LASSO META-REGRESSION AND ITS APPLICATION TO PSYCHIATRIC GENOMICS

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

by

Charles Laurin

______________________________
Gitta Lubke, Director

Graduate Program in Psychology
Notre Dame, Indiana
April 2014
This document is in the public domain.
LASSO META-REGRESSION AND ITS APPLICATION TO PSYCHIATRIC GENOMICS

Abstract

by

Charles Laurin

Meta-analysis is an essential technique in psychiatric genomics. Heterogeneity of the effect sizes included in a meta-analysis can reduce its statistical power. Meta-regression provides a way to identify study-level covariates that are associated with heterogeneity. However, in psychiatric genomics, and other new fields, there is little theory to guide the choice of study-level covariates to be included in a meta-regression. We propose the use of shrinkage regression methods, such as the Lasso and the SCAD, as means to select study-level covariates automatically. We used simulated data to compare these methods to simple regression, a current standard approach. The criterion used in the comparison was the precision of the method in selecting covariates. We found that precision was highly dependent on the number of study level covariates included, the number of studies included, and the relative sizes of the included studies. When many studies were included, a bootstrapped version of the Lasso was able to select covariates with adequate precision. In moderate-sized meta-analyses, a bootstrapped version of simple regression performed best. The comparison of methods presented in this work is applicable to the design of new consortia of studies in psychiatric genomics, as well as to the evaluation and comparison of existing consortia.
CONTENTS

FIGURES .......................................................................................................... v

TABLES ........................................................................................................ vii

SYMBOLS ....................................................................................................... viii

ACKNOWLEDGMENTS .................................................................................... x

CHAPTER 1: INTRODUCTION: PSYCHIATRIC GENOMICS RELIES ON
META-ANALYSIS BUT IS STILL DEVELOPING ............................................. 1

CHAPTER 2: META-REGRESSION: ESTIMATING AN OVERALL EFFECT
CONDITIONAL ON STUDY-LEVEL COVARIATES ........................................ 5
2.1 Fixed-effects meta-regression ................................................................. 6
2.1.1 Fixed-effects meta-regression models are estimated using weighted
least squares regression .............................................................................. 7
2.2 Random-effects meta-regression is used for greater generalizability .... 7
2.2.1 Random-effects meta-regression requires estimating between-
study variance .......................................................................................... 8
2.2.1.1 Random-effects meta-regression as used in linearly
heteroscedastic models ........................................................................... 9
2.3 Automatic SLC selection is uncommon, but often based on significance
tests ........................................................................................................... 10

CHAPTER 3: REGRESSION WITH SHRINKAGE: METHODS FOR VARIABLE
SELECTION WHEN COVARIATES OUTNUMBER OBSERVATIONS ............. 13
3.1 Approach .................................................................................................. 16
3.1.1 Vector and Residual Bootstrapping .................................................. 17
3.1.1.1 The vector bootstrap .................................................................. 17
3.1.1.2 The residual bootstrap ................................................................. 18
3.1.2 Previous research in bootstrapping the Lasso ................................. 18
3.1.2.1 Research with residual bootstrap ............................................. 18
3.1.2.2 Research with vector bootstrap .............................................. 19
3.1.2.3 Research with cross-validated selection of $\lambda$ .................... 19
3.1.3 Role of $\lambda$ when fitting Lasso models ........................................ 20
  3.1.3.1 Selection of $\lambda$ through $K$-fold cross-validation .................. 21
  3.1.3.2 Lasso variance estimates: Contribution of $\lambda$ selection ........ 21
  3.1.3.3 Lambda selection in bootstrapping ........................................ 22
3.2 Methods ............................................................................................... 23
  3.2.1 Lasso CIs: Secondary selection of predictors ................................. 24
3.3 Simulation Studies .................................................................................. 25
  3.3.1 Simulation Design .......................................................................... 26
  3.3.2 Evaluation Criteria ......................................................................... 27
  3.3.3 Simulation Results ......................................................................... 28
    3.3.3.1 Improved Variable Selection Precision ................................. 28
3.4 Empirical Illustration ............................................................................ 28
  3.4.1 Empirical illustration: Results ....................................................... 30
3.5 Conclusion ............................................................................................. 32

CHAPTER 4: SIMULATION STUDIES: CHARACTERISTICS OF SAMPLE
DATA SETS THAT INFLUENCE META-REGRESSION PERFORMANCE 35
4.1 Introduction: Comparing methods for selecting variables in meta-
regression ..................................................................................................... 35
4.2 Design of the simulations ...................................................................... 36
  4.2.1 Data generation: a simple model .................................................. 36
  4.2.2 Factors that influence the probability of correct SLC selection ......... 37
    4.2.2.1 The sample size factor ......................................................... 37
    4.2.2.2 The association strength factor .......................................... 38
  4.2.3 Manipulating false positive selections: The number of SLCs
    factor ............................................................................................... 39
  4.2.4 Empirical concerns: Study weights and heteroscedasticity ............ 41
    4.2.4.1 The study weights factor reflects the differences in
    study size that are observed in practice .......................................... 41
    4.2.4.2 Random effects and study weights ...................................... 42
    4.2.4.3 Random effects: the linear heteroscedasticity model .......... 44
  4.2.5 Expectations before simulating ..................................................... 45
    4.2.5.1 Applying SCAD and bootstrap CIs .................................... 47
4.3 Implementation details .......................................................................... 47
4.4 Results .................................................................................................... 49
  4.4.1 Main effect of the method factor ................................................... 49
  4.4.2 Factors that influence the probability of correct SLC selection:
    Sample size and association strength .............................................. 50
  4.4.3 Manipulating false positive selections: The number of SLCs
    factor ............................................................................................... 53
  4.4.4 Study weights ................................................................................ 55
  4.4.5 Linear heteroscedasticity model .................................................... 57
    4.4.5.1 Random effects and study weights .................................... 57
    4.4.5.2 Selecting random effects ................................................... 57
FIGURES

3.1 Default approach to Lasso variable selection using $K$-fold cross-validation. 23
3.2 Vector bootstrap for improved Lasso variable selection precision with
$\lambda$ treated as fixed. ................................................................. 24
3.3 Vector bootstrapping with metaparameter $\lambda$ treated as random. .... 25
3.4 Bootstrap means and SEs of 77 SNPs, 2926 pairwise interactions; light
gray diamonds represent predictors selected without bootstrapping;
The lines represent mean= $\pm 2 \times SE^*$: points and diamonds outside
the $V$-shape (i.e. in lower corners) are promising signals. A cube-root
transformation was used on both axes. ........................................ 31
4.1 A quantile-quantile plot shows the that the observed distribution of
weights in Wray et al. (2012) has a similar shape to an Exponential(1)
distribution. The line has a slope of 1/2. ................................. 42
4.2 Boxplots of VSP values for each method suggest that, overall, the
Lasso has poor performance, bootstrapped simple regression good per-
formance, and bootstrapped Lasso highly variable performance. .... 51
4.3 A boxplot of VSP values at each sample size suggests that boot-
strapped methods perform best at large sample sizes and that simple
regression performs best at small sample sizes. .......................... 52
4.4 A boxplot of VSP values at each number of SLCs ($p$) shows that
increasing $p$ degrades VSP for all methods, but suggests that boot-
strapped Lasso regression is the least affected. ......................... 54
4.5 Skewed study weights decrease the VSP of all selection methods. The
decrease in performance is greater than the decrease resulting from
considering a large number of SLCs. ........................................ 55
4.6 For all methods, VSP was much larger for data generated under ran-
dom effects, rather than fixed-effects, models. .......................... 58
4.7 The Lasso and simple meta-regression both have relatively low VSP
for selecting an important predictor of random-effects variance. The
random-effects variance in this example represents 40% of the total
variance in effect sizes across studies. ................................. 59
4.8 Boxplots of VSP values in the random-effects conditions, showing the effects of association strength and sample size. In larger data sets and/or stronger association, the bootstrapped Lasso performs well; with moderate power, bootstrapped simple regression performs well. Identical data as in Figure A.4.

A.1 A boxplot of VSP values at each level of association strength ($r^2$) suggests that bootstrapped methods perform best across $r^2$ values.

A.2 Boxplots of VSP values in the fixed-effects condition, showing the effects of association strength and sample size. Identical data as in Figure A.3.

A.3 Scatterplots of VSP values in the fixed-effects condition, showing the effects of association strength and sample size. Identical data as in Figure A.2.

A.4 Boxplots of VSP values in the random-effects conditions, showing the effects of association strength and sample size. Identical data as in Figure 4.8.
TABLES

3.1 VARIABLE SELECTION PRECISION OF DEFAULT LASSO VS. BOOTSTRAPPED PERCENTILE CIS

3.2 PROMISING SNP-SNP INTERACTIONS

4.1 POWER TO DETECT EFFECTS USING A TWO-SIDED TEST OF SIMPLE REGRESSION AT $\alpha = 0.05$

4.2 EXPECTED NUMBER OF FALSE POSITIVES IN MULTIPLE SIMPLE META-REGRESSION AT $\alpha = 0.05$

4.3 SIMPLE META-REGRESSION: EXPECTED VSP BY $N$, $p$, AND $r^2$

4.4 SIMULATION DESIGN SHOWING WHICH CELLS USED SCAD OR BOOTSTRAPPING

4.5 VSP IN THE BEST PERFORMING CELL FOR EACH METHOD

5.1 SQUARED CORRELATIONS BETWEEN SLCS AND EFFECT SIZES SUGGEST SMALL-TO-MODERATE ASSOCIATIONS
SYMBOLS

\( \alpha \)  Significance level of a hypothesis test

\( \beta \)  \( p \)-vector of regression coefficients

\text{bootLas}  \text{Lasso with bootstrapped CIs}

\text{bootSimp}  \text{Simple regression with bootstrapped CIs}

\( \gamma \)  \( q \)-vector of coefficients in the linear heteroscedasticity model

\text{CI}  \text{Confidence interval}

\( \delta_i \)  Unobservable random effect underlying effect size in study \( i \)

\( \varepsilon_i \)  Within-study (sampling) error in estimating study \( i \) effect size

\( \mathcal{E}(Y) \)  Expected value of random variable \( Y \)

\( \hat{F}(\hat{\beta}_j) \)  Empirical marginal distribution of coefficient estimate \( \beta_j \)

\( \hat{F}^*\left(\hat{\beta}_j\right) \)  Monte Carlo estimate of \( \hat{F}^*\left(\hat{\beta}_j\right) \)

\( H^2_{y|x} \)  Heterogeneity: residual variance of effect sizes \( y \) given SLC \( x \)

\( \lambda \)  Lasso metaparameter (parameter of parameters) that determines shrinkage of coefficients

\text{MC}  \text{Monte Carlo}

\( N \)  Number of studies in a meta-analysis

\( p \)  Number of SLCs in a meta-analysis

\( r^2 \)  Squared correlation between one predictor and the outcome

\( \sigma^2 \)  Within-study variance of effect size

\( s_i \)  Estimated standard error of study \( i \) effect size

\( s \)  Set of indices of selected predictors (SLCs)
SE Standard error
SLC Study-level covariate, e.g. % smokers in study, publication year, etc.
   SLCs are the predictor variables in meta-regression
SNP Single nucleotide polymorphism; binomial random variable that represents the genotypes at a particular location in the genome
\( \tau \) Between-study variance of effect size
\( V(Y) \) Variance of random variable \( Y \)
VSP Variable selection precision: proportion of selected variables that are important
\( \omega_i \) The weight given to study \( i \) in random-effects meta-analysis
\( w_i \) The weight given to study \( i \) in fixed-effects meta-analysis
\( W \) \( N \times N \) matrix of weights used in fixed-effects WLS regression
\( x_{ij}, u_{ij} \) Observed value of study-level covariate (SLC) \( j \) in study \( i \)
\( y_i \) Effect size of study \( i \); effect sizes are the outcome variable in meta-regression
\( y \) Overall or combined effect size of studies in a meta-analysis
ACKNOWLEDGMENTS

I wish to thank my committee members, Jiahan Li, Ke-Hai Yuan, and Zhiyong “Johnny” Zhang for their helpful feedback on the first version of this manuscript and all of the other help that they’ve given me over the past half-decade. Next, I will thank my lab-mates Dan McArtor, Patrick Miller, and Raymond Walters, for listening to the ideas presented here since my first vague descriptions of them. Then, I need to thank my adviser and indefatigable writing tutor, Gitta Lubke, for her guidance and support in all aspects of doing research. Thanks to my friends in Haggar Hall and in the Transitorium, who helped me in many ways. Finally, the biggest thanks goes to my family, especially to my wife, Sarah.

This dissertation research was supported by a Presidential Fellowship and by a Professional Development Award from the Graduate School of the University of Notre Dame, and by a Graduate Research Award from Notre Dame’s Institute for Scholarship in the Liberal Arts. Development of the research was aided by conference attendance that was made possible by travel awards from the Behavior Genetics Association and the Society for Multivariate Experimental Psychology. A visit to the Department of Biological Psychology at VU University Amsterdam and presentations at the Quantitative Studies Group seminar at Notre Dame led to helpful feedback on early presentations of the material.
CHAPTER 1

INTRODUCTION: PSYCHIATRIC GENOMICS RELIES ON META-ANALYSIS BUT IS STILL DEVELOPING

Psychiatric genomics is a relatively new area of study that is highly dependent on meta-analysis. Psychiatric genomics is the search for associations between individual genetic markers and psychiatric symptoms. The data required to test such associations only became available in about 2005 (Psychiatric GWAS Consortium Coordinating Committee and others, 2009). To date, the existing reliable associations between individual genetic markers and psychiatric symptoms have been found to have very small effect sizes. It is assumed that very large sample sizes will be required for future studies to have adequate statistical power (Park et al., 2011a). Meta-analysis provides a principled way to increase statistical power by combining samples from different studies. Because of this, it has become a crucial method in psychiatric genomics (Corvin et al., 2010). However, researchers in psychiatric genomics could observe yet larger gains in power by using meta-regression to address the heterogeneity of the studies included in their meta-analyses (Higgins and Thompson, 2002).

Meta-analyses combine effect sizes from several studies to produce an overall effect size estimate. If the studies’ effect sizes are independently and identically distributed, the test of the overall effect size will be more powerful than any of the studies’ hypothesis tests (Cohn and Becker, 2003). The test of overall effect may not have increased power if there is un-addressed variability among the individual studies’ effect sizes.
We define heterogeneity as variability of effect sizes between the studies in a meta-analysis, based on usage in Higgins and Thompson (2002) and Hedges and Olkin (1985, pg. 122):

\[
H_y^2 = N^{-1} \sum_i (y_i - y.)^2 ,
\]

(1.1)

where \( H^2 \) is heterogeneity, \( y_i , i = 1, \ldots, N \) are the effect sizes from studies in a meta-analysis, and \( y. \) is the overall effect size. For example, the \( y_i \) could be Fisher-transformed correlation coefficients and \( y. \) their sample mean. In practice, the \( y_i \) are weighted based on the sample size and \( y. \) is a weighted average; we omit this for now to simplify presentation.

The purpose of meta-regression is to identify variables that explain the heterogeneity of effect sizes between the studies in a meta-analysis (Thompson, 1994; Knapp and Hartung, 2003). Meta-regression models the conditional means of studies’ effect sizes as linear functions of study-level covariates (SLCs), such as their design and sample characteristics (Hedges and Olkin, 1983). In a meta-regression, the regression mean square is interpreted as the portion of the heterogeneity that can be explained by the SLCs. The residual variance is interpreted as heterogeneity not explained by the SLCs in the meta-regression. Equation 1.2 presents this relationship for a single SLC:

\[
\hat{y}_i = E (y | x_i) = \beta x_i \\
H_y^2 = N^{-1} \sum_i (y_k - \hat{y}_i)^2 + N^{-1} \sum_i (\hat{y}_i - y.)^2 \\
H_y^2 - N^{-1} \sum_i (\hat{y}_i - y.)^2 = N^{-1} \sum_i (y_i - \hat{y}_i)^2 \\
= H_{y|x}^2,
\]

(1.2)

\( x_i \) is the observed value of a study-level covariate (SLC) \( X \), such as the percentage
of women in study \( i \), the year of publication, a dummy variable representing country of origin, etc, \( \beta \) is a regression coefficient, and \( H^2_{y|x} \) is the residual variance, interpreted as heterogeneity that is not attributable to SLCs.

Meta-regression coefficients are typically estimated using the method of least squares; this makes them functions of the \( N \) observed study effect sizes \( y_i \) and corresponding SLCs \( x'_i \). Careful selection of SLCs is vital to avoid capitalizing on chance in meta-regression with multiple SLCs: when an additional SLC is included in the meta-regression, the residual variance decreases or stays the same; it cannot increase. The expected proportion of heterogeneity that is explained by an additional SLC with no true effect on the outcome is \( 1/(N - 1) \) (Helland, 1987). When the number of SLCs equals the number of studies (i.e., the number of predictors equals the number of observations), the meta-regression model must fit perfectly (Mulaik, 2001). Consequently, when the number of SLCs is high relative to the number of studies, the proportion of heterogeneity “explained” may be high even if each SLC has little or no association with studies’ effect sizes. In highly-developed areas of research, this problem can be avoided: the potential sources of heterogeneity can be determined by theory and common practice, and a small number of the most important SLCs can be selected for use in a meta-regression (Berlin, 1995).

In newer fields, the factors that influence study quality are less known, and there are also fewer studies to include in a meta-analysis. Lacking theory to suggest SLCs, meta-analysts often present numerous SLCs that may be important sources of heterogeneity. The number of SLCs presented often approaches or exceeds the number of studies in the meta-analysis. For example, in an influential meta-analysis of the genetics of major depressive disorder, Wray et al. analyzed 5 studies, but presented descriptive statistics for 10 study-level covariates in their Table 2 (2012, p. 39). Heterogeneity was observed across the five studies, and may have been a factor in the inability of the meta-analysis to identify significant genetic predictors of depression.
In a recent meta-analysis of the association between genetic markers and amygdala activity, [Murphy et al. (2013)](#), examined 34 studies and attempted to explain heterogeneity by performing simple meta-regressions of 20 SLCs. [Ligthart et al. (2011)](#) performed a meta-analysis of six studies associating genetic markers with migraine, and included a table of 8 SLCs (though did not perform meta-regression). These cases show the need for a method to search for a small set of important SLCs as causes of heterogeneity in meta-analyses in psychiatric genomics. This need is also present in other fields that lack sufficient theory to guide choice of SLCs.

Therefore, we propose using the Lasso as a meta-regression method. The Lasso is a variant of linear regression that selects variables automatically and that is able to produce unique models when the number of covariates is greater than or equal to the sample size [Tibshirani (2011)](#). We will show that by using the Lasso to fit multiple regression models to meta-analytic data, researchers can explore a large number of SLCs to identify a small subset that explains heterogeneity. We will also demonstrate that bootstrapping and cross-validating Lasso models reduces the problem of capitalizing on chance that arises from this model search.

We begin with a review of meta-regression, including the current standard approaches for identifying important SLCs. Next, we review the Lasso and our relevant previous research. Then, we present simulation studies evaluating the use of the Lasso as a meta-regression tool. Finally, we apply the Lasso and other meta-regression methods to data published by [Murphy et al. (2013)](#).
A primary goal of meta-analysis is to combine the effect sizes \( (y_i) \) observed in several studies in order to produce an overall effect size estimate \( (y) \) (Knapp and Hartung, 2003). Meta-regression is the estimation of \( y \) conditional on study level covariates (SLCs) \( x \), (Hedges and Olkin, 1983). Conditioning on SLCs allows meta-analysts to: 1) estimate the overall effect size \( y \) with more precision (Rencher and Schaalje, 2008); 2) estimate a linear association between a SLC \( x \) and study effect size \( y \); and 3) test the hypothesis that a set of SLCs \( x \) is a source of heterogeneity across studies’ effect sizes by testing whether the SLCs contribute significantly to the observed variability in study effect sizes \( H^2_y \).

Meta-regression models are estimated through weighted least squares regression, where the studies’ effect sizes \( y_i \) are the outcome, their SLCs \( x_i' \) are the predictors, and their within-study standard errors \( s_i \) are the weights. If the SLCs are centered, the intercept in a meta-regression model provides the overall effect size estimate (Higgins et al., 2009).

Estimation and significance testing in meta-regression analyses take two forms: fixed-effects models, which are based on assuming that the observed variability in study effect sizes is due entirely to sampling error and to the SLCs (Equation 2.1); and random-effects models, which are based on assuming that unobserved variables lead to additional variability between studies’ effect sizes (Equation 2.4) (Raudenbush and Bryk, 1985).
Significance testing in meta-regression is used differently in older areas of research than in newer ones. In mature areas of research, the choice of SLCs to include in a meta-regression is largely determined by theory (Berlin [1995]). Hence, significance testing of multiple regression models is used to evaluate the theory. In newer areas, it is common to investigate more SLCs than studies. Significance testing of marginal regression models is used to identify important SLCs and to generate hypotheses.

2.1 Fixed-effects meta-regression

The model underlying fixed-effects meta-regression is:

\[ y = \mu + \varepsilon = 1 \beta_0 + X \beta + \varepsilon \] (2.1)

\( y \) is an \( N \)-element column vector of studies’ effect sizes. \( \mu \) is an \( N \)-vector of true effect sizes. \( \varepsilon \) is an \( N \)-element random vector of within-study errors that represents the effect of sampling error on measuring the studies’ effects. \( \mathcal{E}(\varepsilon) = 0 \). \( \mathcal{V}(\varepsilon) = \sigma^2 I \), where \( \sigma^2 \) is the variance of each \( \varepsilon_i \) and \( I \) is the \( N \times N \) identity matrix.

\( X \) is an \( N \times p \) matrix of SLCs with columns \( X_j \) and rows \( x'_i \). The fixed effects of each SLC are in the \( p \times 1 \) vector \( \beta \). The intercept, \( \beta_0 \), gives the average effect size considered over all levels of the covariate. \( 1 \) is a \( N \)-element column vector of 1s. Hence, the mean effect size vector has elements

\[ \mu_i = \beta_0 + x'_i \beta \]

The overall effect size is given by \( N^{-1} \sigma^{-1} \sum \mu_i \), so \( H^2_{y|x} \) is unity.
2.1.1 Fixed-effects meta-regression models are estimated using weighted least squares regression.

The most important statistic in a meta-regression is \((\beta_0, \hat{\beta}')'\), the \(p + 1\)-vector containing the estimated intercept and regression coefficients. This vector is calculated using weighted least squares regression \((\text{Aitken 1935})\). The studies in the meta-regression are treated as independent and, to put them on a common scale, are weighted by their standard errors, which depend on their sample sizes \((\text{Hedges and Olkin 1983})\). Thus, the weight matrix, \(W\) is diagonal with entry \(w_{ii}\) equal to
\[
\frac{1}{\sum_i \left\{ (s_i^2)^{-1} \right\} } \left( s_i^2 \right),
\]
where \(s_i^2\) is the estimated variance of effect size \(y_i\). Then, the coefficient estimates are:

\[
\hat{\beta} = (X'WX)^{-1} X'Wy \quad (2.2)
\]
\[
\beta_0 = y - N^{-1} \left( W^{1/2} X \hat{\beta} \right)'1 \quad (2.3)
\]

Testing whether \(\beta\) is significantly different from \(0\) can then be done, most often through a \(\chi^2\) test, which requires that \(\varepsilon\) is assumed to be normally distributed and \(s_i^2\) is assumed to be a consistent estimator of \(\sigma^2\). Alternatively, permutation testing can be used \((\text{Huizenga et al. 2011})\). Both hypothesis tests will be described in more detail later in the chapter.

2.2 Random-effects meta-regression is used for greater generalizability

Fixed-effects meta-regression is a way of identifying SLCs associated with the observed variability of effect sizes among studies in a meta-analysis. It is possible that there are additional sources of variability across studies that can’t be attributed to a measured SLC or to sampling error. These are modeled as unobservable random effects \((\text{Hedges 1983})\). Usually, random effects for individual studies are not
estimated. Rather, their overall contribution to the heterogeneity of effect sizes is estimated. Estimating the contribution of random effects to heterogeneity enables generalization of the results beyond the N studies included in the meta-regression (Huizenga et al., 2011). Estimating the heterogeneity due to random effects makes generalization possible because the variance due to random effects is incorporated in the prediction interval for the effect size of a new study (Higgins et al., 2009).

This is called random-effects meta-regression (although a more precise name would be “mixed-effects meta-regression,” because fixed and random effects both are estimated) (Higgins and Thompson, 2004).

The model underlying random-effects meta-regression is:

\[
\begin{align*}
y &= \mu + \varepsilon \\
&= 1\beta_0 + X\beta + \delta + \varepsilon
\end{align*}
\]

(2.4)

where \( \delta \) is a \( N \)-vector of study-specific random-effects \( \delta_i \), with the assumption that \( \delta \) is uncorrelated with \( \varepsilon \) and \( X \) and that \( \delta \sim N(0, \tau I) \). As a result,

\[
\begin{align*}
\mathbb{E}(y_i) &= \beta_0 + x_i'\beta + \delta_i \\
\mathbb{V}(y_i) &= \sigma^2 + \tau.
\end{align*}
\]

(2.5) (2.6)

2.2.1 Random-effects meta-regression requires estimating between-study variance

In fixed-effects meta-regression, the within-study error variances are estimated using data from each study. Each study has its own estimate of this variance, which reflects study size (number of participants in the study) and individual differences between participants. In contrast, in random-effects meta-regression, the between-
studies error variance is estimated using data from all studies:

\[ \hat{V}(y_i | x_i') = s_i^2 + \hat{\tau}, \]

where \( s_i^2 \) is the squared standard error of the effect size \( y_i \), and \( \hat{\tau} \) is an estimate of the between-studies error variance (Berkey et al., 1995). The practical result of this is larger estimates of variance for each study’s effect size than when using fixed-effects models. Using these variance estimates tends to result in wider confidence intervals around the estimated overall effect as well (Senn, 2007; Poole and Greenland, 1999). The corresponding study weights are

\[ \omega_i = \frac{\hat{V}(y_i | x_i')^{-1}}{\sum_i \{\hat{V}(y_i | x_i')^{-1}\}}. \]

Estimation in random-effects meta-analysis is most commonly done using either the method of moments (DerSimonian and Laird, 1986) or restricted maximum likelihood estimation (Higgins and Thompson, 2004), though full-information MLE (Huizenga et al., 2011) and Bayesian estimation (Berkey et al., 1995) are also used.

2.2.1.1 Random-effects meta-regression as used in linearly heteroscedastic models

Random-effects meta-regression is also useful in cases where SLCs are assumed to influence the variance of studies’ effect sizes without influencing the mean value. For example, the initial studies in an area of research may be of lower quality than newer ones because of the development of best practices in study design. Hence, year of publication is a SLC that influences between-study variance in effect sizes without affecting the mean. Modeling the influence of publication year on between-study variance can be done using the linear heteroscedasticity model given in Equation 2.7 (Amemiya, 1977, 1978).
\( \mathcal{V}(y_i | u_i) = \sigma^2 + \tau_i \)

\[ = \sigma^2 + \gamma_0 + u'_i \gamma \]

\[ = \tilde{\gamma}_0 + u'_i \gamma \quad (2.7) \]

In Equation 2.7, \( u'_i \) is a vector of SLC values observed for study \( i \), where each \( u_i \) is assumed to be a cause of heteroscedasticity, \( \gamma_0 \) is the between-studies variance for studies with \( u = 0 \), \( \gamma \) is the vector of regression coefficients of variance in effect sizes on the SLCs in \( u \), and \( \tilde{\gamma}_0 \) is \( \sigma^2 + \gamma_0 \).

The linear heteroscedasticity model can be estimated using the method of moments, which requires the assumption that the SLCs \( u \) that influence \( \mathcal{V}(y) \) are uncorrelated with those SLCs \( X \) that influence \( \mathcal{E}(y) \) (Hasbrouck, 1986). Then, the intercept \( \tilde{\gamma}_0 \) and coefficients \( \gamma \) can be estimated by regressing \( y \) on \( X \), calculating the squared residuals \( \hat{\varepsilon}^2_i \), and regressing the \( \hat{\varepsilon}^2_i \) on \( u \). Expressions for the asymptotic variance of \( \hat{\gamma} \) are given in Amemiya (1977, 1978). These can be used to perform significance testing.

2.3 Automatic SLC selection is uncommon, but often based on significance tests

In most fields, meta-regressions are done using a small number of SLCs that have been selected on theoretical grounds prior to analysis; for example, in a study of the genetics of kidney disease, Lim et al. (2014) used 4—ethnicity, percentage male, ethnicity \( \times \) sex interaction, and percentage hypertensive participants. As a result, methodological research on meta-regression has focused on such models, often using simple regressions in derivations and simulations (e.g., Huizenga et al. 2011, López-López et al. 2014). Permutation tests and \( t \)-tests based on conservative \( df \) (i.e., \( N - 4 \) in the test of a single predictor) are used to test the associations between a SLC and
studies’ effect sizes (Knapp and Hartung, 2003). Possibly, this is the reason that, in psychiatric genomics, multiple simple regressions have been used to select SLCs when the number of SLCs approaches or exceeds the number of studies. For example, Murphy et al. (2013) performed simple regressions on 10 SLCs in their meta-analysis of 34 studies.

In the less-common case of multiple meta-regression, stepwise model fitting has been used (e.g., Gagnier et al. 2012; Umpierre et al. 2011). Shadish and Sweeney (1991) and Cheung (2008) represented meta-regression as a structural equation model, and used fit indices ($\chi^2$, NFI, and CFI, which are likelihood-based) to perform model search and selection. Gagnier et al. (2012) applied permutation testing as part of a backward stepwise selection procedure. Bootstrapped meta-regression has been proposed and used in practice, but few if any comparisons to other methods have been done (Stanley and Jarrell, 2005; Koskinen et al., 2010).

When using hypothesis testing to select SLCs, rather than to test a pre-specified, theory-based model, it is necessary to correct the p-values of test statistics for multiple testing (Higgins and Thompson, 2004; Hedges and Pigott, 2004). Conservative tests are also helpful to avoid capitalizing on chance when the number of SLCs approaches the number of studies in the meta-regression (i.e. $N \approx p$) (Knapp and Hartung, 2003). When $N \leq p$, it is not possible to fit a multiple regression model containing all predictors, hence multiple single regressions with corrected significance tests has been the method of choice for selecting SLCs. The disadvantage of this approach is that only the marginal association between each SLC and the outcome is tested. In the next chapter, we introduce methods for data with $N \leq p$. This is an important case in practice because in psychiatric genomics, the number of studies can be small, and there is little theory to relate SLCs to studies’ effect sizes. To begin building this theory, it is necessary to evaluate a large number of SLCs that may be relevant. These methods perform automatic selection of SLCs based on the strength of a SLC’s
association with the outcome *conditioned on* the relationships between other SLCs and the outcome.
CHAPTER 3

REGRESSION WITH SHRINKAGE: METHODS FOR VARIABLE SELECTION WHEN COVARIATES OUTNUMBER OBSERVATIONS

Multiple linear regression is most useful when it is applied to samples in which the number of predictors is small relative to the number of observations. However, if the number of predictors is large relative to the number of observations, there will be considerable sampling variability in the estimated coefficients. Under those conditions, the increased variability of coefficients can be managed by applying shrinkage methods. Shrinkage means estimating coefficients under a constraint that leads them to have reduced absolute values, drawing them toward 0, thus reducing sampling variability. There are many such constraints, hence many shrinkage methods. One of the most important shrinkage methods is the Lasso (Tibshirani, 2011).

When the Lasso is applied, small-valued coefficient estimates are reduced to 0 and the remaining coefficient estimates are shrunk by a fixed amount (Tibshirani, 2013). Because of this property, the Lasso is often used for variable selection (Tibshirani, 1996). In Lasso variable selection, the predictors having nonzero coefficient estimates after shrinkage are selected into the regression model, and those that are shrunk to 0 are excluded from the model.

Variable selection with the Lasso, or with any other method, is a classification problem. The problem is to classify the predictor variables that could be included in a regression model either as “unimportant” or as “important.” Ideally, the unimportant, or “noise,” predictors are associated with the outcome only because of sampling error, while important predictors are consistently associated with the outcome over
independent samples. Unimportant predictors should be excluded from the regression model (i.e., their coefficient estimates shrunk to 0), while important predictors should be included in the model.

Using the Lasso for variable selection translates to classifying the excluded predictors as unimportant and the included predictors as important. The accuracy of this claim has been studied in simulation studies, and the Lasso has often been found to have low variable selection precision (VSP), meaning that among the included predictors, a relatively large proportion (> 50%) were false positives (Devlin et al., 2003; Ayers and Cordell, 2010; He and Lin, 2011). In this chapter, we describe a method for controlling the Lasso’s VSP. In Chapters 4 and 5, we apply this method to meta-regression using shrinkage: SLCs will be identified as important or as unimportant by using the Lasso and the SCAD, a related method that is introduced at the end of this chapter.

We build on a variety of research that has investigated how to control the Lasso’s VSP. A main current in this research has been to estimate standard errors (SEs) and confidence intervals (CIs) for Lasso coefficient estimates with the bootstrap (e.g. Sartori, 2009; Chatterjee, 2011). Lasso SEs and CIs are used in a secondary variable selection step based on analogy with hypothesis testing (Freedman and Lane, 1983). The SEs are used to generate $t$-statistics; predictors having $t$-statistics below a user-set cutoff (e.g. $|t| \leq 2$) are excluded. Similarly, predictors with CIs that contain 0 are excluded.

Successfully applying a second variable selection step requires knowledge of why the Lasso includes too many variables in the model during the first step. Lasso variable selection depends on its degree of shrinkage. The degree of shrinkage in a Lasso model is controlled by a user-set metaparameter called $\lambda$.

The shrinkage metaparameter $\lambda$ is a scalar; its value determines the number of variables that the Lasso selects. The larger the value of $\lambda$, the more conservative the
model. The choice of $\lambda$ is thus related to the Lasso’s VSP. In general, each value of $\lambda$ is associated with a single Lasso model. $\lambda$ is commonly chosen through model-comparison methods: the user proposes a set of candidate $\lambda$s and chooses one based on, e.g., BIC or cross-validation indices.

Of these common methods, using cross-validation to choose $\lambda$ has been associated with overfitting, leading to an excess of false positives and low VSP (Meinshausen and Bühlmann 2010; James and Radchenko 2009). Fan and colleagues (2008; 2012) attribute this to the large sample correlations that can arise between noise predictors and the outcome. They demonstrate empirically that noise correlations increase in magnitude with increasing numbers of noise predictors. Further, when sample size $N$ is less than number of predictors $p$, the largest noise correlation can easily exceed true correlations in magnitude. In cross-validation, the sample size used for model-fitting is always smaller than that in the entire sample, exacerbating the problems identified by Fan et al.

There has been little or no investigation into using bootstrap SEs and CIs to mitigate low VSP associated with cross-validated selection of $\lambda$. Under bootstrap resampling, Lasso coefficients for noise predictors are expected to fluctuate between positive and negative values (Chatterjee 2011). This should lead noise predictors to have larger bootstrap SEs and CIs than important predictors, which suggests that these statistics are useful for identifying false positive associations. Hence, SLCs with large bootstrap SEs or CIs should be eliminated from meta-regression models.

To investigate this expectation in a general regression context, we introduced and evaluated a bootstrap-based method with the goal of improving Lasso VSP by excluding false positive predictor selections under cross-validation. We investigated bootstrap SEs and CIs for the Lasso when cross-validated selection of $\lambda$ is done before bootstrapping, which is the standard approach (e.g. Sartori 2009; D’Angelo et al. 2009). An alternative is to nest cross-validated selection of $\lambda$ within each
bootstrap replication, which could lead to larger SEs and wider CIs than in the standard approach (Bühlmann et al. [2011]). Such larger SEs and wider CIs lead to more conservative variable selection, and possibly to improved VSP. We compared the VSP resulting from the standard approach to bootstrapping to the VSP resulting from nested selection of $\lambda$.

We did this comparison in simulated and empirical data. The simulated data were generated to resemble data observed in statistical genetics. The simulated data were high-dimensional, with a small number of weak predictors and a large number of noise predictors. The empirical data were drawn from a genome-wide association study (GWAS). We made this choice because of the prominence of genetics as a context for the application and development of the Lasso (Waldron et al. [2011], Lange et al. [2014]). This context differs slightly from that of meta-regression, in which small sample size is an important concern; this will be addressed in the simulations described in Chapter 4.

3.1 Approach

Estimation of SEs and CIs for nonzero Lasso coefficient estimates may provide a way to control the Lasso’s VSP as well as to assess the relative importance of predictors. In general, Lasso SEs or CIs cannot be estimated using a closed form (Osborne et al. [2000]). Consequently, a variety of approaches has been tried to estimate Lasso SEs and CIs: see reviews in e.g. Lockhart, Taylor, Tibshirani, and Tibshirani [2013], Chatterjee (2011), Kyung, Gill, Ghosh, and Casella (2010). Bootstrap estimation has received substantial interest, but many issues remain less explored, most importantly, the effects of choosing $\lambda$ through cross-validation when using the bootstrap.

Next, we briefly review the two most common approaches to bootstrapping Lasso SEs and CIs: vector bootstrapping and residual bootstrapping. We follow this with a selective review of applied and methodological research using these approaches. We
review both approaches to show that the behavior of the residual bootstrap has been studied in detail, but that comparatively little methodological research has been done on the vector bootstrap, despite its popularity in applied research. Consequently, this chapter focuses on the vector bootstrap.

3.1.1 Vector and Residual Bootstrapping

In nonparametric bootstrapping, samples are repeatedly drawn from observed sample data. The statistic of interest is calculated in each bootstrap sample. In this case, Lasso coefficient estimates are calculated, leading to an approximate sampling distribution. The approximate sampling distribution of Lasso coefficient estimates then permits calculation of SEs and CIs (Efron and Tibshirani [1994]).

We denote an estimate of the sampling distribution for the Lasso coefficient for predictor \( j \) as \( \hat{F}(\hat{\beta}_j) \). \( \hat{F}(\hat{\beta}_j) \) represents the marginal distribution of \( \hat{\beta}_j \) values, as estimated using the nonparametric bootstrap. Two ways to use nonparametric bootstrapping to find \( \hat{F}(\hat{\beta}_j) \) are vector bootstrapping and residual bootstrapping (Sartori [2009]).

3.1.1.1 The vector bootstrap

Vector bootstrapping begins with an observed sample of \( N \) observations (i.e. studies) measured on \( p \) predictors (i.e. SLCs) \( x_1, \ldots, x_p \) and an outcome \( y \) (i.e. the studies’ effect sizes). Each observation in the sample is considered as a row vector \( z_i = (x_{i1}, \ldots, x_{ip}, y_i) \), which consists of \( p \) predictor values \( x_{ij} \) and a single outcome value \( y_i \). A Lasso model can be fit to every bootstrap sample of \( N \) observations \( z_i \), yielding a set of Lasso coefficient estimates. \( \hat{F}(\hat{\beta}_j) \) is defined as the distribution of \( \hat{\beta}_j \) values from each of all possible bootstrap samples of size \( N \). The number of unique bootstrap samples increases faster than exponentially in \( N \); to save computa-
tion time in practice, $\hat{F}(\hat{\beta}_j)$ is estimated using Monte Carlo simulation. The Monte Carlo estimate of $\hat{F}(\hat{\beta}_j)$ is written $\hat{F}^*(\hat{\beta}_j)$.

3.1.1.2 The residual bootstrap

Residual bootstrapping begins by fitting a linear regression model to a sample of $N$ observations measured on $p$ predictors $x_1, \ldots, x_p$ with outcome $y$, and generating the $N$ residuals $\hat{\varepsilon}$. Residual bootstrapping uses the sampling distribution of residuals to simulate the distribution of $y$ values about their conditional means $X\beta$. Importantly, this requires treating the observed predictor values $x_1, \ldots, x_p$ as fixed and assuming that only the correct predictors are in the model [Efron and Gong, 1983], likely an inappropriate assumption in the context of variable selection. The residuals are then resampled.

Each bootstrap sample of $N$ residuals, stored in the vector $\hat{\varepsilon}^*$, can be used to generate $N$ outcomes $y^*$, defined as $y^* = X\hat{\beta} + \hat{\varepsilon}^*$. The $y^*$ values are regressed on $X$ using the Lasso, yielding, as in vector bootstrapping, a set of Lasso coefficient estimates for each resample. This set of coefficient estimates is used to define $\hat{F}(\hat{\beta}_j)$, which is typically approximated through Monte Carlo simulation, as $\hat{F}^*(\hat{\beta}_j)$.

3.1.2 Previous research in bootstrapping the Lasso

3.1.2.1 Research with residual bootstrap

Residual bootstrapping of Lasso SEs and CIs has received more methodological research interest than has vector bootstrapping. Detailed investigations of the residual bootstrapped Lasso have been undertaken by Chatterjee (2011), Minnier, Tian, and Cai (2011), Kyung, Gill, Ghosh, and Casella (2010), Bach (2009), and Knight and Fu (2000), among others, with the theoretical results in Chatterjee (2011) synthesizing much of the previous work.
By comparison, the behavior of Lasso SEs and CIs under vector bootstrapping has been under-studied, particularly with $\lambda$ selected through cross-validation.

3.1.2.2 Research with vector bootstrap

The vector-bootstrapped Lasso has often been applied in statistical genetics, sometimes with $\lambda$ selected through cross-validation. Further, the vector-bootstrapped Lasso is closely related to several other prominent variable-selection methods proposed in statistical genetics (Valdar et al., 2012; Cho et al., 2010; Motyer et al., 2011).

D’Angelo, Rao, and Gu (2009) proposed using vector resampling to estimate SEs of Lasso coefficients of SNP-SNP and gene-gene interaction terms. Sartori (2009) compared residual and vector bootstrapping of Lasso CIs and SEs in the context of statistical genetics, including the selection of $\lambda$ through cross-validation before resampling. She found that: 1) residual and vector bootstrap SEs of Lasso coefficient estimates had similar degrees of bias in linear models; and 2) vector bootstrap CIs had superior coverage in linear and in logistic models. The present study builds on this comparison by comparing versions of the vector bootstrap with and without cross-validation nested within bootstrap replications. Both of these methods are straightforward to implement in existing statistical software. This makes them attractive and approachable to applied researchers, who would benefit from understanding the tradeoff in VSP involved in choosing one over the other.

3.1.2.3 Research with cross-validated selection of $\lambda$

Many methods have been proposed to increase the VSP of Lasso regression and related methods (Chatterjee, 2011; Lockhart et al., 2013; Fan and Li, 2012). However, little methodological research has addressed the combination of vector bootstrapping and cross-validated selection of $\lambda$ that is used in practice (Sartori 2009; Cho et al.)
Despite the suggestion, in a recent textbook, that cross-validated selection of \( \lambda \) should be nested within bootstrap samples when applying the Lasso (Bühlmann et al., 2011), little or no published work has evaluated this procedure. The goal of the present chapter is to address this deficiency and to investigate the conditions in which nested selection of \( \lambda \) leads to improved VSP.

### 3.1.3 Role of \( \lambda \) when fitting Lasso models

The metaparameter \( \lambda \) controls the bias and parsimony of a fitted Lasso model. Equation 3.1 gives the definition of a Lasso model for predictors \( X \) and outcome \( y \) when \( \lambda \) is known (Tibshirani, 1996).

\[
\hat{\beta} = \arg\min_\beta \frac{1}{2} \sum_i \left( y_i - \sum_j x_{ij} \beta_j \right)^2 + \lambda \sum_j |\beta_j| \tag{3.1}
\]

The value of \( \lambda \) determines the degree of shrinkage toward 0 and serves as a threshold for variable selection. A predictor is selected if the absolute value of its covariance with the outcome is larger than \( \lambda \), and excluded otherwise (Efron et al., 2004). This thresholding property limits the useful range of values that \( \lambda \) can take. The minimum value that \( \lambda \) can take is 0, where the Lasso fit is the same as that of OLS regression. Such solutions are unbiased, but, because of the improbability of any OLS coefficients equaling 0 exactly, they are also unparsimonious.

The maximum value that \( \lambda \) can take depends on the largest sample covariance of any predictor with the outcome. More specifically, when \( \lambda \) is equal to or greater than that covariance, all coefficients are shrunk to 0, and the fitted model is intercept-only, thus parsimonious but biased (Friedman et al., 2007).
3.1.3.1 Selection of $\lambda$ through $K$-fold cross-validation

In general, each value of $\lambda$ is associated with a single Lasso model (Tibshirani, 2013; Efron et al., 2004). $K$-fold cross-validation (Zhang, 1993) is often used to select the best-performing model. Lasso model fitting, $\lambda$ selection, and $K$-fold cross-validation has been described in detail for its implementation in the R package glmnet (Friedman et al., 2010b).

Following this procedure, the $\lambda$ value that is associated with the best-performing model is selected. The best-performing model is the one having the minimum cross-validation index, which is computed as the sum of squared residuals averaged over the $K$ cross-validations. The selected $\lambda$ value is then used to fit a finalized model by solving Equation 3.1 in the entire sample. This produces a set of selected predictors that are then indexed in set $s$.

Different $\lambda$ values might be selected in different samples from the same population due to the influence of noise correlations (Fan and Li, 2012). In the next section, we interpret selection of $\lambda$ as a source of variation in Lasso coefficient estimates.

3.1.3.2 Lasso variance estimates: Contribution of $\lambda$ selection

The variance of Lasso coefficient estimates depends on the joint distribution of the $p$ predictor variables $X$ and the outcome $y$, as well as on the value of $\lambda$ (Pötscher and Leeb, 2009). The conditional distribution of estimates for a single predictor, denoted $g_j(\hat{\beta} \mid \lambda)$, is the distribution of $\hat{\beta}_j$ coefficients at a fixed $\lambda$ value. The marginal distribution, $h_j(\hat{\beta})$, is the distribution of $\hat{\beta}_j$ averaged over $\lambda$ values.

The variance of $\hat{\beta}_j$ can be found using either $g_j$ or $h_j$. Heuristically, $h_j$ treats the selected $\lambda$ value as a realization of a random variable (Zhang, 1993). We argue that using $h_j$ might improve VSP because using $g_j$ treats $\lambda$ as fixed, which can underestimate the variance of coefficients.
To support this claim, consider the inequality:

\[
\begin{align*}
V(\hat{\beta}_j) &= \mathcal{E}_\lambda \{ V_\beta (\hat{\beta}_j|\lambda) \} + \mathcal{V}_\lambda \{ \mathcal{E}_\beta (\hat{\beta}_j|\lambda) \} \\
&\geq V_\beta (\hat{\beta}_j|\lambda)
\end{align*}
\] (3.2)

(Chatfield 1995; Casella and Berger 2002). If \( \hat{\beta}_j \) and \( \lambda \) were independent, then \( V(\hat{\beta}_j) = V_{\beta}(\hat{\beta}_j|\lambda) \), and there would be little difference between the fixed and random \( \lambda \) approaches in practice. However, \( \hat{\beta}_j \) and \( \lambda \) are not necessarily independent: the range of possible \( \lambda \) values is bounded by \((0, \text{cov}_{\text{max}})\), where \( \text{cov}_{\text{max}} \) is the largest sample covariance between any predictor and the outcome. Thus, although using \( g_j \) (treating \( \hat{\beta}_j \) and \( \lambda \) as independent) has the practical advantage of using fewer computational resources, it will only be acceptable if the resulting underestimate of the standard error of \( \hat{\beta}_j \) is small.

3.1.3.3 Lambda selection in bootstrapping

In practice, both \( g_j \), the conditional distribution of \( \hat{\beta}_j \) given \( \lambda \), and \( h_j \), the distribution of \( \hat{\beta}_j \) averaged over all \( \lambda \)s, are unknown. Both distributions can be estimated using the vector bootstrap.

Finding the bootstrap estimate of the conditional distribution, \( \hat{g}_j^* \) is done by fitting Lasso models to resampled \( X \) and \( y \) values, given the \( \lambda \) value chosen through \( K \)-fold cross-validation in the original sample.

Finding the bootstrap estimate of the marginal distribution \( \hat{h}_j^* \) requires treating the selected \( \lambda \) value as random. Nesting \( \lambda \)-selection within each bootstrap replication approximates the effect of sampling error on the value of \( \lambda \) selected.

Our simulations compared the fixed- and random-\( \lambda \) approaches with respect to VSP, and suggest effect sizes at which the increased computational burden of the random approach is worthwhile.
3.2 Methods

The purpose of the current chapter is to propose and to evaluate the use of the vector bootstrap, with \( \lambda \) selected through \( K \)-fold cross-validation, as a method for estimating Lasso CIs. In particular, we compared three variants of this approach: a software default approach to variable selection using the Lasso (Method 3.1); an approach involving selection of \( \lambda \) before resampling (Method 3.2); and a third approach where \( \lambda \)-selection is nested within bootstrap samples (Method 3.3). Our evaluation was in terms of VSP.

In the first step of each variable selection method, a Lasso model is fit to the entire sample. This requires selection of \( \lambda \), which is done through \( K \)-fold cross-validation. The initial model fit produces a set of selected predictors, which are indexed in the set \( s \). This step is the default application of the Lasso in the \texttt{R} packages \texttt{glmnet} and \texttt{grpreg}. We denote it Method 3.1.

---

Given \( N \) individuals measured on \( p \) standardized predictors \( x_j, \ j = 1, \ldots, p \), with outcome \( y \):

1. Fit Lasso model to select predictors
   1. Identify candidate \( \lambda \) values using sample covariances
   2. Use \( K \)-fold cross-validation (\( K = 10 \)) to select finalized \( \lambda \)
2. Finalized \( \lambda \) selects predictors, indexes them in \( s \subset \{1, \ldots, j\} \)
3. Exclude predictors \( x_k \), where \( k \notin s \)

Method 3.1. Default approach to Lasso variable selection using \( K \)-fold cross-validation.

Methods 3.2 and 3.3 differ from Method 3.1 by having a second variable selection
step. In this step, further reduction of the set of selected predictors is done using bootstrap CIs. All predictors are used in vector bootstrap resampling, but, to save computational resources, CIs are not calculated for predictors that were excluded in the initial variable selection step.

1. Fit Lasso models, select set of predictors $s$ as in Method 3.1
2. Estimate marginal sampling distributions $\hat{F}^*\left(\hat{\beta}_k\right)$ for $k \in s$
   1. Draw $B$ vector bootstrap samples ($B \geq 1000$)
   2. In each bootstrap sample:
      1. Fit a Lasso model with $\lambda$ at finalized value from Step 1 (Fixed $\lambda$)
      3. Use $\hat{F}^*\left(\hat{\beta}_k\right)$ to calculate mean, SE, and CI for each $\hat{\beta}_k$
   4. If CI for $\hat{\beta}_k$ includes 0, exclude $x_k$

---

Method 3.2. Vector bootstrap for improved Lasso variable selection precision with $\lambda$ treated as fixed.

Method 3.2 uses the same value of $\lambda$ in every bootstrap sample. Method 3.3 differs from Method 3.2 by re-selecting $\lambda$ in each bootstrap sample. In both methods, after CIs are calculated, predictors that include 0 in their confidence intervals are excluded.

3.2.1 Lasso CIs: Secondary selection of predictors

Bootstrap CI or SE estimates improve Lasso models through a second step of selecting or ranking predictors. CI and SE estimates are both directly related to the sampling variance of a Lasso coefficient estimate, discussed above. A $1 - \alpha$ CI
1. Fit Lasso models, select set of predictors $s$ as in Method 3.1
2. Estimate marginal sampling distributions $\hat{F}^*\left(\hat{\beta}_k\right)$ for $k \in s$

1. Draw $B$ vector bootstrap samples ($B \geq 1000$)
2. In each bootstrap sample:
   1. Fit a Lasso model exactly as in Step 1 (Random $\lambda$)
3. Use $\hat{F}^*\left(\hat{\beta}_k\right)$ to calculate mean, SE, and CI for each $\hat{\beta}_k$
4. If CI for $\hat{\beta}_k$ includes 0, exclude $x_k$

---

Method 3.3. Vector bootstrapping with metaparameter $\lambda$ treated as random.

for the coefficient estimate $\hat{\beta}_j$ is generated either using the $\frac{\alpha}{2}$, $1 - \frac{\alpha}{2}$ quantiles of the bootstrap distribution $\hat{g}_j^*$, or using an approximate inverted $z$-test, which gives the interval $\hat{\beta}_j \pm z_{\alpha/2}SE^*\left(\hat{\beta}_j\right)$, where $z_{\alpha/2}$ is the $\frac{\alpha}{2}$ quantile of a standard normal distribution and $SE^*\left(\hat{\beta}_j\right)$ is the bootstrap estimate of the standard error of $\hat{\beta}_j$.

Using either CI method, predictors that have CIs that contain 0 are excluded since this can be regarded as evidence that predictor $x_j$ is a false positive selection.

In the next section, we describe simulations that were used to evaluate Methods 3.1–3.3

3.3 Simulation Studies

We first compared Methods 3.1–3.3 using a factorial simulation study. The simulation had two goals: first, evaluating the Methods’ ability to distinguish signal from noise; second, evaluating their ability to correctly order signals of differing strengths.

The data generation models were as simple as possible while still representing two
empirically interesting scenarios based on statistical genetics: 1) a low probability of selecting important predictors at random; and 2) a spectrum of small true effect sizes. To this end, data was generated under a linear model with under 5% of predictors having true effects, and with effect sizes (given in $r^2$) representing 2.5% or less of outcome variance attributable to any important predictor.

2500 Monte Carlo (MC) replications were used in each cell of the design. This number of replications was chosen based on pilot studies, in which at least 2500 replications were required in order to generate relatively smooth empirical distributions of coefficient estimates (see also Sartori 2009).

$B = 1000$ bootstrap replications were used within each MC replication. The average performances of the three methods across samples were compared; within each MC replication, each method was employed on an independent sample drawn from the population distribution. This was done in order to avoid creating dependence among results that might arise from fitting the methods to the same data.

The next sections describe the simulation design and outcome in more detail.

3.3.1 Simulation Design

Three factors were manipulated in the simulations: method, data-generating model, and effect size.

As described above, the methods compared were the fixed- (Method 3.2) and random-λ (Method 3.3) variants of the vector bootstrapped Lasso, with the default application, Method 3.1.

The second factor manipulated in the simulation study was the data-generating model. Two data-generating models were used: one with 1 important predictor and 99 noise predictors; and the other with 5 important predictors and 99 noise predictors, with the important predictors having different $r^2$ values, enabling us to rank them.

Each data generating model was a linear regression model having a standard
normal outcome and binomial(2, 0.5) distributed predictors; N = 5000 was used as the sample size, as in Hoffman, Logsdon, and Mezey (2013). This was chosen as a rough approximation of the sample size and predictor structure of genome-wide association studies of quantitative phenotypes (Balding, 2006). However, a relatively small number of predictors was used in order to maintain the computational feasibility of bootstrapping with random $\lambda$.

The third factor manipulated in the simulation study was effect size. Effect sizes of $r^2 = 0.01, 0.0033, 0.001$ were used in the single-important-predictor analyses. In the five-important-predictors analyses, each important predictor had a different effect size: the set $r^2 = 0.01, 0.0067, 0.0033, 0.0022, 0.001$ was used.

We manipulated the effect size of important predictors for two reasons: we used effect size as a measure of the “difficulty” of correctly selecting important predictors, giving us a way to use the data to influence the methods’ VSP; and because previous simulation studies (e.g. Tibshirani, 2011; Leng, Lin, and Wahba, 2006; Meinshausen and Bühlmann, 2010), used effect sizes that are now considered to be unrealistically large in the context of statistical genetics (Stefansson et al., 2009; Park et al., 2011b).

In the next section, we describe the evaluation criteria.

3.3.2 Evaluation Criteria

The main question asked in this chapter is: when does nested selection of $\lambda$ lead to improved VSP over other approaches?

Addressing these questions requires quantifying the performance of the different methods.

We used variable selection precision (VSP) to quantify the methods’ performance.

\[
\text{VSP} = \frac{\# \text{Important Predictors Selected}}{\# \text{Selected Predictors}} = \frac{\# \text{True Positives}}{\# \text{Positives}}.
\]
VSP is set to 0 if there are no positives. Thus, in each replication of the factorial simulation, VSP ranged from 0, $\frac{1}{104}$, $\frac{1}{103}$, \ldots, $\frac{1}{12}$, 1.

A positive in the vector bootstrap Lasso (Methods 3.2 and 3.3) was defined as a predictor for which the $(1 - \alpha) \times 100\%$ bootstrap percentile confidence interval (Efron and Tibshirani, 1994) excluded 0. $\alpha$ values of 0.02, 0.05, and 0.20 were used. A positive in the default Lasso (Method 3.1) was defined as a predictor having a coefficient in the finalized Lasso model (i.e. a predictor indexed in $s$).

3.3.3 Simulation Results

3.3.3.1 Improved Variable Selection Precision

Use of the vector bootstrap (Methods 3.2 and 3.3) was associated with increased VSP at all effect sizes. This is shown in Table 3.1 for nominal $\alpha = 0.05$. At the smaller effect sizes, this advantage was only apparent with random $\lambda$, and, at the smallest effect size, the magnitude of this advantage was small. The nominal coverage rate $\alpha$ chosen for confidence intervals interacted with the bootstrapping method used. For example, at $r^2 = 0.0033$, the random approach (Method 3.3) is more precise than the fixed approach (Method 3.2) at $\alpha = 0.01, 0.05$ (Table 3.1), but less precise at $\alpha = 0.10$ (not shown).

The results of the simulation studies quantified the methods’ relative performance in idealized data. For a more critical evaluation of their practical utility, we applied them in a GWAS data set, using them to select pairwise interactions as well as main effects.

3.4 Empirical Illustration

We used data gathered for a genome-wide association study (GWAS) of Borderline Personality Disorder features to compare the vector bootstrap with $\lambda$-selection
### TABLE 3.1

VARIABLE SELECTION PRECISION OF DEFAULT LASSO VS. BOOTSTRAPPED PERCENTILE CIS

<table>
<thead>
<tr>
<th>$r^2$</th>
<th>Method</th>
<th>Default</th>
<th>Fixed $\lambda$</th>
<th>Random $\lambda$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.3854</td>
<td><strong>0.9797</strong></td>
<td>0.8885</td>
<td></td>
</tr>
<tr>
<td>0.0033</td>
<td>0.35</td>
<td>0.5185</td>
<td><strong>0.7149</strong></td>
<td></td>
</tr>
<tr>
<td>0.001</td>
<td>0.1284</td>
<td>0.0642</td>
<td><strong>0.1766</strong></td>
<td></td>
</tr>
</tbody>
</table>

Bold text indicates the largest VSP in each row. Nominal $\alpha = 0.05$. The random-$\lambda$ bootstrap CI (Method 3.3) outperforms the default at each effect size, but improvement is marginal for the very smallest effects nested within bootstrap samples to the standard Lasso. This comparison serves as a representative analysis for possible applications of Method 3.3. The original GWAS was based on a sample of $N = 7124$ individuals who participated in a twin-family study of mental and somatic health [Boomsma et al., 2006; Willemsen et al., 2013]; see Willemsen et al. for detailed methods including IRB approval, genotyping, and quality control procedures. Responses to a psychiatric inventory measuring Borderline Personality features were used as outcomes because they have shown a promising signal that was replicated in an independent sample [Lubke et al., In Press].

Borderline features were measured using total scores on the PAI-BOR inventory, a 24-item test [Morey, 1991]. More specifically, the outcome we used in this study was the residual of PAI-BOR score after OLS regression on age, gender, and their interaction, as well as a principal component score representing ancestry [Abdellaoui et al., 2013; Price et al., 2010]. Following up on D’Angelo, Rao, and Gu (2009)’s proposal, we fit Lasso multiple regressions of Borderline features on SNP main effects and SNP-SNP interactions. To control the computational resources required, we
limited the analysis to pairwise interactions and main effects of the 125 SNPs having the strongest univariate association with the Borderline features phenotype, as listed in Lubke et al. In Press. R’s memory limitations limit the application of bootstrapping to interactions between 1500 or fewer variables (1500$^2 \times 1000$ bootstrap samples $\approx 2^{31}$ objects) (R Core Team, 2013). To avoid multicollinearity, these 125 were then pruned to the set of 77 SNPs that had pairwise correlations of less than $r = 0.6$ among each other. From these, 2926 pairwise interaction terms were calculated, resulting in a total $p = 3003$, and hence up to 3003 Lasso partial regression weights needing CIs.

Results from applying the vector bootstrap with random $\lambda$ (Method 3.3) were compared to those from the default Lasso (Method 3.1). Method 3.3 was used to generate percentile CIs. The resulting variable selections and rankings were compared to Method 3.1’s selections.

3.4.1 Empirical illustration: Results

The empirical illustration concerned the application of the bootstrapped Lasso to pairwise interaction effects. The bootstrapped Lasso produced different variable selections and importance rankings that did the default Lasso. The bootstrapped Lasso tended to select better quality predictors than did the default Lasso. The default Lasso selected predictors with low minor allele frequency (MAF), hence had very large standard error estimates. This suggests that the default Lasso can ignore important aspects of the data.

Figure 3.4 and Table 3.2 present comparisons of the default Lasso and the approach using vector bootstrapping. Figure 3.4 plots bootstrap mean estimates on the horizontal axis and bootstrap standard error estimates on the vertical axis. Points falling outside of the dark gray $V$ have bootstrapped $t$-statistics greater than 2. Predictors that were selected by the default Lasso are plotted as light gray diamonds. There is no obvious relationship between default Lasso selection of a predictor and
Figure 3.4. Bootstrap means and SEs of 77 SNPs, 2926 pairwise interactions; light gray diamonds represent predictors selected without bootstrapping; The lines represent mean= $\pm 2 \times SE^*$: points and diamonds outside the V-shape (i.e. in lower corners) are promising signals. A cube-root transformation was used on both axes.

its bootstrap moments. dbSNP look-up of the predictors in Table 3.2 showed that the bootstrapped Lasso was less prone to selecting interactions between SNPs having low MAF than was the default Lasso. The default Lasso, in selecting these interactions, was in effect including interactions between binomial predictors that have low success probabilities. These interactions tended to have large bootstrap standard errors. Interestingly, the low-MAF SNPs involved these interactions tended to have moderately strong main effects in the conventional GWAS analyses.

Using vector bootstrapping of Lasso coefficients suggested a single interaction for followup.
TABLE 3.2

PROMISING SNP-SNP INTERACTIONS

<table>
<thead>
<tr>
<th>Chrs</th>
<th>rsIDs</th>
<th>MAFs</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>16, 1</td>
<td>rs118160379×rs199104015</td>
<td>.05, .27</td>
<td>[−.078, −.002]</td>
</tr>
</tbody>
</table>

Selected by default but rejected by bootstrap:

<table>
<thead>
<tr>
<th>Chrs</th>
<th>rsIDs</th>
<th>MAFs</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>9, 4</td>
<td>rs112188788×rs139344595</td>
<td>.02, .01</td>
<td>[0, .501]</td>
</tr>
<tr>
<td>6, 1</td>
<td>rs117266484×rs73008417</td>
<td>.01, .01</td>
<td>[0, .562]</td>
</tr>
<tr>
<td>9, 1</td>
<td>rs112188788×rs73008417</td>
<td>.02, .01</td>
<td>[−2.904, 0]</td>
</tr>
<tr>
<td>12, 9</td>
<td>rs117256451×rs112188788</td>
<td>.02, .02</td>
<td>[−.056, .451]</td>
</tr>
</tbody>
</table>

3.5 Conclusion

Using vector bootstrap CIs on Lasso regression coefficients offers a valid way to distinguish false positive selections from true positives. Percentile CIs were associated with increased precision of variable selection at all effect sizes. At the smallest effect sizes, gains were only achieved when using Method 3.3 which treated the meta-parameter $\lambda$ as random. This suggests that, if small effects are expected, treating $\lambda$ as random justifies increased computational cost.

The default Lasso was better able to detect small effects when larger effects were present than when solitary small effects were considered. We observed this “rising tide lifts all boats” effect in pilot simulations, considering that the more complex models tended to cause lower (more lenient) thresholds $\lambda$ to be selected by cross-validation. A possible explanation is that a $\lambda$ causing inclusion of a solitary small effect might not be able to consistently decrease the residual sum of squares in different cross-
validation subsamples, but that a $\lambda$ that admits multiple small effects could.

In the empirical analysis, the vector-bootstrapped Lasso excluded unreliable predictors that had been selected by the default Lasso. Two follow-up studies are suggested by this result: a simulation study comparing the two Lasso methods after manipulating the reliability of predictors; and an attempt to replicate the promising interaction between rs118160379 and rs59194015.

There were several limitations to this study. The data generating model had predictors that were independently and identically distributed as well as a normally distributed outcome. These attributes are unlikely to hold in empirical data. We plan to extend the current simulations with correlated and differently scaled predictors as well as skewed outcomes. Second, the simulated effect sizes, while small, were still somewhat larger than those that are typically observed in the statistical genetics of complex traits (Stefansson et al., 2009). The number of important predictors used was also much smaller than the number of genetic loci expected to influence complex phenotypes (Sivakumaran et al., 2011). Finally, the argument used to justify nesting $\lambda$-selection within bootstrapping was intuitive. A more rigorous argument might be able to identify specific conditions on $X$ or $y$ that would lead to Method 3.3 consistently outperforming Method 3.2 or vice-versa.

Vector bootstrapping CIs of Lasso coefficients led to increased variable selection precision, especially at small effect sizes. Our illustration with empirical data showed that this is also an effective approach to select important interactions between predictors. In consequence, vector bootstrapping CIs is a very promising approach for identifying sets of SNP-SNP and SNP-environment interactions.

There are several implications of these results that are especially relevant to meta-regression, which is characterized by much smaller sample sizes. First, in some meta-regressions, the sample size may be so small that Monte Carlo bootstrapping within $K$-fold cross-validation may have many tied observations, leading to inappro-
appropriate bootstrap quantiles. This problem is expected to occur for sample sizes of $N/K \leq 5$ and can be addressed by using the exact distribution, $\hat{F}(\hat{\beta}_j)$, rather than $\hat{F}^*(\hat{\beta}_j)$ (Fisher and Hall, 1991, p. 161). Second, the small sample sizes observed in meta-regression suggest that only large effect sizes can be detected. This slightly reduces the applicability of the results presented in this chapter; however, we expect much larger effect sizes in meta-regression than in SNP-phenotype associations.

We also propose two extensions of this method for use with meta-regression: 1) using the SCAD estimator; and 2) applying shrinkage regression to linear heteroscedasticity models. The SCAD (Smoothly Clipped Absolute Deviation) estimator is a shrinkage regression method which is similar to the Lasso but does not apply shrinkage to large coefficient estimates (Fan and Li, 2001). In SCAD estimation, coefficient estimates with absolute values greater than $M\lambda$ (where $M$ is a user-set constant, conventionally set at 3.7) are not shrunk and are hence estimated without bias (assuming independent predictors). Because the SCAD does not apply downward bias to the “best” predictors, it may outperform the Lasso as a tool for meta-regression.

Applying shrinkage regression to linear heteroscedasticity models may be a way to identify SLCs that affect the variance of effect sizes between studies. We assume that these SLCs are uncorrelated with SLCs that affect the conditional mean value of effect sizes. This enables us to use the linear heteroscedasticity model as an alternative way of estimating the contribution of random effects to heterogeneity. Because of the bias induced by applying shrinkage, the resulting variance estimates (e.g. Amemiya (1977, 1978)) will also be biased. However it is quite possible that shrinkage regression can be used to identify SLCs that influence heterogeneity with precision. We evaluated this claim in simulation studies.

In the following chapter, we describe the simulation experiments used to compare the Lasso and the SCAD to simple meta-regression, and to evaluate shrinkage regression in the linear heteroscedasticity model.
CHAPTER 4

SIMULATION STUDIES: CHARACTERISTICS OF SAMPLE DATA SETS THAT INFLUENCE META-REGRESSION PERFORMANCE

4.1 Introduction: Comparing methods for selecting variables in meta-regression

Meta-regression is used to identify study-level covariates (SLCs) that are associated with heterogeneity of effect sizes between the studies in a meta-analysis. In newer fields, such as psychiatric genomics, the pool of SLCs that may be used for meta-regression is large, approaching or exceeding the number of studies that can be analyzed. In this situation, the choice of which SLCs to include in a meta-regression is an open question that can be addressed using automatic methods for variable selection.

In this text, we have introduced three methods for automatic variable selection: 1) simple regression of effect size on each SLC separately, with selection based on hypothesis testing; 2) the Lasso, a shrinkage regression method; and 3) the SCAD, which is similar to the Lasso but provides unbiased estimates of large regression coefficients.

In this chapter, we describe simulations that were used to compare these methods with respect to their variable selection precision (VSP). VSP is a concise representation of the criterion that a good variable selection method should make trustworthy selections. VSP is calculated as the proportion of important SLCs out of selected SLCs. Hence, VSP is a function of the number of true positive selections and of the number of false positive selections. Consequently, in the simulations, we varied
factors that are likely to influence the probability of selecting important SLCs: the number of studies to be included in the meta-analysis, denoted $N$, and the squared correlation between a SLC and studies’ effect sizes, denoted $r^2$. We also varied a factor that is likely to influence the probability of selecting unimportant SLCs: the total number of SLCs considered, denoted $p$.

We also intend to generalize the simulations’ results to empirical meta-analyses. In empirical meta-analyses, studies are weighted using the standard error (SE) of their effect size estimates, denoted $s$. The value of $s_i$ is determined by individual differences between the participants in study $i$ as well as the study’s size (the number of participants). In psychiatric genomics, as in other fields, studies tend to have differing sizes. Because of this, we varied a factor that represents the pattern of weights across studies. Another feature of empirical meta-analyses is the possible presence of unobserved random effects. Random effects cause additional heterogeneity of effect sizes across studies. Because of this, we also performed a set of simulations having random effects. In some of these, we simulated these effects under a linear heteroscedasticity model.

4.2 Design of the simulations

4.2.1 Data generation: a simple model

The data generating model was linear with standard normal outcome, zero intercept term, and independent normal sampling error. A single important SLC was simulated. The remaining $p-1$ SLCs were simulated as i.i.d. standard normal variables. Equation 2.1 modified for the simulation, is:

$$y = X\beta + \epsilon = X_{1}\beta + \epsilon + X_{2,...,p}0.$$  \hspace{1cm} (4.1)
\( \mathbf{y} \) is an \( N \)-element column vector of studies’ effect sizes. \( \boldsymbol{\varepsilon} \) is an \( N \)-element random vector of within-study errors that represents the effect of sampling error on measuring the studies’ effects. \( \mathcal{E}(\boldsymbol{\varepsilon}) = \mathbf{0}. \ \mathcal{V}(\boldsymbol{\varepsilon}) = s\mathbf{I} \). \( \mathbf{X} \) is an \( N \times p \) matrix of SLCs with columns \( \mathbf{X}_j \) and rows \( \mathbf{x}'_i \). \( \mathbf{X}_1 \) is the single important SLC, and has coefficient \( \beta \), while the unimportant SLCs, \( \mathbf{X}_{2,...,p} \), have zero association with \( \mathbf{y} \).

We chose this model for two reasons: 1) effect size measures are often normally distributed, either by design (e.g., Hedges’ \( g \), \( z \)-transformed correlation coefficient), or because they are computed as sums or averages of individual observations and the Central Limit Theorem applies; 2) using a single important predictor means that the most important differences in VSP across conditions will be due to differences in numbers of false positive selections.

4.2.2 Factors that influence the probability of correct SLC selection

4.2.2.1 The sample size factor

Sample size, \( N \), is the number of simulated studies that were included in each meta-regression. We manipulated this factor because correct selection of important predictors in any of the methods depends on \( N \): if the true association, \( \beta \), when standardized, is below \( 1/\sqrt{N} \), the shrinkage methods are unlikely to detect it reliably, while the simple regression approach will be under-powered if the true \( \beta \) is below \( 1.28/\sqrt{N} - 3 \) (Pötscher and Leeb 2009; Hedges and Olkin 1985).

However, rather than setting \( N \) to ensure a certain power to detect the true effect, \( N \)-s were chosen to represent typical numbers of studies included in meta-analyses and simulation studies of meta-analyses. We used \( N = 6, 18, 54, 100 \). These sample sizes fall within the range of several recent meta-analyses in psychiatric genetics, which range in size from small \( (N \leq 10) \), e.g. (Amin et al. 2012; Major Depressive Disorder Working Group of the Psychiatric GWAS Consortium 2012), to moderate \( N = 10–20 \), e.g. (Anttila et al. 2013; de Moor et al. 2010), to large \( N \geq 50 \), e.g. (Berndt...
These sample sizes also overlap with those used in simulations of meta-regression, such as Huizenga et al. (2011), Bondell et al. (2010) and López-López et al. (2014).

4.2.2.2 The association strength factor

The strength of association, $r^2$, gives the proportion of heterogeneity in effect size that is explained by the important predictor $X_1$. It is given by dividing the squared regression coefficient by the weighted variance of $X_1$:

$$r^2 = N \frac{\beta^2 \sum_i w_i (x_{i1} - \bar{X}_1)^2}{\sum_i w_i (y_i - \bar{y})^2},$$

where $\beta$ is the meta-regression coefficient, $x_{i1}$ is the observed value of $X_1$ in study $i$, $\bar{X}_1$ is the weighted mean of $X_1$ values, the weights are:

$$w_i = (s_i^2)^{-1} / \sum_i \left\{ (s_i^2)^{-1} \right\},$$

where $s_i$ is the known standard error of effect size $y_i$, and $\bar{y}$ is the weighted mean of effect sizes.

We chose to manipulate this factor because it is directly related to the expected probability of selecting the important predictor. In simple meta-regression, the procedures in Hedges and Pigott (2004) can be used to calculate this expectation exactly. For the Lasso and the SCAD, it is clear that as $r^2$ increases, so must the chance of selecting the true important predictor, and we use simulation to estimate the expected probability of making this selection.

We used $r^2$ values of 0.2 and 0.5. These values were chosen to represent effects from marginally detectable to likely detectable, and are close to the values used in other meta-regression simulations (López-López et al., 2014).

Using the procedure in Hedges and Pigott (2004), we calculated the power for a
two-sided test of simple regression with significance level 0.05 to detect the important predictor. These power values are listed in Table 4.1 and represent expected values of the probability of selecting the important predictor in simple regression.

**TABLE 4.1**

POWER TO DETECT EFFECTS USING A TWO-SIDED TEST OF
SIMPLE REGRESSION AT $\alpha = 0.05$

<table>
<thead>
<tr>
<th>$N$ $r^2$</th>
<th>0.2</th>
<th>0.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>0.054</td>
<td>0.167</td>
</tr>
<tr>
<td>18</td>
<td>0.120</td>
<td>0.567</td>
</tr>
<tr>
<td>54</td>
<td>0.304</td>
<td>0.975</td>
</tr>
<tr>
<td>100</td>
<td>0.515</td>
<td>1.000</td>
</tr>
</tbody>
</table>

4.2.3 Manipulating false positive selections: The number of SLCs factor

The number of SLCs, $p$, is the size of the pool of possible covariates considered in the simulated meta-regression. We manipulated this factor to control the expected number of unimportant variables that could be selected.

We chose $p = 5, 10, 20$ to represent numbers of SLCs that arise in different scenarios. Most published meta-analyses present a few broad demographic factors, while others synthesize studies that have very different methods, requiring many dummy variables to represent the different approaches. A small $p$ represents demographic
comparison of samples in a meta-analysis, e.g. average age, sex, ethnicity, SES, and whether the sample was clinical or population, information often given in published meta-analyses, for example the tables in Wray et al. (2012) ($p = 10$), and Ligthart et al. (2011) ($p = 8$). The larger values of $p$ are used when dummy coding is required to represent a diverse collection of studies; for example, Table 1 in Murphy et al. (2013, p. 515) can be coded into $p = 25$ variables, including 13 dummies.

We used $p\alpha$ as an estimate of the expected number of unimportant variables selected using simple regression, yielding the values in Table 4.2.

\[
\begin{array}{c c c c}
\hline
p & \# FP \\
5 & 0.25 \\
10 & 0.5 \\
20 & 1 \\
\hline
\end{array}
\]

TABLE 4.2

EXPECTED NUMBER OF FALSE POSITIVES IN MULTIPLE SIMPLE META-REGRESSION AT $\alpha = 0.05$
4.2.4 Empirical concerns: Study weights and heteroscedasticity

4.2.4.1 The study weights factor reflects the differences in study size that are observed in practice

The study weights factor, \( wts \), represents the differences in precision of estimated effect sizes that are observed in practice. It is used to make the simulation results generalizable to empirical meta-analyses. In fixed-effects meta-analysis, studies are weighted by the squared SE of their effect size estimates, \( s_i^2 \). In random-effects meta-analysis, studies are weighted by the sum of the squared SEs of their effect size estimates with the estimate of random-effects variance, \( s_i^2 + \hat{\tau} \). Both sets of weights depend on the SE of effect size estimates. SEs are largely determined by the size of a study; larger studies have more influence on a meta-regression than do smaller studies. The distribution of study sizes in meta-analyses are typically skewed (Osburn and Callender, 1992; Sanchez-Meca and Marin-Martinez, 1998). This means that published meta-regressions have a small number of influential observations. The SLC values of these influential observations have a strong impact on the estimated meta-regression coefficients. We expect that this will influence VSP: if a large study has a large effect size and an outlier value on a noise SLC, the estimated regression coefficient may be spuriously large, leading to a false positive selection.

We used two levels of \( wts \): equal and skewed. The equal-\( wts \) condition is the ideal case, representing a situation with equal study sizes. The skew-\( wts \) condition is more realistic. In the skew-\( wts \) condition, \( N \) weights were drawn from the Exponential(1) distribution, then normalized to sum to 1. We chose the Exponential(1) distribution because it has a single parameter and is skewed. It also has similar shape to the distribution of weights in the meta-analysis performed by Wray et al. (2012). This is illustrated in Figure 4.1, a quantile-quantile plot of the Wray et al. weights.

\(^1\)For some effect sizes, such as Hedges’ \( g \), the standard error is a function of the effect size—equal study sizes do not imply equal standard errors (Hedges, 1983).
Figure 4.1. A quantile-quantile plot shows the that the observed distribution of weights in Wray et al. (2012) has a similar shape to an Exponential(1) distribution. The line has a slope of 1/2.

against the quantiles of an Exponential(1) distribution.

4.2.4.2 Random effects and study weights

Another aspect of the simulation is exploring the sensitivity of variable selection to mis-specification of the meta-regression model. We did this by fitting fixed-effects meta-regression models to data that were generated to have random effects. We simulated random effects with a mean of 0 and a standard deviation of 0.64, as well as using the linear heteroscedasticity model. When generating random effects data, we used Equation 4.2 to calculate \( \beta \) values given \( N \), association strength \( r^2 \), standard errors \( s \), and random effects variance \( \tau \). In these calculations, the simulated studies were weighted by \( \omega_i = \hat{V}(y_i | x_i')^{-1} / \sum_i \{ \hat{V}(y_i | x_i')^{-1} \} \).

However, when fitting models to select SLCs, we weighted studies based on the
standard errors of their effect size estimates, which is what occurs in practice when
fixed-effects models are fit to data that has random effects. In general, for a given
analysis, the study weights based on random-effects models tend to be closer in
value to each other than those based on fixed-effects models (Berkey et al., 1995).
Failure to include the random effects variance, $\tau$, when weighting studies leads to an
incorrect pattern of weights. The studies that have the smallest standard errors will
be disproportionately up-weighted. The studies that have the largest standard errors
will be disproportionately down-weighted. If effect sizes are correlated with standard
errors, a fixed-effects analysis may lead to inaccurate estimates.

A concise expression for the difference in weighted means that results from ap-
plying two sets of weights to a data set is given by Senn (2007):

$$z_{.,w} - z_{.,\omega} = s_z \left( s_w r_{w,z} - s_\omega r_{\omega,z} \right),$$

where $z_i, i = 1, \ldots, N$, are a set of observations with standard deviation $s_z$, $w_1, \ldots, w_N$ are one set of weights, having standard deviation $s_w$, $\omega_i, \ldots, \omega_N$ are a different set of weights with standard deviation $s_\omega$, $z_{.,w}$ is the weighted mean under the $w$ weights, $z_{.,\omega}$ is the weighted mean under the $\omega$ weights, $r_{w,z}$ is the sample correlation between the $z$ and the $w$ weights, and $r_{\omega,z}$ the sample correlation between the $z$ and the $\omega$ weights. We have used the notation $z$ because Equation 4.2.4.2 applies to weighted variance estimates (the $z_i$ are squared deviations from a mean) and weighted covariance estimates (the $z_i$ are cross-products of deviations).

If, as suggested by Berkey et al. (1995), the fixed-effects weights are more variable
than the random-effects weights, then fixed-effects estimates (including estimates of
regression coefficients) may have an upward bias. This assumes that the correlations
between $z$ and the weights are approximately equal for both fixed- and random-effects
weights.
We next describe our use of a special form of the random-effects model, the linear heteroscedasticity model.

4.2.4.3 Random effects: the linear heteroscedasticity model

We simulated data under the linear heteroscedasticity model as initial step in shrinkage meta-regression for random effects, because random-effects meta-analysis is widely used in practice (Huizenga et al., 2011). In this simulation, we used the Lasso and simple meta-regression to select SLCs associated with $\mathcal{V}(y_i)$; these SLCs are denoted $U$. In the linear heteroscedasticity model, the first step is to estimate $E(y_i | x_i)$. Then, the residuals that result from this estimation are squared and regressed on $U$. When a Lasso model is fit to a data set using cross-validation, the residual variance is biased upward (Wasserman and Roeder, 2009). This means that, on average, squared residuals are biased upward, which could lead to an inflated intercept term when they are regressed on $U$. We have not explored this sufficiently to justify performing variable selection on $X$, then on $U$. To ensure that the residuals did not depend on variable selection of fixed effects, we estimated them by regressing $y$ on all of the SLCs in $X$. When there were fewer studies than SLCs in $X$, we used the rank $N - 1$ Moore-Penrose inverse of the $p \times p$ matrix $X'X$. All rank $N - 1$ generalized inverses of $X'X$ yield the same residuals when applied to a given sample (Rencher and Schaalje, 2008).

The linear heteroscedasticity model is a straightforward way to specify the strength of association in simulated relationships between SLCs and random effect variance. Strength of association is determined by the value of the coefficients in the regression of residual variance on SLCs.

We generated data using a modified version of Equation 2.7.
\[ V(y_i \mid u_i) = \tilde{\gamma}_0 + u_i \gamma. \]  \hfill (4.3)

We set \( \tilde{\gamma}_0 \) to equal 1.04s_i^2; generated the \( u_i \) as i.i.d. from a Uniform (0, 1) distribution, and set \( \gamma = 1.28 \). Thus, we would expect that \( V(y_i \mid u_i) = 1.04s_i^2 + 0.64 \). This means that, in the equal-wts condition, 38% of the variance in \( y \) is attributable to the important SLC; in the skew-wts condition, this should hold on average. In addition to the single important \( u \), we also simulated a single noise SLC from the uniform distribution.

We fit linear heteroscedasticity models using simple regressions and using the Lasso. For both methods, we first calculated the vector of squared residuals. Next, to fit simple regressions, we applied the GLS estimator of \( \gamma \) from Amemiya (1977). This is a two-step estimator: first, \( \gamma \) is estimated using OLS; second, \( \gamma \) is re-estimated using GLS with the diagonal weighting matrix \( u\hat{\gamma}_{OLS}I_{N \times N} \), where \( \hat{\gamma}_{OLS} \) is the estimator from the first step. To fit Lasso models, we replaced the OLS and GLS estimates with estimates under Lasso shrinkage.

4.2.5 Expectations before simulating

The most important factor in this set of simulations is the method factor. However, it seems extremely unlikely that any method will categorically outperform the others. Based on pilot simulations, we expect that the number of SLCs \( p \) factor will be associated with the greatest differences in performance among the methods, with simple regression doing best for large \( p \). Based on the results in Chapter 3, we anticipated that bootstrapped Lasso would perform as well as simple regression in this condition. The pilot simulations also suggested that the Lasso methods would outperform simple regression in the skew-wts condition.
In our pilots, the number of studies ($N$) and strength of association ($r^2$) factors did not seem crucial to distinguishing the methods; rather, varying them was useful to identify scenarios where none of the SLC selection methods was likely to be informative. We used the information in Tables 4.1 and 4.2 to estimate the expected VSP for simple meta-regression at each combination of $N$, $p$, and $r^2$, which is presented in Table 4.3. The table suggests that in the $N = 6$, $p = 20$ cell, poor performance should be expected from every approach.

**TABLE 4.3**

SIMPLE META-REGRESSION: EXPECTED VSP BY $N$, $p$, AND $r^2$

<table>
<thead>
<tr>
<th></th>
<th>$r^2 = 0.2$</th>
<th></th>
<th>$r^2 = 0.5$</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$N \backslash p$</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td>5</td>
<td>10</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----</td>
<td>-----</td>
<td>-------------</td>
<td>------</td>
<td>-----</td>
<td>-----</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.18</td>
<td>0.10</td>
<td>0.05</td>
<td>0.40</td>
<td>0.25</td>
<td>0.14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>0.32</td>
<td>0.19</td>
<td>0.11</td>
<td>0.69</td>
<td>0.53</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>0.55</td>
<td>0.38</td>
<td>0.23</td>
<td>0.80</td>
<td>0.66</td>
<td>0.49</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>0.67</td>
<td>0.51</td>
<td>0.34</td>
<td>0.80</td>
<td>0.67</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2.5.1 Applying SCAD and bootstrap CIs

We did not use the SCAD or bootstrap CIs in each cell of the design. The SCAD is expected to perform best with moderate numbers of SLCs and studies. The use of the SCAD was limited to cells with sample sizes of 18 or larger and having 10 or more SLCs. Additionally, the SCAD was not applied to the linear heteroscedasticity model.

Bootstrap CIs are expected to be too narrow with very small sample sizes, because ties are expected to be common in bootstrap samples [Fisher and Hall, 1991]. When the number of SLCs is large, calculation of bootstrap CIs can require substantial computing time. So, bootstrap CIs were limited to cells with sample sizes of 18 or larger and having 10 or fewer SLCs. Table 4.4 shows the overall design of simulations and identifies the cells in which the SCAD or bootstrap CIs were used. Bootstrap CIs were not applied to the linear heteroscedasticity model.

4.3 Implementation details

The simulations were implemented in R [Ihaka and Gentleman, 1996; R Core Team, 2013]. We used the Lasso functions in the package glmnet, the SCAD functions in the package SIS, and bootstrapping wrapper functions from the packages boot and cvTools [Friedman et al., 2010a; Fan et al., 2010; Canty and Ripley, 2014; Alfons, 2012]. Scripts are available from the author.

The number of MC replications and bootstrap samples used were the same as those in Chapter 3: 1500 MC replications per cell in the design, 1000 bootstrap samples drawn when needed. Using R version 2.15.1 on Linux, data generation and model fitting for each cell required approximately 400 Mb of RAM, 5 Mb of HDD space, and 45 minutes of processor time on a single AMD Opteron 2.3 GHz processor.
TABLE 4.4

SIMULATION DESIGN SHOWING WHICH CELLS USED SCAD OR BOOTSTRAPPING

<table>
<thead>
<tr>
<th>N\p</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>18</td>
<td>,B</td>
<td>,B,S</td>
<td>,S</td>
</tr>
<tr>
<td>54</td>
<td>,B</td>
<td>,B,S</td>
<td>,S</td>
</tr>
<tr>
<td>100</td>
<td>,B</td>
<td>,B,S</td>
<td>,S</td>
</tr>
</tbody>
</table>

“.” indicates that data were simulated under all conditions of the factors $r^2$, $wts$, and linear heteroscedasticity model; B = bootstrap CIs used; S = SCAD used
4.4 Results

The results of our simulations suggest that the bootstrapped methods performed the best across all conditions, and that: 1) the predictions in Table 4.3 for the most favorable conditions were supported; and 2) the simulation design made it so methods had higher VSP when fixed-effects models were fit to random-effects data.

For each of the 504 cells in the design, the mean and standard deviation over the 1500 MC replications were calculated. The detailed results presented below are based on comparing the MC mean values from each cell. Boxplots are used to summarize trends in mean VSP value between cells in the experimental design. For example, Figure 4.3 contains five boxplots. Each one is used to summarize the MC mean values of VSP for all the cells at a certain level of the method factor. The red boxplot, on the far left, summarizes VSP values for the 72 cells in which the bootstrap Lasso was used. The median of the mean VSP over MC replications is 0.522, the maximum is 0.9938, and the minimum, 0.0033. So, in 36 cells, the bootstrapped Lasso had a MC mean VSP under 0.522. In the most favorable cell, the bootstrapped Lasso had an average VSP of 0.9938, while in the worst, the bootstrapped Lasso had an average VSP of 0.0033.

4.4.1 Main effect of the method factor

The boxplots in Figure 4.2 show that the methods clearly differed in VSP, but that no method dominated the others across all conditions. Bootstrapped simple regression had the highest median performance, but there were cells in which the bootstrapped Lasso had better VSP. The SCAD and simple regression had about the same pattern of VSP values, and the Lasso was, overall, the poorest method. The interquartile ranges of VSPs overlapped for all methods at each sample size, so identifying additional differences in performance is difficult. Consequently, we present
results for the methods given other factors; by conditioning on them, we may observe clear differences in performance between the methods.

4.4.2 Factors that influence the probability of correct SLC selection: Sample size and association strength

As sample size increased, every method’s VSP increased, as expected. Figure 4.3 shows boxplots of VSP values for each method at each value of $N$. At large $N$, the bootstrapped methods had the best VSP values. At smaller $N$, simple regression showed the best performance. The Lasso showed poor performance at all sample sizes. Increasing sample size had the greatest effect on the bootstrapped Lasso’s performance: it had the lowest median VSP at $N = 18$, but the highest at $N = 100$. However, it also had the cell with the lowest VSP at $N = 100$.

As $r^2$ increased, every method’s VSP increased, as expected. In general, the results for the association strength factor are consistent with those of the sample size factor. This is sensible because both factors are positively associated with the power to correctly identify an important predictor. However, at sample sizes under 20 or with $r^2 = 0.2$, with any method, the median VSP was under 0.5. Under these conditions, accurate selection in a single sample is unlikely. Results for the $r^2$ are plotted in Figure A.1 located in the Appendix.
Figure 4.2. Boxplots of VSP values for each method suggest that, overall, the Lasso has poor performance, bootstrapped simple regression good performance, and bootstrapped Lasso highly variable performance.
Figure 4.3. A boxplot of VSP values at each sample size suggests that bootstrapped methods perform best at large sample sizes and that simple regression performs best at small sample sizes.
4.4.3 Manipulating false positive selections: The number of SLCs factor

The results for the number of SLCs ($p$) factor also match expectations, with increasing $p$ associated with decreased VSP for all methods. There was substantial overlap in VSPs across methods, as illustrated in Figure 4.4. However, in the best-performing cells for the bootstrapped Lasso and the SCAD, the decrease in VSP was small. The bootstrapped simple regression method showed the highest median VSP, and the Lasso the lowest.
Figure 4.4. A boxplot of VSP values at each number of SLCs ($p$) shows that increasing $p$ degrades VSP for all methods, but suggests that bootstrapped Lasso regression is the least affected.
4.4.4 Study weights

Having skewed study weights degraded the performance of all methods. It was the factor having the largest effect on VSP; additional analyses suggested that having skewed weights both increased the number of false positive selections and decreased the probability of a true positive selection, regardless of method. In the skew-wts condition, bootstrapped simple regression showed the best performance. The Lasso again performed poorly overall, with a median VSP of 0.229 for the default Lasso and 0.282 for the bootstrapped Lasso.
Figure 4.5. Skewed study weights decrease the VSP of all selection methods. The decrease in performance is greater than the decrease resulting from considering a large number of SLCs.
4.4.5 Linear heteroscedasticity model

4.4.5.1 Random effects and study weights

Figure 4.6 shows that, in data generated using random-effects models, all methods have generally increased VSP over data generated using fixed-effects models. This may be due to the upward bias suggested by Equation 4.2.4.2; if so, this bias is apparently high enough to increase the chances of selecting a true positive, but not so high as to increase the false positive selection rate. The linear heteroscedasticity model provided even more favorable conditions for variable selection; this likely is a result of the generation of the $U$ variables, and warrants further investigation.

4.4.5.2 Selecting random effects

In contrast to the results in selecting SLCs that predict $\mathcal{E}(y_i)$, applying the Lasso to select SLCs that predict $\mathcal{V}(y_i)$ resulted in higher median VSP than did simple regression. However, as shown in Figure 4.7, there is substantial overlap between the ranges of VSPs. Additionally, the VSP values of both methods are quite low; even in the best-performing cells, the VSP for the important predictor was under 0.5.
Figure 4.6. For all methods, VSP was much larger for data generated under random effects, rather than fixed-effects, models.
Figure 4.7. The Lasso and simple meta-regression both have relatively low VSP for selecting an important predictor of random-effects variance. The random-effects variance in this example represents 40% of the total variance in effect sizes across studies.
4.4.6 Results conditional on several factors

The random-effects models had the most favorable conditions for all methods, so we focused attention there. A set of scatterplots of each cell’s mean VSP over MC replications is shown in Figure A.4. It suggests that, with low-to-moderate power to select an important predictor, bootstrapped simple regression performed the best. With more power, the bootstrapped Lasso equaled or outperforms bootstrapped simple regression. The SCAD performed at its best in the same cells in which the bootstrapped Lasso did well.

In conditions with very small sample sizes, all methods perform equally poorly.

Table 4.5 lists the results in the best cell of the design for each method. It is clear that larger samples, fewer SLCs, random effects, and equal weighting led to the best performance. The fact that the best performing Lasso and bootstrapped simple regressions were in the association strength $r^2 = 0.2$ condition may be evidence for a ceiling effect, especially for bootstrapped simple regression.

TABLE 4.5

<table>
<thead>
<tr>
<th>Method</th>
<th>VSP</th>
<th>N</th>
<th>p</th>
<th>$r^2$</th>
<th>Model</th>
<th>wts</th>
</tr>
</thead>
<tbody>
<tr>
<td>bootLasso</td>
<td>0.994(7e-04)</td>
<td>100</td>
<td>5</td>
<td>0.5</td>
<td>LinH</td>
<td>equal</td>
</tr>
<tr>
<td>bootSimp</td>
<td>0.946(.0038)</td>
<td>100</td>
<td>5</td>
<td>0.2</td>
<td>LinH</td>
<td>equal</td>
</tr>
<tr>
<td>Lasso</td>
<td>0.678(.0075)</td>
<td>54</td>
<td>5</td>
<td>0.2</td>
<td>LinH</td>
<td>equal</td>
</tr>
<tr>
<td>SCAD</td>
<td>0.966(.0030)</td>
<td>100</td>
<td>10</td>
<td>0.5</td>
<td>Random</td>
<td>equal</td>
</tr>
<tr>
<td>Simple</td>
<td>0.962(.0030)</td>
<td>18</td>
<td>5</td>
<td>0.5</td>
<td>LinH</td>
<td>equal</td>
</tr>
</tbody>
</table>
Figure 4.8. Boxplots of VSP values in the random-effects conditions, showing the effects of association strength and sample size. In larger data sets and/or stronger association, the bootstrapped Lasso performs well; with moderate power, bootstrapped simple regression performs well.

Identical data as in Figure A.4
4.5 Discussion

The simulation results suggest that bootstrap CIs are necessary for successful application of the Lasso as a meta-regression method. The bootstrapped Lasso performed well enough to warrant applying the method to an empirical analysis. Bootstrapped simple regression showed about the same overall VSP as the bootstrapped Lasso, and slightly outperformed it in the presence of skewed weights. However, in this simulation, SLCs were simulated independently of one another. In an empirical application of meta-regression, where SLCs, the Lasso’s ability to estimate conditional associations should result in it outperforming bootstrapped simple regression. The SCAD required larger samples than the other methods to perform well. However, at larger sample sizes, it outperformed the Lasso, which suggests that bootstrapping the SCAD is a promising future direction for meta-regressions that are based on large numbers of studies.

An important practical result was that very small sample sizes and skewed weights are likely to lead to imprecise selection of VSPs regardless of method used. We did not bootstrap small sample sizes in this simulation, but, when applying the Lasso to small samples, we will use the ‘exact’ bootstrap distribution.

We observed the counter-intuitive result that applying fixed-effects models to random-effects data led to better performance. This was likely as a result of bias, which, due to the extreme simplicity of the simulations, increased VSP. In empirical data sets, this seems unlikely.

Applying the linear heteroscedasticity model was less successful; in the simulation results, VSP less than 0.2 was typical. The next step in this line of research is to find a $\gamma$ value that is large enough to be precisely detected by using the Lasso. Alternatively, more powerful random-effects models can be investigated.

The next chapter concerns the application of the (bootstrapped) Lasso and simple meta-regression to published meta-analysis data.
CHAPTER 5

EMPIRICAL ILLUSTRATION: COMPARING META-REGRESSION METHODS IN GENETIC NEUROSCIENCE DATA

The simulation results of the previous chapter suggest that variable selection precision (VSP) in meta-regression is strongly affected by conditions in data sets such as the presence of many SLCs, a skewed distribution of standard errors of the studies in the meta-analysis, and weak associations between SLCs and studies’ effect sizes. In this chapter, we applied the Lasso and simple meta-regression to a recent meta-analysis from the field of genetic neuroscience. The data set used in the meta-analysis has many of the conditions that would be expected to lead to poor VSP. However, we expected that other characteristics of this data set, such as correlations among the SLCs, might lead the Lasso to have better performance.

5.1 In a meta-analysis of neuroscience results, shrinkage meta-regression suggested SLCs that simple regression did not

5.1.1 Background

[Murphy et al.] (2013) performed a meta-analysis of 34 studies which combined neuroimaging and psychiatric genetics. In each of these studies, participants were genotyped in order to identify the presence or absence of a mutation in their serotonin transporter gene. They then viewed an emotionally arousing stimulus while undergoing neuroimaging. The effect tested in each study was the association between the mutation and activation of the amygdala, as measured by the neuroimaging
technique. The strength of this association was then expressed using Hedges $g$; in the meta-regression, $y_i$ represents the $g$ value from study $i$.

In order to gather such a large number of genetic neuroscience studies, Murphy et al. (2013) included studies that used a variety of methodologies. Because of this, Murphy et al. (2013) published a large table (Table 1, p. 515) listing the studies, their effect sizes, and 17 study-level covariates (SLCs). For example, results were aggregated over studies based on clinical as well as on population samples. Studies representing three types of stimuli (words, faces, and pictures) and four types of neuroimaging (PET, fMRI, perfusion MR, and SPECT) were analyzed. In total, four continuous and 13 categorical variables related to study methodology were presented; together, they require 34 dummy variables to represent. Hence, there are 38 SLCs that could be entered into a meta-regression; $p = 38$, while the sample size, $N$, is only 34.

The authors of the paper deal with the problem of having more SLCs than studies by performing simple meta-regressions using a random-effects model. Five unpublished studies were gathered as part of the analysis; the authors found a significant difference in effect sizes between the published and the unpublished studies. The authors performed meta-regressions in both the full set of 34 studies and in the set of 29 published studies. Other than publication status, no SLCs were significantly associated with effect size in the full set of studies. Considering only published studies, the authors identified two SLCs that were significantly associated with effect size: 1) average age of the sample was positively related to effect size (significant with p-value 0.01); 2) whether or not the sample’s genotypes were in Hardy-Weinberg Equilibrium, which is a measure of genotyping quality (significant with p-value 0.04) (Turner et al., 2011). The measure of genotyping quality was negatively associated with effect size; lower quality studies tended to have larger effects. Substantial heterogeneity in effect sizes remained after performing these meta-regression, which is consistent with the
presence of random effects.

We applied the Lasso and simple regression to this data. Bootstrap confidence intervals were used for both methods. We used the same R packages as were described in Chapter 4 to perform these analyses.

5.1.2 Method

The simulations presented in Chapter 4 suggest that the number of SLCs, the skewness of the study weights, and the strength of the association between SLCs and effect size are important factors in making precise selections of SLCs. This is an empirical study; the important SLCs (if any) are unknown. However, noting the number of SLCs, the skewness of the weights, and the strength of association helped to put our results in context.

The main part of our analysis was to fit simple linear regression and the Lasso to the full set and to the published subset of study data presented in Murphy et al. (2013). We generated bootstrap CIs for both methods; the results of the simulations suggest that this will lead to more precise results. For the Lasso, we nested selection of $\lambda$ within bootstrapping; we used 4-fold cross-validation to perform the selection.

5.1.3 Results

5.1.3.1 The data set

There are 38 SLCs in this data set, one of which is publication status. This suggests that false positive selections may lead to low VSP.

We calculated study weights according to Equation 4.2.2.2 and the formula for $g$ (Hedges, 1983). The study weights are skewed; in the full data set, their skewness is 2.17, and in the published subset, the skewness is 2.09.

To assess the strength of the relationship between continuous SLCs and effect size, we calculated the squared correlation between each continuous SLC and effect
size, after weighting each study. To assess the strength of the relationship between categorical SLCs and effect size, we use the $R^2$ resulting from the weighted regression of effect size on all of the SLC’s dummy variables. Table 5.1 gives the five-number summary of $r^2$ and $R^2$ in the full data and the published subset. Associations are stronger in the published subset, as reported in the paper. The largest associations in both sets are $r^2 \approx 0.1 - 0.2$. The $r^2 = 0.2$ condition led to moderately poor performance in the simulations of Chapter 4.

If the simulation results generalize to this data, the majority of SLCs selected by any method are likely to be false positives, though the bootstrapped methods should make more precise selections than the other methods. However, making a SLC selection is useful in this field because it generates a hypothesis that can be tested in independent data.

### TABLE 5.1

SQUARED CORRELATIONS BETWEEN SLCs AND EFFECT SIZES
SUGGEST SMALL-TO-MODERATE ASSOCIATIONS

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Full data $r^2$ Value</th>
<th>Variable</th>
<th>Published subset $r^2$ Value</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>0.003</td>
<td>Stimulus Task</td>
<td>0.013</td>
<td>Stimulus Presentation</td>
</tr>
<tr>
<td>1st Quartile</td>
<td>0.022</td>
<td>-</td>
<td>0.023</td>
<td>-</td>
</tr>
<tr>
<td>Median</td>
<td>0.033</td>
<td>-</td>
<td>0.039</td>
<td>Imaging Method</td>
</tr>
<tr>
<td>3rd Quartile</td>
<td>0.075</td>
<td>-</td>
<td>0.078</td>
<td>-</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.22</td>
<td>Publication Status</td>
<td>0.101</td>
<td>Stimulus Type</td>
</tr>
</tbody>
</table>
5.1.3.2 Simple regression

The results of applying simple regression were close to the results in Murphy et al. (2013), as expected. In the full sample, publication status was selected as an important SLC by simple regression. Simple regression selected no SLCs in the published subset of studies.

We used a slightly different test in our analyses than did Murphy et al. (2013), hence there were small discrepancies. The Hardy-Weinberg Equilibrium SLC was marginally significant based on the 2-tailed test used in simple regression, so was not selected. This SLC was significant in the homogeneity of treatment test used in the published meta-analysis, which is described in DerSimonian and Laird (1986).

We obtained different results when using bootstrapped simple regression. In the full data set, publication status, average age of sample, and Hardy-Weinberg Equilibrium were selected as important SLCs, as were the manner in which the stimulus was presented, the imaging method, and the brain hemisphere that was imaged. In the published subset, Hardy-Weinberg Equilibrium and brain hemisphere were again selected.

5.1.3.3 The Lasso

The Lasso results differed from those of simple regression. When applying the Lasso without bootstrapping, the ancestry of the sample was selected as an important SLC both in the full data set and in the published subset. The bootstrapped Lasso did not identify any important predictors. Interestingly, the Lasso methods did not select publication status or Hardy-Weinberg Equilibrium as important variables.

5.1.4 Discussion

In our re-analysis of the meta-analytic data published by Murphy et al. (2013), simple regression with bootstrapping identified all of the SLCs that the authors sug-
gested were important. However, applying the Lasso suggested that a novel SLC, the ancestry of the individuals in a study. In the published data, studies that had samples of European ancestry had a mean effect size of 0.48, with a standard error of 0.1, based on 21 studies. Samples with mixed or non-European ancestry had a mean effect size of 0.144, with SE 0.3, based on 7 studies. Results in the full data set were similar. The difference in effect sizes between studies having different ancestries was not significant. However, ancestry is moderately associated with other SLCs, most strongly with the stimulus contrast, Kendall’s rank-correlation coefficient 0.36, p-value 0.04. It may be that the ability of the Lasso to estimate conditional relationships has been able to identify an important SLC, ancestry of the sample, that simple regression and similar methods overlook.

An important limitation in this illustration is that the glmnet package requires complete data in order to fit Lasso models. As a result, Lasso model fitting was done using a reduced sample of $N = 19$ studies.
CHAPTER 6

DISCUSSION: REGULARIZED META-REGRESSION SHOULD BE
BOOTSTRAPPED TO BE EFFECTIVE

Meta-regression is used to identify study-level covariates (SLCs) that are associated with heterogeneity of effect sizes between the studies in a meta-analysis. In newer fields, such as psychiatric genomics, the pool of SLCs that may be used for meta-regression is large, approaching or exceeding the number of studies that can be analyzed. In this situation, the choice of which SLCs to include in a meta-regression is an open question that can be addressed using automatic methods for variable selection.

In this text, we have introduced three methods for automatic variable selection: 1) simple regression of effect size on each SLC separately, with selection based on hypothesis testing; 2) the Lasso, a shrinkage regression method; and 3) the SCAD, which is similar to the Lasso but provides unbiased estimates of large regression coefficients.

By comparing these methods using simulated and empirical data, we’ve observed two main results. For meta-regression in which sample sizes are relatively large ($N > 50$ studies), if the studies in the meta-analysis estimate their effects sizes with roughly equal precision, then a bootstrapped version of the Lasso is an appropriate method for variable selection. Under these conditions, more than 75% of the SLCs identified using this method were truly associated with heterogeneity of effect sizes. In situations where sample size or the association between SLCs and studies’ effect sizes were more moderate, bootstrapped simple regression was shown to be an effective
method. In an exploratory meta-regression, it might be quite fruitful to use both methods, as illustrated by applying both to data from a meta-analysis of genetic neuroscience data.

The first workers to apply the Lasso to meta-analysis appear to have been (Vignes et al., 2011); they applied the method to the analysis of gene regulatory networks. The present context is simpler, but our results are informative about the performance of Lasso methods at small sample sizes, which were the source of several limitations to the study. The first was that a “general” approach to bootstrapping and cross-validation was used. It may have been more profitable to use approaches which have been designed to yield more accurate results when applied to small samples, such as exact bootstrap calculations, the balanced bootstrap, and important resampling. A second limitation was that stepwise regression methods were not used in the comparison. The positive results for simple regression (the first step in many stepwise procedures) suggest that it may have been useful in this context. However, the Lasso is related both to stepwise regression and to best subsets regression, and the methods perform similarly in some data sets (Efron et al., 2004; Tibshirani, 1996). Stepwise methods could show performance that is intermediate between that of the Lasso and of simple regressions. A final limitation was that correlated SLCs were not considered. Because the Lasso uses conditional associations in generating its solutions, it might perform much more favorably. The mediocre performance of the Lasso, even at large sample sizes and strong SLC-study effects associations, was an unexpected result. That the SCAD performed substantially better than the Lasso was also unexpected; the first follow-ups to the present analysis will focus on bootstrapping the SCAD as well as on applying other shrinkage regression methods. A second source of follow-up studies involves random effects: 1) finding appropriate values for the regression coefficients in the linear heteroscedasticity model; 2) considering more powerful approaches to estimating random-effects variance (López-López et al., 2014); and 3)
studying the relationship between the size of random-effects variance and expected VSP.

This study is useful to researchers in psychiatric genomics because it suggests methods that can be used to precisely identify important predictors of the heterogeneity of effect sizes across studies. The field relies on large consortia, comprising many studies, to make discoveries. It is nearly inevitable that new consortia, composed of new cohorts, will be founded. When planning the creation of such a consortium, the methods presented here might be used to learn from existing consortia and their results, for example, whether it is beneficial to use a uniform imaging technique, or if later-life studies should be included.
APPENDIX A

ADDITIONAL FIGURES
Figure A.1. A boxplot of VSP values at each level of association strength \( (r^2) \) suggests that bootstrapped methods perform best across \( r^2 \) values.
Figure A.2. Boxplots of VSP values in the fixed-effects condition, showing the effects of association strength and sample size. Identical data as in Figure A.3.
Figure A.3. Scatterplots of VSP values in the fixed-effects condition, showing the effects of association strength and sample size. Identical data as in Figure A.2.
Figure A.4. Boxplots of VSP values in the random-effects conditions, showing the effects of association strength and sample size. Identical data as in Figure 4.8.
REFERENCES


