005) to discuss relative merits of different variable selection and shrinkage procedures. First we choose $\sigma = 3$ and $n = 40$. Then we reduce σ to 1 and finally increase the sample size to 60. We use 5-fold cross-validation to choose the complexity parameter. The mean squared error of each procedure is compared relative to that of OLS. We use the median of the relative mean squared errors over 100 datasets (MRMSE) to compare performance of different methods. We have also compared the performance of an oracle estimator to OLS.

To judge the performance our proposed method in the $p > n$ situation, we simulate data from the same linear model but with fewer observations than predictors. We set $n = 100$, $p = 1000$, and $\sigma = 1$. The non-zero β's constitute a random sample from normal distribution with mean zero and variance σ_β^2. We examine the performance of TCS, both the thresholded and the non-thresholded versions, and compare it with LASSO and ridge regression over a range of simulation setups, generated by varying the number of non-zero predictors p_1 from 5 to 1000 and σ_β from 0.1 to 2. LASSO is known to perform well in situations where we have a few big predictors and ridge usually performs better than LASSO in cases where we have many small predictors. Thus, in our simulation setup, we cover the two extreme situations of having very few big predictors and many small ones. Through our numerical exercise we plan to verify the empirical observations about LASSO and ridge regression and also hope to

understand in which cases our method works better than either of the two or in which cases it fails to perform well. For each simulation setup we compute the estimated test error for all three methods over 100 replications and judge their relative performance on the basis of the average test error. For each replication we have a training set of size n to fit the model over a range of values of the tuning parameter, a validation set of size n on the basis of which we decide on the value of the tuning parameter, and a test set of size 10,000 to estimate the test error of the fitted model. We standardize the covariates and center the response variable before analysis.

5.3 Results

Table 5.1 shows the results for the prostate cancer data for different variable selection and shrinkage methods. We see that ridge regression reduces OLS test error only by a small margin whereas LASSO offers remarkable improvement over OLS. Test error for SCAD is slightly lower than that for LASSO. Our proposed penalty has the lowest test error. If we compare the standard errors of the test error estimates for different methods, TCS penalty has the smallest.

Table 5.2 shows the results for the simulations in the $n > p$ setup. The median of the relative mean squared errors over 100 datasets (MRMSE) are reported in Table 5.2.

When the noise level is high and sample size is small, *i.e.*, $\sigma = 3$ and $n = 40$, LASSO performs the best but deteriorates quickly as signal to noise ratio increases, *i.e.*, as we decrease σ or increase n. When the noise level is reduced, the proposed shrinkage

Table 5.1: Estimated coefficients and test error results for prostate data

Term	OLS	Ridge	LASSO	SCAD	TCS
Intercept	2.480	2.469	2.479	2.484	2.478
lcavol	0.680	0.322	0.552	0.803	0.800
lweight	0.305	0.206	0.192	0.144	0.044
age	-0.141	-0.004	0.000	-0.005	0.000
lbph	0.210	0.135	0.035	0.039	0.024
svi	0.305	0.190	0.115	0.001	0.019
lcp	-0.288	0.061	0.002	-0.002	-0.004
gleason	-0.021	0.048	0.000	0.000	0.005
pgg45	0.267	0.109	0.003	0.000	0.021
Test Error	0.586	0.548	0.486	0.482	0.461
Std. Error	0.184	0.179	0.154	0.135	0.133

Table 5.2: Results for simulated numerical example in $n > p$ scenario

	MRMSE(%)		
Method	n=40, $\sigma = 3$	n=40, $\sigma = 1$	n=60, $\sigma = 1$
Oracle	41.49	62.70	68.23
Ridge	89.59	100.00	100.00
LASSO	81.18	86.34	85.72
SCAD	95.12	72.41	73.63
TCS	85.89	70.55	72.88

method has the lowest MRMSE and it performs as well as the oracle procedure as signal to noise ratio increases. Ridge regression performs extremely poorly. Performance of SCAD is comparable to our penalty function for high or moderate signal to noise ratio but for high noise level and small sample size our shrinkage method clearly outperforms SCAD. Table 5.2 suggests that the proposed penalty performs remarkably well and is indeed a worthy competitor.

Table 5.3 shows the results for TCS, LASSO, and ridge for the simulation in the

$p > n$ situation. The different columns of the table are for the number of non-zero predictors and the rows signify different values of σ_β. For each cell corresponding to a particular number of non-zero predictors and a value of σ_β we have recorded the average test error over 100 datasets for TCS, thresholded and non-thresholded versions of it, LASSO, and ridge in increasing order of magnitude. We have color-coded the different methods so that it is easy to visualize which method does best in a particular situation. The non-thresholded TCS is coded in cyan blue and the thresholded TCS in navy blue. LASSO is in red, ridge in green.

If we are interested in a particular method we can trace the corresponding color across the grid and its positions in various cells, in terms of first, second, third, or fourth, gives us its overall performance. For example, if we compare the red (LASSO) and the green (ridge) paths we observe that in situations where we have fewer non-zero predictors than samples LASSO does better than ridge, while ridge does better than LASSO when the number of non-zero predictors is greater than the sample size. In the latter situation, LASSO is at a disadvantage since it cannot choose more predictors than number of observations. In the case where there are equal number of non-zero predictors and observations, LASSO and ridge are very close, but LASSO does better when the non-zero coefficients are big whereas ridge outperforms LASSO when the non-zero coefficients are small in magnitude.

The first two columns of the table shows that thresholded TCS (navy blue) has the smallest test error among all the methods when $\boldsymbol{\beta}$ is truly sparse. But it performs poorly in situations where we have more non-zero predictors than samples. If we

compare it with LASSO we find that thresholded TCS has smaller average test error in very sparse situations but LASSO takes over as we increase the number of non-zero β's till we saturate $\boldsymbol{\beta}$ with all non-zero components. The difference in the performances of thresholded TCS and LASSO narrows down as we decrease σ_β.

The non-thresholded TCS never is the best choice other than for the situations where $\sigma_\beta = 0.1$ and number of non-zero variables is 50 and 100. But from this table we can definitely conclude that it is the best choice in terms of overall performance. It is almost always the second choice. In sparse situations it is practically the same as thresholded TCS and significantly better than either LASSO or ridge. When we have 50 non-zero predictors, LASSO usually has the smallest test error, but non-thresholded TCS is only second to it and as the β's become smaller in size it approaches LASSO and finally takes over for $\sigma_\beta = 0.1$. In the situation where there are more non-zero predictors than number of observations and ridge outperforms any other method, non-thresholded TCS is close behind and performs remarkably better than LASSO. It is only in the case where we have the same number of non-zero β's as the number of samples that non-thresholded TCS is unable to beat any of the two most popular methods, but the difference between them decreases as we make the β's smaller in magnitude and finally non-thresholded TCS outperforms both LASSO and ridge.

renewcommand21.5

Table 5.3: Simulation results for $p > n$

	Number of non-zero predictors					
σ_β	5	10	50	100	500	1000
2	1.11	1.19	117.50	335.24	1699.79	3379.85
	1.12	1.22	136.87	335.88	1843.06	3628.62
	1.50	2.24	147.90	352.74	1969.42	4019.68
	17.16	32.69	168.85	383.04	2014.58	4032.96
1	1.17	1.37	31.23	84.32	425.84	845.80
	1.19	1.41	36.39	84.85	460.77	906.27
	1.44	2.02	39.64	88.17	494.16	1004.66
	5.14	9.03	43.06	96.15	511.40	1006.20
0.5	1.21	1.50	9.24	22.09	107.34	212.31
	1.22	1.52	10.53	22.14	115.30	227.68
	1.33	1.70	11.12	22.85	123.88	252.74
	2.12	3.09	11.62	24.82	127.01	255.05
0.1	1.05	1.09	1.47	1.93	5.39	9.59
	1.05	1.10	1.49	1.96	5.60	10.08
	1.06	1.10	1.50	1.98	6.00	11.06
	1.15	1.19	1.53	1.99	6.04	11.12

5.4 Discussion

We have presented a model selection approach that can be used in the high-dimensional setting. Our approach can either behave like ridge regression and perform only coefficient shrinkage or it can perform variable selection and coefficient shrinkage simultaneously, depending on the problem that we are interested in. The method is based on the novel penalty that we propose: the TCS penalty. The motivation for this penalty comes from the testing of hypothesis of regression coefficients for selecting the corresponding predictors in the model. We have used this penalty in a penalized likelihood framework

in the $n > p$ situation. In the high-dimensional setup we have used it individually on each predictor and developed an iterative procedure. We have shown via numerical examples that our proposed method performs remarkably well in the $n > p$ situation. In the $p > n$ scenario, TCS has lower prediction error than the popular competing approaches, such as LASSO or ridge, when the true coefficient vector is extremely sparse, a situation where LASSO is known to dominate. When we have a large number of non-zero predictors, TCS performs significantly better than LASSO and is very close to ridge regression in terms of prediction error.

In the high-dimensional setting, we have judged the performance of the estimators in terms of prediction error. But the number of β-coefficients wrongly predicted as non-zero, better known as false-positives, or the number of β-coefficients wrongly predicted as zero, known as false-negatives, can serve as criteria for judging the performance of these methods. Obviously, ridge and the non-thresholded version of TCS select all the candidate predictors and these measures will not mean much in their cases. But they might be informative for comparing the performance of LASSO with the thresholded TCS.

In the usual $n > p$ setting we have used the TCS penalty in the penalized likelihood framework and implemented coefficient shrinkage, but the penalty is not a thresholding one and as a result does not participate in variable selection. We can implement a thresholded version of this by investigating each predictor individually as we did in the high-dimensional case. But we would have to deviate from the penalized likelihood

framework where we optimize a single objective function over all the coefficients simultaneously. For future work, we plan to modify the TCS penalty to be a thresholding rule.

In the high-dimensional case we describe an iterative procedure that examines each predictor separately. So, the method for the high-dimensional model is not an automatic extension of the method in the $n > p$ situation. For future work in this area it would be important to build an unified algorithm that can applied to both the situations. Also, it would be help us understand the method better if we are able to compare it with other high-dimensional model selection techniques in the regression setup. In most cases, it is difficult to so in the absence of published code. We plan to develop a code that we could distribute for everybody to use, That would also help us understand how our method performs in various situations and how does it compare with other techniques.

Appendix A

$$
\begin{aligned}
E(Z\big||Z|>c) &= K^{-1}\left[\int_{-\infty}^{-c} z\phi(z-\mu)dz + \int_{c}^{\infty} z\phi(z-\mu)dz\right], \\
&\qquad\qquad \text{where } K = \Phi(-c+\mu) + \Phi(-c-\mu) \\
&= \mu + K^{-1}\left[\int_{-\infty}^{-c-\mu} x\phi(x)dx + \int_{c-\mu}^{\infty} x\phi(x)dx\right], \quad x = z-\mu \\
&= \mu + K^{-1}\left[(2\pi)^{-1}\int_{\infty}^{\frac{1}{2}(c+\mu)^2} e^{-y}dy + (2\pi)^{-1}\int_{\frac{1}{2}(c-\mu)^2}^{\infty} e^{-y}dy\right], \\
&\qquad\qquad y = \frac{1}{2}x^2 \\
&= \mu + K^{-1}\left[(2\pi)^{-1}e^{-\frac{1}{2}(c-\mu)^2} - (2\pi)^{-1}e^{-\frac{1}{2}(c+\mu)^2}\right] \\
&= \mu + \frac{\phi(c-\mu) - \phi(c+\mu)}{K}
\end{aligned}
$$

Appendix B

B1. $E_{\boldsymbol{\mu}}\left(\mathbf{Z}|\ |Z_1|>c\right)=\boldsymbol{\mu}+\begin{pmatrix}1\\ \\ \boldsymbol{\rho}\end{pmatrix}\frac{\phi(c-\mu_1)-\phi(c+\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}$

$$E_{\boldsymbol{\mu}}\left(\mathbf{Z}|\ |Z_1|>c\right)=\begin{pmatrix}E_{\boldsymbol{\mu}}\left(Z_1||Z_1|>c\right)\\ \\ E_{\boldsymbol{\mu}}\left(\mathbf{Z_{-1}}|\ |Z_1|>c\right)\end{pmatrix}.$$

Now,

$$\begin{aligned}
& E_{\boldsymbol{\mu}}\left(Z_1||Z_1|>c\right)\\
= & \int_{\mathbf{z}_{-1}}\int_{|z_1|>c} z_1 L_c(\boldsymbol{\mu})dz_1 d\mathbf{z}_{-1}\\
= & \int_{|z_1|>c} z_1\frac{\phi(z_1-\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\ X\\
& \qquad\qquad \int_{\mathbf{z}_{-1}} N_{p-1}\left(\mathbf{z}_{-1};\boldsymbol{\mu}_{-1}+\boldsymbol{\rho}(z_1-\mu_1),R_{22}-\rho\rho'\right)d\mathbf{z}_{-1}dz_1\\
= & \int_{|z_1|>c} z_1\frac{\phi(z_1-\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}dz_1\\
= & \mu_1+\frac{\phi(c-\mu_1)-\phi(c+\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}(2.2)\ ,
\end{aligned}$$

and

$$
\begin{aligned}
& E_{\boldsymbol{\mu}}\left(\mathbf{Z}_{-1}|\ |Z_1|>c\right) \\
= & \int_{\mathbf{z}_{-1}}\int_{|z_1|>c} \mathbf{z}_{-1} L_c(\boldsymbol{\mu}) dz_1 d\mathbf{z}_{-1} \\
= & \int_{|z_1|>c} \frac{\phi(z_1-\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\ X \\
& \qquad \int_{\mathbf{z}_{-1}} \mathbf{z}_{-1} N_{p-1}\left(\mathbf{z}_{-1}; \boldsymbol{\mu}_{-1}+\boldsymbol{\rho}(z_1-\mu_1), R_{22}-\rho\rho'\right) d\mathbf{z}_{-1} dz_1 \\
= & \boldsymbol{\mu}_{-1} + \boldsymbol{\rho} \int_{|z_1|>c} (z_1-\mu_1) \frac{\phi(z_1-\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)} dz_1 \\
= & \boldsymbol{\mu}_{-1} + \boldsymbol{\rho} \frac{\phi(c-\mu_1)-\phi(c+\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\ .
\end{aligned}
$$

Hence

$$
E_{\boldsymbol{\mu}}\left(\mathbf{Z}|\ |Z_1|>c\right) = \boldsymbol{\mu} + \begin{pmatrix} 1 \\ \\ \boldsymbol{\rho} \end{pmatrix} \frac{\phi(c-\mu_1)-\phi(c+\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)}\ .
$$

B2. Proof of $\begin{cases} \widetilde{\mu}_1^{(2)} = \frac{\int \mu_1 L_c(\mu_1) d\mu_1}{\int L_c(\mu_1) d\mu_1} \\ \widetilde{\boldsymbol{\mu}}_{-1}^{(2)} = \mathbf{z}_{-1} - \boldsymbol{\rho}(z_1 - \widetilde{\mu}_1^{(2)}) \end{cases}$:

$$
\begin{aligned}
\widetilde{\mu}_1^{(2)} &= \frac{\int \mu_1 L_c(\boldsymbol{\mu}) d\boldsymbol{\mu}}{\int L_c(\boldsymbol{\mu}) d\boldsymbol{\mu}} \\
&= \frac{\int_{\mu_1} \mu_1 \frac{\phi(z_1-\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)} \int_{\boldsymbol{\mu}_{-1}} N_{p-1}\left(\mathbf{z}_{-1}; \boldsymbol{\mu}_{-1} + \boldsymbol{\rho}(z_1-\mu_1), R_{22} - \rho\rho'\right) d\boldsymbol{\mu}_{-1} d\mu_1}{\int_{\mu_1} \frac{\phi(z_1-\mu_1)}{\Phi(-c-\mu_1)+\Phi(-c+\mu_1)} \int_{\boldsymbol{\mu}_{-1}} N_{p-1}\left(\mathbf{z}_{-1}; \boldsymbol{\mu}_{-1} + \boldsymbol{\rho}(z_1-\mu_1), R_{22} - \rho\rho'\right) d\boldsymbol{\mu}_{-1} d\mu_1} \\
&= \frac{\int_{\mu_1} \mu_1 L_c(\mu_1) \int_{\boldsymbol{\mu}_{-1}} N_{p-1}\left(\boldsymbol{\mu}_{-1}; \mathbf{z}_{-1} - \boldsymbol{\rho}(z_1-\mu_1), R_{22} - \rho\rho'\right) d\boldsymbol{\mu}_{-1} d\mu_1}{\int_{\mu_1} L_c(\mu_1) \int_{\boldsymbol{\mu}_{-1}} N_{p-1}\left(\boldsymbol{\mu}_{-1}; \mathbf{z}_{-1} - \boldsymbol{\rho}(z_1-\mu_1), R_{22} - \rho\rho'\right) d\boldsymbol{\mu}_{-1} d\mu_1} \\
&= \frac{\int_{\mu_1} \mu_1 L_c(\mu_1) d\mu_1}{\int_{\mu_1} L_c(\mu_1) d\mu_1} ,
\end{aligned}
$$

and

$$
\begin{aligned}
\widetilde{\boldsymbol{\mu}}_{-1}^{(2)} &= \frac{\int \boldsymbol{\mu}_{-1} L_c(\boldsymbol{\mu}) d\boldsymbol{\mu}}{\int L_c(\boldsymbol{\mu}) d\boldsymbol{\mu}} \\
&= \frac{\int_{\mu_1} L_c(\mu_1) \int_{\boldsymbol{\mu}_{-1}} \boldsymbol{\mu}_{-1} N_{p-1}\left(\mathbf{z}_{-1}; \boldsymbol{\mu}_{-1} + \boldsymbol{\rho}(z_1-\mu_1), R_{22} - \rho\rho'\right) d\boldsymbol{\mu}_{-1} d\mu_1}{\int_{\mu_1} L_c(\mu_1) d\mu_1} \\
&= \frac{\int_{\mu_1} L_c(\mu_1) \int_{\boldsymbol{\mu}_{-1}} \boldsymbol{\mu}_{-1} N_{p-1}\left(\boldsymbol{\mu}_{-1}; \mathbf{z}_{-1} - \boldsymbol{\rho}(z_1-\mu_1), R_{22} - \rho\rho'\right) d\boldsymbol{\mu}_{-1} d\mu_1}{\int_{\mu_1} L_c(\mu_1) d\mu_1} \\
&= \mathbf{z}_{-1} - \boldsymbol{\rho}(z_1 - \widetilde{\mu}_1^{(2)}) .
\end{aligned}
$$

B3. Formulae for estimates of covariances between estimates of effect sizes for disease phenotype and binary secondary phenotype for prospective and retrospective study designs:

For data collected prospectively, we fit separate logistic regressions for D and Y,

$$\begin{aligned} logit\ P(D=1 \mid g) &= \alpha_1 + \beta_1 g \\ logit\ P(Y=1 \mid g) &= \alpha_2 + \beta_2 g\ . \end{aligned}$$

We need $\hat{cov}(\hat{\beta}_1, \hat{\beta}_2)$ to perform the bias correction. We begin with some basic results. Let $l(\theta)$ be any function of θ and $\dot{l}(\hat{\theta}) = \frac{\partial}{\partial\theta} l(\theta)\,|_{\theta=\hat{\theta}} = 0$. Using Taylor expansion of $\dot{l}(\hat{\theta})$ about the true parameter value θ,

$$\dot{l}(\theta) + (\hat{\theta} - \theta)\ddot{l}(\theta^*) = 0\ , \quad \text{where } |\theta - \theta^*| < |\theta - \hat{\theta}|\ .$$

Hence,

$$(\hat{\theta} - \theta) = -\left\{\ddot{l}(\theta^*)\right\}^{-1} \dot{l}(\theta)\ ,$$

or approximately,

$$(\hat{\theta} - \theta) = -\left\{\ddot{l}(\theta)\right\}^{-1} \dot{l}(\theta)\ .$$

If $l(\theta) = \sum \log f(data_i|\theta)$ then $(\hat{\theta} - \theta)$ is asymptotically equivalent to $Var(\hat{\theta})\dot{l}(\theta)$, where $Var(\hat{\theta}) = I(\theta)^{-1}$, $I(\theta)$ being the Fisher information matrix. Applying this for $\hat{\boldsymbol{\theta}}_1 = (\hat{\alpha}_1, \hat{\beta}_1)'$ and $\hat{\boldsymbol{\theta}}_2 = (\hat{\alpha}_2, \hat{\beta}_2)'$ we get

$$cov((\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_1), (\hat{\boldsymbol{\theta}}_2 - \boldsymbol{\theta}_2)) = Var(\hat{\boldsymbol{\theta}}_1) cov(\dot{l}_1(\boldsymbol{\theta}_1), \dot{l}_2(\boldsymbol{\theta}_2)) Var(\hat{\boldsymbol{\theta}}_2) \ ,$$

$$\dot{l}_1(\boldsymbol{\theta}_1) = \begin{pmatrix} \sum \left\{ d_i - \frac{\exp(\alpha_1+\beta_1 g_i)}{1+\exp(\alpha_1+\beta_1 g_i)} \right\} \\ \sum \left\{ d_i g_i - \frac{g_i \exp(\alpha_1+\beta_1 g_i)}{1+\exp(\alpha_1+\beta_1 g_i)} \right\} \end{pmatrix}, \text{ and } \dot{l}_2(\boldsymbol{\theta}_2) = \begin{pmatrix} \sum \left\{ y_i - \frac{\exp(\alpha_2+\beta_2 g_i)}{1+\exp(\alpha_2+\beta_2 g_i)} \right\} \\ \sum \left\{ y_i g_i - \frac{g_i \exp(\alpha_2+\beta_2 g_i)}{1+\exp(\alpha_2+\beta_2 g_i)} \right\} \end{pmatrix}.$$

Therefore $cov(\dot{l}_1(\boldsymbol{\theta}_1), \dot{l}_2(\boldsymbol{\theta}_2)) = \begin{pmatrix} \sum cov(d_i, y_i) & \sum g_i cov(d_i, y_i) \\ \sum g_i cov(d_i, y_i) & \sum g_i^2 cov(d_i, y_i) \end{pmatrix}$. We finally define $\hat{cov}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$ as,

$$\hat{cov}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) = \hat{Var}(\hat{\boldsymbol{\theta}}_1) \hat{cov}(\dot{l}_1(\boldsymbol{\theta}_1), \dot{l}_2(\boldsymbol{\theta}_2)) \hat{Var}(\hat{\boldsymbol{\theta}}_2) \ ,$$

where $\hat{cov}(\dot{l}_1(\boldsymbol{\theta}_1), \dot{l}_2(\boldsymbol{\theta}_2)) = \hat{cov}(d, y) \begin{pmatrix} n & \sum g_i \\ \sum g_i & \sum g_i^2 \end{pmatrix}$. Thus $\hat{cov}(\hat{\beta}_1, \hat{\beta}_2)$ would be the $(2,2)^{\text{th}}$ element of the matrix $\hat{cov}(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)$.

For case-control data we fit logistic regression for D, but for Y we have to consider the retrospective likelihood. Since we perform the selection based on the effect size estimate obtained from the logistic regression of D on g, for the bias correction calculations we need an estimate of the covariance between $\hat{\beta}_1$ obtained from fitting $logit\ P(D = 1 \mid g) = \alpha_1 + \beta_1 g$ and $\hat{\beta}_2$ obtained from fitting the retrospective log-likelihood $l = \sum \log P(Y = y_i, G = g_i | D = d_i)$. Let $\boldsymbol{\psi} = (\alpha_1, \beta_1)'$ and

$\boldsymbol{\eta} = (\beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3, p(g))'$. From earlier theoretical discussion we can conclude that

$$cov(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}) = Var(\hat{\boldsymbol{\psi}}) cov(\dot{g}(\boldsymbol{\psi}), \dot{l}(\boldsymbol{\eta})) Var(\hat{\boldsymbol{\eta}}) \ ,$$

where $g(\boldsymbol{\psi}) = \sum \log P(D = d_i | G = g_i)$. Then

$$\hat{cov}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}}) = \hat{Var}(\hat{\boldsymbol{\psi}}) \hat{cov}(\dot{g}(\boldsymbol{\psi}), \dot{l}(\boldsymbol{\eta})) \hat{Var}(\hat{\boldsymbol{\eta}}) \ .$$

Since any standard statistical software would provide us with $\hat{Var}(\hat{\boldsymbol{\psi}})$ and $\hat{Var}(\hat{\boldsymbol{\eta}})$ we only need to get $\hat{cov}(\dot{g}(\boldsymbol{\psi}), \dot{l}(\boldsymbol{\eta}))$. We use

$$\hat{cov}(\dot{g}(\boldsymbol{\psi}), \dot{l}(\boldsymbol{\eta})) = \sum_{i=1}^{n} \dot{g}_i(\hat{\boldsymbol{\psi}}) \dot{l}_i(\hat{\boldsymbol{\eta}})'$$

. Finally, we get $cov(\hat{\beta}_1, \hat{\beta}_2)$ as the $(2,3)^{\text{th}}$ element of the matrix $\hat{cov}(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\eta}})$.

B4. Proof of $cov(\hat{\beta}_G^{(s)}, \hat{\beta}_{GE}^{(b)}) = 0$:

The log-likelihood for the full-model is,

$$\begin{aligned} l &= \sum_{i=1}^{n} l_i \\ &= \sum_{i=1}^{n} \log P(D = d_i | G = g_i, E = e_i) \\ &= \sum_{i=1}^{n} \Big\{ d_i(\beta_0^{(b)} + \beta_G^{(b)} g_i + \beta_E^{(b)} e_i + \beta_{GE}^{(b)} g_i e_i) \\ &\qquad - \ln(1 + \exp(\beta_0^{(b)} + \beta_G^{(b)} g_i + \beta_E^{(b)} e_i + \beta_{GE}^{(b)} g_i e_i)) \Big\} \\ &= \sum_{i=1}^{n} \{ d_i r_i(\boldsymbol{\theta}; g_i, e_i) - \ln(1 + \exp(r_i(\boldsymbol{\theta}; g_i, e_i)) \} \\ &= \sum_{i=1}^{n} \{ d_i r_i - \ln(1 + \exp(r_i) \} \ . \end{aligned}$$

$\hat{\boldsymbol{\theta}}^{(b)}$ is obtained by solving the likelihood equation $\dot{l}(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} l = \mathbf{0}$. From multivariate Taylor expansion of the likelihood equation about the true parameter value $\boldsymbol{\theta}$ it follows that

$$\left(\hat{\boldsymbol{\theta}}^{(b)} - \boldsymbol{\theta}^{(b)} \right) = -\ddot{l}(\boldsymbol{\theta})^{-1} \dot{l}(\boldsymbol{\theta}) \ , \text{ where}$$

$$\dot{l}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \dot{l}_i(\boldsymbol{\theta}) \ , \ \dot{l}_i(\boldsymbol{\theta}) = \begin{pmatrix} d_i - \frac{\exp(r_i)}{1+\exp(r_i)} \\ g_i d_i - \frac{g_i \exp(r_i)}{1+\exp(r_i)} \\ e_i d_i - \frac{e_i \exp(r_i)}{1+\exp(r_i)} \\ g_i e_i d_i - \frac{g_i e_i \exp(r_i)}{1+\exp(r_i)} \end{pmatrix} \text{, and}$$

$$-\ddot{l}(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \ddot{l}_i(\boldsymbol{\theta}) = \sum_{i=1}^{n} \begin{pmatrix} c_i & g_i c_i & e_i c_i & g_i e_i c_i \\ g_i c_i & g_i^2 c_i & g_i e_i c_i & g_i^2 e_i c_i \\ e_i c_i & e_i g_i c_i & e_i^2 c_i & g_i e_i^2 c_i \\ e_i g_i c_i & e_i g_i^2 c_i & g_i e_i^2 c_i & g_i^2 e_i^2 c_i \end{pmatrix} \text{, where } c_i = \frac{\exp(r_i)}{(1+\exp(r_i))^2} \text{ .}$$

Hence,

$$\begin{aligned} \left(\hat{\beta}_{GE}^{(b)} - \beta_{GE}^{(b)}\right) = \frac{1}{\left|-\ddot{l}(\boldsymbol{\theta})\right|} \Bigg(& b_{41} \sum_{i=1}^{n} \left(d_i - \frac{\exp(r_i)}{1+\exp(r_i)} \right) \\ & + b_{42} \sum_{i=1}^{n} \left(g_i d_i - \frac{g_i \exp(r_i)}{1+\exp(r_i)} \right) \\ & + b_{43} \sum_{i=1}^{n} \left(e_i d_i - \frac{e_i \exp(r_i)}{1+\exp(r_i)} \right) \\ & + b_{44} \sum_{i=1}^{n} \left(g_i e_i d_i - \frac{g_i e_i \exp(r_i)}{1+\exp(r_i)} \right) \Bigg) \text{ ,} \end{aligned}$$

where b_{41}, b_{42}, b_{43}, and b_{44} are the elements of the 4^{th} row of the conjugate matrix of $-\ddot{l}(\boldsymbol{\theta})$. The analytical expressions for b_{41}, b_{42}, b_{43}, and b_{44} are as follows:

$$\begin{aligned} b_{41} \;=\; & \sum gc \sum e^2 c \sum g^2 ec + \sum g^2 c \sum ec \sum ge^2 c + \left(\sum gec\right)^3 \\ & - \sum gc \sum gec \sum ge^2 c - \sum g^2 c \sum e^2 c \sum gec - \sum gec \sum ec \sum g^2 ec \end{aligned}$$

$$\begin{aligned} b_{42} \;=\; & \sum c \sum gec \sum ge^2 c + \sum gc \sum e^2 c \sum gec + \left(\sum ec\right)^2 \sum g^2 ec \\ & - \sum c \sum e^2 c \sum g^2 ec - \sum gc \sum ec \sum ge^2 c - \sum ec \left(\sum gec\right)^2 \end{aligned}$$

$$
\begin{aligned}
b_{43} &= \sum c \sum gec \sum g^2ec + \left(\sum gc\right)^2 \sum ge^2c + \sum ec \sum g^2c \sum gec \\
&\quad - \sum c \sum g^2c \sum ge^2c - \sum gc \sum gec^2 - \sum ec \sum gc \sum g^2ec
\end{aligned}
$$

$$
\begin{aligned}
b_{44} &= \sum c \sum g^2c \sum e^2c + 2 \sum gc \sum gec \sum ec - \sum c \left(\sum gec\right)^2 \\
&\quad - \left(\sum gc\right)^2 \sum e^2c - \left(\sum ec\right)^2 \sum g^2c
\end{aligned}
$$

$\hat{\beta}_G^{(s)}$ is the estimate of the disease risk effect obtained from fitting a logistic regression of D on g. Hence $\hat{\beta}_G^{(s)}$ is defined as

$$
\hat{\boldsymbol{\theta}}^{(s)} = \begin{pmatrix} \hat{\beta}_0^{(s)} \\ \hat{\beta}_G^{(s)} \end{pmatrix} = arg\ max \sum_{i=1}^{n} \left\{ d_i(\beta_0^{(s)} + \beta_G^{(s)} g_i) - \ln(1 + \exp(\beta_0^{(s)} + \beta_G^{(s)} g_i)) \right\} .
$$

Hence,

$$
\left(\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}^{(s)}\right) = \frac{1}{a_1 h_1 - b_1^2} \begin{pmatrix} h_1 & -b_1 \\ -b_1 & a_1 \end{pmatrix} \begin{pmatrix} \sum \left\{ d_i - \frac{\exp(\beta_0^{(s)} + \beta_G^{(s)} g_i)}{1 + \exp(\beta_0^{(s)} + \beta_G^{(s)} g_i)} \right\} \\ \sum \left\{ d_i g_i - \frac{g_i \exp(\beta_0^{(s)} + \beta_G^{(s)} g_i)}{1 + \exp(\beta_0^{(s)} + \beta_G^{(s)} g_i)} \right\} \end{pmatrix} ,
$$

where $a_1 = \sum \frac{\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}{1+\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}$, $b_1 = \sum \frac{g_i \exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}{1+\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}$, and $h_1 = \sum \frac{g_i^2 \exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}{1+\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}$.

This gives us

$$\begin{aligned}\left(\hat{\beta}_G^{(s)} - \beta_G^{(s)}\right) &= \frac{a_1}{a_1h_1 - b_1^2}\sum\left\{d_ig_i - \frac{g_i\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}{1+\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}\right\} \\ &\quad -\frac{b_1}{a_1h_1 - b_1^2}\sum\left\{d_i - \frac{\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}{1+\exp(\beta_0^{(s)}+\beta_G^{(s)}g_i)}\right\} .\end{aligned}$$

$\Rightarrow$

$$\begin{aligned} & cov(\hat{\beta}_G^{(s)}, \hat{\beta}_{GE}^{(b)}) \\ = \; & \frac{1}{\left|-\ddot{l}(\boldsymbol{\theta})\right|}\frac{1}{a_1h_1 - b_1^2} cov\left(\sum\left(b_{41}+b_{42}g_i+b_{43}e_i+b_{44}g_ie_i\right)d_i, \sum\left(a_1g_i - b_1\right)d_i\right) \\ \propto \; & cov\left(\sum w_i^{(b)}d_i, \sum w_i^{(s)}d_i\right) , \\ & \qquad \text{where } w_i^{(b)} = \left(b_{41}+b_{42}g_i+b_{43}e_i+b_{44}g_ie_i\right) , \; w_i^{(s)} = \left(a_1g_i - b_1\right) \\ = \; & \sum w_i^{(b)}w_i^{(s)}Var(d_i) \\ = \; & \sum w_i^{(b)}w_i^{(s)}c_i . \end{aligned}$$

Now,

$$\begin{aligned}\sum w_i^{(b)}w_i^{(s)}c_i &= a_1\left(b_{41}\sum g_ic_i + b_{42}\sum g_i^2c_i + b_{43}\sum g_ie_ic_i + b_{44}\sum g_i^2e_ic_i\right) \\ &\quad -b_1\left(b_{41}\sum c_i + b_{42}\sum g_ic_i + b_{43}\sum e_ic_i + b_{44}\sum g_ie_ic_i\right)\end{aligned}$$

It can be shown that

$$b_{41}\sum g_i c_i + b_{42}\sum g_i^2 c_i + b_{43}\sum g_i e_i c_i + b_{44}\sum g_i^2 e_i c_i = 0 \ ,$$

and

$$b_{41}\sum c_i + b_{42}\sum g_i c_i + b_{43}\sum e_i c_i + b_{44}\sum g_i e_i c_i = 0 \ .$$

This leads to the result $cov(\hat{\beta}_G^{(s)}, \hat{\beta}_{GE}^{(b)}) = 0$.

Bibliography

Agresti, A. (2007). *An introduction to categorical data analysis*. Wiley-Interscience.

Allison, D., Fernandez, J., Heo, M., Zhu, S., Etzel, C., Beasley, T. and Amos, C. (2002). Bias in estimates of quantitative-trait–locus effect in genome scans: demonstration of the phenomenon and a method-of-moments procedure for reducing bias. *The American Journal of Human Genetics*, **70** 575–585.

Aschengrau, A. and Seage, G. (2003). *Essentials of epidemiology in public health*. Jones & Bartlett Publishers.

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7** 781–791.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37** 373–384.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, **35** 2313–2351.

Chanock, S., Manolio, T., Boehnke, M., Boerwinkle, E., Hunter, D., Thomas, G., Hirschhorn, J., Abecasis, G., Altshuler, D., Bailey-Wilson, J. et al. (2007). Replicating genotype-phenotype associations. *Nature a-z index*, **447** 655–660.

Cox, D. and Snell, E. (1989). *Analysis of binary data*. Chapman & Hall/CRC.

Donoho, D. and JOHNSTONE, J. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81** 425–455.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of statistics* 407–451.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96** 1348–1360.

Fan, J. and Lv, J. (2006). Sure independence screening for ultra-high dimensional feature space. *Arxiv preprint math.ST/0612857*.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Annals of statistics* 928–961.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics* 109–135.

Frayling, T., Timpson, N., Weedon, M., Zeggini, E., Freathy, R., Lindgren, C., Perry, J., Elliott, K., Lango, H., Rayner, N. et al. (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316** 889.

Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1** 302–332.

Fu, W. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics* 397–416.

Garner, C. (2007). Upward bias in odds ratio estimates from genome-wide association studies. *Genetic Epidemiology*, **31**.

Ghosh, A., Zou, F. and Wright, F. (2008). Estimating odds ratios in genome scans: an approximate conditional likelihood approach. *The American Journal of Human Genetics*, **82** 1064–1074.

Gong, G. and Samaniego, F. (1981). Pseudo maximum likelihood estimation: theory and applications. *The Annals of Statistics* 861–869.

Göring, H., Terwilliger, J. and Blangero, J. (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *The American Journal of Human Genetics*, **69** 1357–1369.

Gudbjartsson, D., Walters, G., Thorleifsson, G., Stefansson, H., Halldorsson, B., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S. et al. (2008). Many sequence variants affecting diversity of adult human height. *Nature Genetics*, **40** 609–615.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Friedman, J. and Tibshirani, R. (2001). *The elements of statistical learning*. Springer New York.

Hirschhorn, J., Lohmueller, K., Byrne, E. and Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, **4** 45.

Hoerl, A. and Kennard, R. (2000). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* 80–86.

Hu, X. and Lawless, J. (1997). Pseudolikelihood estimation in a class of problems with response-related missing covariates. *The Canadian Journal of Statistics/La Revue Canadienne de Statistique* 125–142.

Huang, J., Horowitz, J. and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*, **36** 587–613.

Huang, J., Ma, S. and Zhang, C. (2007). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica*.

Ioannidis, J., Ntzani, E., Trikalinos, T. and Contopoulos-Ioannidis, D. (2001). Replication validity of genetic association studies. *Nature Genetics*, **29** 306–309.

Jiang, Y., Scott, A. J. and Wild, C. J. (2006). Secondary analysis of case-control data. *Stat Med*, **25** 1323–1339.

Lander, E. and Kruglyak, L. (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet*, **11** 241–247.

Lee, A., McMURCHY, L. and Scott, A. (1997). Re-using data from case-control studies. *Statistics in medicine*, **16**.

Lehmann, E. and Casella, G. (1983). *Theory of point estimation*. Wiley New York.

Leonard, T. and Hsu, J. (1999). *Bayesian methods: an analysis for statisticians and interdisciplinary researchers*. Cambridge University Press.

Lettre, G., Jackson, A., Gieger, C., Schumacher, F., Berndt, S., Sanna, S., Eyheramendy, S., Voight, B., Butler, J., Guiducci, C. et al. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genetics*, **40** 584–591.

Levy, D., Ehret, G., Rice, K., Verwoert, G., Launer, L., Dehghan, A., Glazer, N., Morrison, A., Johnson, A., Aspelund, T. et al. (2009). Genome-wide association study of blood pressure and hypertension. *Nature Genetics*.

Lin, D. Y. and Zeng, D. (2008). Proper analysis of secondary phenotype data in case-control association studies. *Genet. Epidemiol.*

Liu, A., Troendle, J., Yu, K. and Yuan, V. (2004). Conditional maximum likelihood estimation following a group sequential test. *Biometrical Journal*, **46**.

Lohmueller, K., Pearce, C., Pike, M., Lander, E. and Hirschhorn, J. (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature genetics*, **33** 177–182.

Loos, R., Lindgren, C., Li, S., Wheeler, E., Zhao, J., Prokopenko, I., Inouye, M., Freathy, R., Attwood, A., Beckmann, J. et al. (2008). Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics*, **40** 768–775.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall/CRC.

Monsees, G., Tamimi, R. and Kraft, P. (2009). Genome-wide association scans for secondary traits using case-control samples. *Genetic Epidemiology*.

Nagelkerke, N., Moses, S., Plummer, F., Brunham, R. and Fish, D. (1995). Logistic regression in case-control studies: the effect of using independent as dependent variables. *Statistics in medicine*, **14**.

Palmgren, J. (1989). Regression models for bivariate binary responses. *UW Biostatistics Working Paper Series* 101.

Radchenko, P. and James, G. (2008). Variable inclusion and shrinkage algorithms. *Journal of the American Statistical Association*, **103** 1304–1315.

Rao, C. (1973). *Linear statistical inference and its applications*. Wiley New York.

Richardson, D. B., Rzehak, P., Klenk, J. and Weiland, S. K. (2007). Analyses of case-control data for additional outcomes. *Epidemiology*, **18** 441–445.

Risch, N. and Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273** 1516.

Rothman, N., Skibola, C., Wang, S., Morgan, G., Lan, Q., Smith, M., Spinelli, J., Willett, E., De Sanjose, S., Cocco, P. et al. (2006). Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. *Lancet Oncology*, **7** 27–38.

Sanna, S., Jackson, A., Nagaraja, R., Willer, C., Chen, W., Bonnycastle, L., Shen, H., Timpson, N., Lettre, G., Usala, G. et al. (2008). Common variants in the GDF5-UQCC region are associated with variation in human height. *Nature Genetics*, **40** 198–203.

Scott, A. and Wild, C. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics* 497–510.

Scott, A. and Wild, C. (2001a). Maximum likelihood for generalised case-control studies. *Journal of Statistical Planning and Inference*, **96** 3–27.

Scott, A. and Wild, C. (2002). On the robustness of weighted methods for fitting models to case-control data. *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 207–219.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika*, **84** 57–71.

Scott, A. J. and Wild, C. J. (2001b). Maximum likelihood for generalised case-control studies. *J. Statist. Plann. Inference*, **96** 3–27. Statistical design of medical experiments, II.

Scott, L., Mohlke, K., Bonnycastle, L., Willer, C., Li, Y., Duren, W., Erdos, M., Stringham, H., Chines, P., Jackson, A. et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316** 1341.

Siegmund, D. (2002). Upward bias in estimation of genetic effects. *The American Journal of Human Genetics*, **71** 1183–1188.

Stamey, T., Kabalin, J. and Ferrari, M. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. III. Radiation treated patients. *J Urol*, **141** 1084–1087.

Sun, L. and Bull, S. (2005). Reduction of selection bias in genomewide studies by resampling. *Genetic epidemiology*, **28**.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Todd, J., Walker, N., Cooper, J., Smyth, D., Downes, K., Plagnol, V., Bailey, R., Nejentsev, S., Field, S., Payne, F. et al. (2007). Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature genetics*, **39** 857–864.

Wald, A. (1943). Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large. *Transactions of the American Mathematical Society*, **54** 426–482.

Wang, L., Chen, G. and Li, H. (2007). Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics*, **23** 1486.

Wang, W., Barratt, B., Clayton, D. and Todd, J. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, **6** 109–118.

Wasserman, L., Roeder, K. and Pittsburgh, P. (2007). Multi-Stage Variable Selection: Screen and Clean. *Arxiv preprint arXiv:0704.1139*.

Weedon, M., Lango, H., Lindgren, C., Wallace, C., Evans, D., Mangino, M., Freathy, R., Perry, J., Stevens, S., Hall, A. et al. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genetics*, **40** 575–583.

Weedon, M., Lettre, G., Freathy, R., Lindgren, C., Voight, B., Perry, J., Elliott, K., Hackett, R., Guiducci, C., Shields, B. et al. (2007). A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nature genetics*, **39** 1245–1250.

Wild, C. (1991). Fitting prospective regression models to case-control data. *Biometrika*, **78** 705–717.

Yu, K., Chatterjee, N., Wheeler, W., Li, Q., Wang, S., Rothman, N. and Wacholder, S. (2007). Flexible design for following up positive findings. *The American Journal of Human Genetics*, **81** 540–551.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68** 49–67.

Zöllner, S. and Pritchard, J. (2007). Overcoming the winners curse: estimating penetrance parameters from case-control data. *The American Journal of Human Genetics*, **80** 605–615.

Zondervan, K. and Cardon, L. (2007). Designing candidate gene and genome-wide case–control association studies. *Nature Protocols*, **2** 2492–2501.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, **67** 301–320.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, **36** 1509–1533.