12-2018

Matching potential in randomized complete block designs

Elaine Sotherden
Clemson University, elainesotherden@gmail.com

Matching potential in randomized complete block designs

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Science

---

by
Elaine Sotherden
December 2018

Accepted by:
Dr. William C. Bridges, Jr., Committee Chair
Dr. Patrick Gerard
Dr. Brook Russell
Dr. Matthew Saltzman

Abstract

A small sheep experiment (nobs=32) planned to use a randomized complete block design (RCBD) treatment assignment of two binary factors. Complications creating the RCBD blocks prompted the researchers to discard the original blocks from the initial analysis plan and to rearrange their experimental units into new groups using linear covariate adjustment. We compare the blocks from the experiment's initial analysis plan and the groups from the researcher's linear covariate adjustment to groups formed by potential matching methods. We evaluate these three analysis approaches on the original sheep dataset and on simulated sheep datasets. We find that the groups created using matching methods produce less precise estimates and that further, those estimates may be biased. Additionally, the matching methods may alter the experiment's size and thus, its overall power. When small RCBD experiments have complications forming the desired blocks, we recommend the joint use of well-established preliminary testing and post-stratification procedures. This acts as a more formalized version of the sheep researchers' use of linear covariate adjustment and implicit model selection.

Dedication

This work is dedicated to God, for his honor and glory, who knew its contents before it was conceived and who granted me the strength to persevere.

Acknowledgments

My most sincere praise and gratitude to my advisor, Dr. William Bridges, without whose extensive knowledge and experience with statistics, guidance throughout the years, good advice, excellent editing, and valuable investment of time, this process could never be completed.

To my committee, Dr. Patrick Gerard, Dr. Brook Russell, and Dr. Matthew Saltzman, my great appreciativeness for your interest in this dissertation and for your patience and direction in helping me to complete it. To fellow colleague and friend, Garrett Dranichak, my thanks for your clear explanations of networks and assignment algorithms in the foreign realm of operations research. In the department, I would also like to thank Dr. Judith McKnew and Dr. Christy Brown, for their training, mentorship, and encouragement for instructing undergraduates. Additionally, to the department staff over the years, especially Kris Hunnicutt and Carol Lund, my acknowledgement of both your friendliness and helpfulness navigating graduate student life.

I am personally indebted to my parents, James and Joan Sotherden, for their unconditional support and for giving me all the tools needed to foster my love of learning. I am especially grateful to my brother, Edward, for keeping me grounded in life outside academics, to his family, for their constant support, and to my friends Elizabeth Ryder and Allison Heck, for always considering success as a given. I am so grateful to the many family members who provided encouragement throughout the years and to the many, many others who prayed earnestly for me. Thank you especially to my friends, Rachel Smith and Tony Nguyen without were unfailingly patient and encouraging in the last part of this process and to Rachel especially, who let me stay with her through unexpected developments.

Finally, a special acknowledgement of my uncle, Edward Carey, whose own education set my on this path. His example introduced me to the concept of doctorates and made me sure I would pursue one before I began kindergarten.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

To investigate the effect of endophyte (*Neotyphodium coenophialum*) infected fescue on sheep reproduction, a sheep study (Suffolk ewes) was conducted during 2015–16 at Clemson University. The study sought to confirm and quantify the detrimental effect of consuming this infected grass species on sheep's fetal development, a detriment supported by extensive anecdotal evidence. Reasonably equivalent ewes (nobs=40) were sorted into groups of equal size (nblocks=10) and impregnated by a single ram. After gestation, several weights recorded on each pregnant ewe measured fetal development.

To form blocks for the planned analysis, the researchers used physical impregnation groups, stratifying ewes on their initial body condition score (BCS), a livestock measurement in which healthier ewes tend to score more highly. After confirming pregnancy, the researchers randomly assigned pregnant ewes to one of four treatments within each block, resulting in an experiment with a randomized complete block design. The experiment used a two by two factorial treatment structure consisting of regulated seed diets applying the toxic fescue treatment (i.e., the seed either did or did not contain the toxic endophyte) at two times (i.e., the pregnancy's second and third trimesters). Further details of the study protocol can be found in [8].

The sheep researchers considered a variety of variables that may respond to the treatments, but the basic analysis plan was identical for each of the response variables. Initially, the researchers used ordinary least squares (OLS) to estimate the model parameters and the response variable means for each treatment. They also conducted an analysis of variance (ANOVA) to evaluate hypothesis tests about the overall effects of the two treatment factors on the response variable

means. Unfortunately, unforeseen complications in the study's execution caused challenges for the basic analysis plan.

The first complication in the study's execution concerned the impregnation of the ewes. Impregnation occurred based on the sets of ewes comprising the blocks. Specifically, the first set of impregnations was made in the first block (i.e., the block containing the ewes with the five highest initial BCS values), the second set of impregnations was made on the second block (the block containing the ewes with the next five highest initial BCS values, allowing these later ewes to increase their weight and BCS before attempting impregnation), and so forth until the tenth and final block. Then, after a (suspected) pregnancy, the four treatments were (randomly) assigned to the ewes within each block, in such a way that a block contained at least one ewe of each treatment whenever possible.

The major complication occurred when impregnation was not successful; ewes assigned to receive the treatment involving no infected seed during the second trimester and that were also later found to be not pregnant were "recycled" for a subsequent impregnation attempt in another block. Additionally, some ewes were removed from the study due to complications during their pregnancy. Then, as some of the ten blocks were incomplete at the analysis time, experimental units assigned to the some blocks were redistributed to complete some of the incomplete blocks, or to "fill in the gaps" (see Table 1.1). After rearranging, the dataset had the same structure as a randomized complete block design (RCBD) in which (nobs=32) were sorted into groups of size four (nblocks=8).

The initial study protocol grouped the ewes into complete blocks defined by initial BCS, but after sheep recycling (and other complications), the initial BCS values were mixed within the blocks, so the "block effect" was not a simple function of the initial BCS. The block effect was "not significant" in the ANOVA for many of the response variables, and the response variable ANOVAs suggested that the endophyte infected fescue diet was significant in the third trimester but not significant in the second trimester. Based on these issues, the researchers decided to remove the estimation of the block effect from their final analysis and to disregard the second trimester effect and its interaction with the third trimester. Also, as imcomplete blocks were completed ad-hoc from a supply of "extra" experimental units, the study took on characteristics of an observational study as opposed to a randomized experiment. This prompted our consideration of matching methods.

The second complication in the study's execution involved important covariates measured after the treatment assignment ("post-hoc"), lamb number and genotype, that the study protocol

Table 1.1: Original block assignments

Original block assignments for each ewe were changed after the experiment to form a total of eight complete blocks for statistical analysis. Five ewes were removed during the study, and four ewes were redistributed to different blocks. That is,

- ewe numbers *37, 16, 25, 20*, and *48* were *removed* altogether, and
- several ewes were **redistributed** to form complete blocks. Specifically,
  - ewe number **44** was **redistributed** from block 2 to block 7;
  - ewe number **36** was **redistributed** from block 9 to group 3; block ewe number **49** was **redistributed** from block 7 to block 4; and
  - ewe number **2** was **redistributed** from block 10 to block 6.

Original Blocks	Ewe ID
1	4
1	18
1	38
1	50
2	15
2	31
2	32
2	**44**
2	45
3	22
3	*25*
3	46
4	14
4	34
4	*37*
4	52
5	11
5	12
5	47
5	57
6	7
6	*16*
6	24
6	39
6	42
7	41
7	43
7	48
7	**49**
7	55
8	9
8	51
8	56
8	58
9	**36**
9	*20*
10	**2**

Updated Blocks	Ewe ID
1	4
1	18
1	38
1	50
2	15
2	31
2	32
2	45
3	22
3	25
3	**36**
3	46
4	14
4	34
4	**49**
4	52
5	11
5	12
5	47
5	57
6	**2**
6	7
6	39
6	42
7	41
7	43
7	**44**
7	55
8	9
8	51
8	56
8	58

failed to anticipate. Lamb number is the number of fetuses gestated by each ewe during the experiment. While strongly related to many of the response variable considered, it is not detectable in early sheep pregnancy and is thus unavailable until after treatment assignment. Genotype refers to the value of a genetic marker that has been found in recent literature to be associated with resistance to the detrimental effects of endophyte infected fescue; each ewe's genotype was also recorded after treatment assignment. The initial block definitions were based on BCS only, as no information about lamb number or genotype was initially available for use. Thus, the researchers decided to add estimates of the effects these important covariates to their analysis.

Altering the intended blocks and measuring important covariates after randomization caused the researcher's final analysis method to differ from their original specified protocol. For a new analysis strategy that could potentially better address these complications, the researchers chose covariate adjustment with the added covariates of lamb number and genotype directly forming a new set of blocks. We consider the idea of creating a new set of blocks more generally, still replacing the original blocks with groups based on the post-hoc covariates, but creating the new groups with matching methods originally developed for causal inference in observational studies. Note that we use "blocks" to refer solely to the researchers' originally created groups, and "groups" to refer to all other new sets of blocks, both those created implicitly through covariate adjustment and those created explicitly with matching methods.

Overall, our objective is to compare three analysis approaches, the original planned approach of data analysis (denoted "protocol"), the modified approach of data analysis the researchers chose (denoted "researcher"), and some new analysis approaches based on post-hoc matching (denoted "potential"). The analysis approaches are evaluated for several of the response variables to determine whether or not the researcher or potential methods offer any important advantages for addressing a study's complications. *We compare the three methods on the original data set and with a small simulation study.* Criteria to determine possible method advantages include (1) the estimate of the effect of the endophyte infected fescue diet on the response variable means and (2) the standard error of the effect of the endophyte infected fescue diet (i.e., the standard deviation of the difference in response variable means).

Chapter 2

Background

Including covariate information when estimating differences among true population means is foundational to data analysis. When planning an analysis, we can use covariate information before treatment assignment ("pre-experimental"), usually as some type of blocking experiment design with the blocks included in the analysis of variance (ANOVA), or after treatment assignment ("post-experimental" or "post-hoc"), usually as some type of analysis of covariance (ANCOVA). The former case allows the number of observations within each covariate category, or block, to be numerically balanced. In the latter case, as treatment levels contain varying proportions of treated units across covariate categories, the number of observations within each category can be unbalanced. In both cases, when we block or stratify on a single covariate in a linear model (i.e., linear covariate adjustment), we are effectively accounting for variation in the measured response that can be (directly) explained by the given covariate.

Often more than one covariate is available for inclusion in a linear model, but including too many covariates can leave fewer degrees of freedom and can insinuate false connections between covariates and a measured response. Additionally, using too many covariates increases our observations' "sparsity" relative to the range of possible covariate combinations. Researchers thus aim to use a minimal list of the covariates they believe most strongly explain variation in their measured response. (Direct) linear adjustment of more than one covariate leads to the researcher models, so for our potential models, matching methods provide alternate (indirect) ways to include multiple covariates in a linear model.

2.1 Researcher approach: Covariate adjustment

Mimicking the researchers, we ignore the original blocks and also ignore the effect of an infected seed diet in the second trimester and its interaction with an infected seed diet in the third trimester. The result is a completely unadjusted linear model. Then, we add lamb number, genotype, and initial BCS to the linear model as post-hoc covariates. Note that lamb number and genotype were determined by the researcher, bypassing a much larger question of variable selection.

Recall that lamb number is the number of fetuses each ewe gestates. To add each ewe's lamb number to the linear model, lamb number is a discrete integer covariate with three levels. Researchers know that lamb number is very strongly related to the considered response variables. In the sheep study, lamb number is a post-hoc covariate because it is not fully detectable in early sheep pregnancy and so is unavailable until after treatment assignment. Note, however, that lamb number should be unrelated to (is independent of) the assigned treatment through the randomized treatment assignment mechanism Although we cannot numerically balance lamb number across treatment levels, its value is fixed before treatment, so an ewe's lamb number is unlikely to be influenced by the actual treatment (or so the researchers believe).

Recall that genotype refers to the value of a genetic marker that has been found in recent livestock literature to be associated with resistance to the detrimental effects of endophyte infected fescue. To add each ewe's genotype to the linear model, genotype is a nominal covariate with three levels. Researchers believe that ewes of different genotypes should respond differently to the fescue toxicity during in fetal development. In the sheep study, genotype is a post-hoc covariate because physical constraints prevented the researchers from genotyping the ewes in advance of treatment assignment. Although the explicit study goal was to investigate and quantify the effects of fescue toxicity on sheep fetal development, the researchers were at least as interested, implicitly, in identifying covariate markers that indicate sheep more suited to resist fescue toxicity. That is, researchers included genotype in their linear model to investigate which genotype, if any, breeds more successfully in the presence of toxic fescue and because, in sheep science, there should be a definite genotype effect on sheeps' fetal development in the presence of endophyte infested fescue.

Recall that BCS is a livestock measurement in which healthier ewes tend to score more highly (i.e., higher values correspond to sheep more likely to have a healthy pregnancy) and that each ewe's initial BCS was the covariate indirectly used to create the experiment's original analysis of

variance blocks. To add each ewe's initial BCS to the linear model, initial BCS is an integer covariate with levels from one to four. We choose to investigate this covariate, although the researchers did not include it in their final model, because the researchers initially believed to be related to most of the response variables. Further, if BCS was not actually related to the response variables, it may more generally represent a cost associated with adjustment by a falsely significant factor.

2.2 Potential approach: Matching

We want to create "better" blocks of experimental units by using matching methods as an alternate way to include information from multiple covariates in a linear model. Generally, matching methods were designed to correct (observational studies') response bias by creating groups of experimental units with similar covariate distributions. The idea of matching, however, is extremely useful beyond response bias correction; matching can also improve estimates' precision. As improving estimates' precision is a main use of blocks in experiments, using matching to "create" groups has potential in a study design.

Although a single experimental unit cannot simultaneously both receive and not receive any given treatment, it is possible for a researcher to have experimental units receiving a treatment which are believed to be comparable to experimental units not receiving the treatment. The most basic expression of comparable experimental units is pairwise matching, or when a single experimental unit is paired with a single "twin," one receiving a certain treatment and the other not receiving it. Pairwise matching can be extended both to multiple treatment levels and to multiple "twins" for each treated experimental unit. There remains, of course, complex questions of how many and which such matches are permissible and of how to select among several potential matches for an individual experimental unit.

In [16], matching methods are determined by specifying combinations of distance, structure, and algorithm, where distance characterizes the distributional difference between matched populations, structure characterizes the allowable matches, and algorithm characterizes specific rules for matching. Common types of algorithms are greedy matching, which includes the use of nearest neighbor, expectation maximization, and genetic matching procedures, and optimized matching [26]. Optimized matching typically uses network flow algorithms to estimate a "best" (optimized) matching assignment, or an assignment that minimizes the total distance between experimental units across all groups. Common structures are intuitive integer generalizations of one-to-one, or pairwise, matching and may be done with or without allowing replacement of already-assigned observations [31]. The popular default for the optimized algorithm is special structure known as full matching, when the assignment algorithm determines the best ratio of treated to control observations made within each match [**?**]. Finally, common distances generally make two choices, whether to seek matches only for finite moments or for full distributional information and whether to con-

sider marginal or joint covariate information when matching using multiple covariates [31]. Potential distances are generated by a estimated propensity score and by a Gaussian kernel.

Historically, many matching methods were designed to reduce observational study bias via an estimate of a sampled population's propensity score. The propensity score, or the respective (conditional) probability of each experimental unit being assigned to a treatment level (assuming covariate values are known) is a distance that can be used for matching. Propensity score theory was developed by Rubin [43] and Rosenbaum [41]. The propensity score is also a summary of covariate information that can be included in general adjustment methods [17, 26], and it can be considered a "balancing score" because it creates covariate subsets that are equally distributed across treated and untreated populations [42]. In practice, propensity scores are employed in four main approaches of weighting, stratifying, matching, and adjusting [26]. The authors of [17, 26] catalogue propensity score use, which is often seen in medical research [45] and in social sciences [33].

Most recently, kernel-based matching approaches, or matching methods with kernel distances, have been developed. Reproducing kernel Hilbert spaces (RKHS) are widely used in practice, and a kernel function can capture a covariate's full distributional information while simultaneously acknowledging the relationship between multiple covariates [14, 40]. Kernel-based matching methods have been developed in the literature [18, 21, 35, 48], but they have received significantly less attention in practice. Kallus [30, 31] studied the intersection of optimization and statistics and showed that many popular matching algorithms seek to minimize (optimize) an implicit definition of covariate imbalance; this also works in reverse, leading to more useful definitions, or metrics, of imbalance. As an example, Kallus [6] demonstrates that forming subgroups via simultaneous optimization of multiple covariates' observed moments can be more useful than matching on a single covariate. This last observation suggests that future experiments should be blocked not on a single covariate or on a small covariate subset but on all available covariate information, both pre-experimental and post-hoc.

Chapter 3

Methods

We analyze the sheep dataset in a sequence of three main approaches, the *protocol*, *researcher*, and *potential* model structures, as explained in the introduction. Each approach generates several linear models using the original researcher blocks and the covariates of lamb number, genotype, and initial BCS in different ways. In this section, we develop notation for the *protocol*, *researcher*, and *potential* model structures (with additional details in Appendix A). We also explain the intention of each model, for which Appendix E provides a summary. Finally, we discuss the creation of datasets simulated from the original sheep dataset.

3.1 Models

The *protocol* approach generates a linear model specified in advance by the experiment's protocol. It consists of the experiment's two binary treatment factors, the presence or absence of endophyte infested fescue seed in ewes' second and third pregnancy trimesters; the interaction between these two binary treatment factors; and the experiment's blocks which were originally used to assign treatments. The researchers planned the experiment using this model but discarded this model upon consideration of their observed results.

The *researcher* approach generates several linear models from post-hoc covariate adjustment on lamb number, genotype, and initial BCS. From the available researcher models, we highlight the completely unadjusted model, which uses only the binary third trimester treatment indicator, and the final researcher model, which combines the binary third trimester treatment with the two post-hoc covariates of lamb number and genotype. The researchers formally analyzed their experiment under the final researcher model.

The *potential* approach generates linear models by combining the binary third trimester treatment with different ways of grouping a dataset's experimental units. From the available potential models, we highlight those created by three matching methods, three variations of a propensity-score based method and one kernel-based method. We analyzed the experiment under these models to investigate their potential to improve the estimate of the third trimester effect.

In each linear model, the error term ε represents experimental error, or variation in the response that the model does not explain. As each model is linear in its unknown coefficients, each assumes that this unexplained error is additive relative to other included treatment and covariate indicators. That is, each model assumes that the unexplained error not a function of other model terms. Further, traditional blocked ANOVA procedures, as in the protocol approach, carry the implicit assumption that the treatments and the blocks created for treatment assignment do not interact.

Note especially that from the protocol model, ignoring the experiment's original blocks combines the response variation associated with the blocks with the protocol model's error term. Ignoring the second trimester effect and its interaction with the third trimester also combines the response variation associated with those terms with the protocol model's error term. The end result is a completely unadjusted model whose error term encompasses all response variation not

associated with the third trimester effect. Similarly, any added post-hoc covariates removes the response variation associated with those terms from the completely unadjusted model's error term.

Over the three approaches, we develop notation for total of 22 linear models, name each model, and describe each model's intent; a summary of the last is available in Appendix E.

3.1.1 Protocol model

The protocol approach uses a linear model based on a two by two factorial treatment design and a randomized complete block experiment design. The corresponding `protocol` model is

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{t2} z_{i,t2} + \beta_{int} z_{i,int} + BLK_i + \varepsilon_i. \tag{3.1}$$

for $i = 1, 2, \ldots, n = 32$. Here, the ith observation represents one ewe, with a total of 32 observed ewes, and BLK_i is a series of indicator variables (see Appendix A) representing the "block effect," or the effect of the block to which the ith ewe is assigned. Further, the constant β_0 is a "baseline" mean of the measured response, or a mean defined without considering other possible sources of the variation in the measured response.

The terms $z_{i,t2}$ and $z_{i,t3}$ are binary indicator variables for the second and third trimesters of each ewe's pregnancy, respectively. That is, $z_{i,t2} = 1$ if the ewe ate a diet of endophyte infected fescue in its second trimester and $z_{i,t2} = 0$ if not; similarly, $z_{i,t3} = 1$ if the ewe ate the infected diet in its third trimester and $z_{i,t3} = 0$ if not. Thus, the coefficients β_{t2} and β_{t3} respectively represent the change in the mean of the measured response from including binary indicator variables $z_{i,t2}$ and $z_{i,t3}$. Based on these definitions of $z_{i,t2}$ and $z_{i,t3}$, the coefficients β_{t2} and β_{t3} can be considered the effects of an endophyte infected fescue diet in the second and third trimesters, respectively, and we call these the "second trimester" and "third trimester" effects. Further, the coefficient β_{int} represents the change in the mean of the measured response y_i for a one unit change in the product of $z_{i,t2}$ and $z_{i,t3}$, or $z_{i,int} = (z_{i,t2} \cdot z_{i,t3})$. Thus, the coefficient β_{int} can be considered the interaction of an infected diet across both trimesters, or the change in the severity of an overall diet effect from the second to the third trimester; we call this the "interaction effect."

Equation (3.1) models a two factor factorial treatment design in a randomized complete block experiment design. This was the analysis specified by the initial study protocol, which investigated the relative effect of two binary factors and their mutual interaction on the means of several measured responses. The protocol model intended to use blocks to increase precision when estimating differences in the means of the measured response(s).

For the hypothesis tests associated with the ANOVA for the `protocol` model, the researchers suspected rejecting the null hypothesis for the second trimester effect ($H_0 : \beta_{t2} = 0$), for the third trimester effect ($H_0 : \beta_{t3} = 0$), and for the interaction effect between the two trimeters

($H_0 : \beta_{int} = 0$). In each case, the associated F-tests are represented by the "partial" (Type III) ANOVA sums of squares, or the additional contribution of each source of variation, assuming all the other sources are included. The "block effect" is estimated but not formally tested.

3.1.2 Researcher models

The researcher approach first ignores the effect of the experiment's original blocks from the randomized complete block design, and then ignores the second trimester effect and its interaction to create an unadjusted model. The researchers choose this avenue based on the ANOVA results from the `protocol` estimated on the original sheep dataset.

First, although the original blocks are presupposed to have an effect on the means of the measured response(s), this is not necessarily a correct assumption. If incorrect, then using the original blocks can actually decrease the overall precision of estimated differences in the measures response means by reducing the degrees of freedom available for estimating error. A small (partial ANOVA) F statistic for the blocking factor suggests that blocking factor may not be important; because of this, the researchers ignored the originally created blocks. Ignoring the block effect, the corresponding reduced model (compared to Equation (3.1)) becomes the `noblock` model,

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{t2} z_{i,t2} + \beta_{int} z_{i,int} + \varepsilon_i \tag{3.2}$$

Second, the researcher was initially interested in investigating which trimester (if either) produced stronger evidence of infected fescue's detriments to fetal development. Again, small (partial ANOVA) F statistics for the second trimester effect and for the interaction effect suggests that only in the third trimester does fescue have a significant effect on the measured response(s). Simultaneously ignoring the second trimester and interaction effects, the corresponding model becomes the completely unadjusted model (`t3only`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \varepsilon_i. \tag{3.3}$$

Note that the researchers also considered a model with no interaction effect (`noint`)

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{t3} z_{i,t2} + BLK_i + \varepsilon_i \tag{3.4}$$

and a model with neither block nor interaction effects (`t3 + t2`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{t2} z_{i,t2} + \varepsilon_i. \tag{3.5}$$

Observe that the error term in Equation (3.2) combines the error term from Equation (3.1) with the BLK_{sub} term from Equation (3.2). Similarly, from Equation (3.2) to Equation (3.3), the terms $\beta_{t2}z_{i,t2} + \beta_{int}z_{i,int}$ combine with the error term from Equation (3.1). Considering Equation (3.2) nested within Equation (3.1) and Equation (3.3) nested within Equation (3.2), we also note that decreasing the number of estimated quantities (our β's), naturally increases the information available to estimate our remaining quantities.

Even more notably, observe that by ignoring the original blocks, the resulting dataset is structurally indistinguishable from that of a two-factor analysis of variance with interaction for which blocks were never made; a difference remains in the treatment assignment mechanism of this situation. Even though the blocks are ignored in Equation (3.2) , they were not ignored when assigning treatment levels (treatments); indeed, by design, treatments were assign at random to experimental units only within their pre-determined blocks.

Similarly, observe that Equation (3.3) is structurally identical to one-way fixed analysis of variance with neither interaction nor block effects, as though the initial design had been a completely randomized design (CRD) with a single binary treatment. Again, the actual treatment assignments did consider the trimester and interaction factors.

Third, the researcher approach uses linear covariate adjustment to include three covariates as potential predictors. As lamb number is $x_{i,lmno} \in \{1, 2, 3\}$, the coefficient β_{lmno} in the shorthand $\beta_{lmno}x_{i,lmno}$ can be considered a "lamb number effect." Similarly, as genotype $x_{i,geno} \in \{AA, AG, GG\}$, the coefficient β_{geno} in the shorthand $\beta_{geno}x_{i,geno}$ can be considered a "genotype effect," and as initial BCS is $x_{i,ibcs} \in \{1, 2, 3, 4\}$, the coefficient β_{ibcs} in the shorthand $\beta_{ibcs}x_{i,ibcs}$ can be considered a "initial BCS effect." (See Appendix A for details.) Using lamb number and genotype, the final researcher model (`lgcov`) is

$$y_i = \beta_0 + \beta_{t3}z_{i,t3} + \beta_{lmno}x_{i,lmno} + \beta_{geno}x_{i,geno} + \varepsilon_i, \tag{3.6}$$

and the model including all three covariates of lamb number, genotype, and initial BCS (`allcov`) is

$$y_i = \beta_0 + \beta_{t3}z_{i,t3} + \beta_{lmno}x_{i,lmno} + \beta_{geno}x_{i,geno} + \beta_{ibcs}x_{i,ibcs} + \varepsilon_i. \tag{3.7}$$

Note that one could also consider models using only lamb number (`lcov`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{lmno} x_{i,lmno} + \varepsilon_i, \tag{3.8}$$

using only genotype (`gcov`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{geno} x_{i,geno} + \varepsilon_i, \tag{3.9}$$

using only initial BCS (`bcov`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{ibcs} x_{i,ibcs} + \varepsilon_i, \tag{3.10}$$

using only lamb number and initial BCS (`lbcov`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{lmno} x_{i,lmno} + \beta_{ibcs} x_{i,ibcs} + \varepsilon_i, \tag{3.11}$$

or using only genotype and initial BCS (`gbcov`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{geno} x_{i,geno} + \beta_{ibcs} x_{i,ibcs} + \varepsilon_i, \tag{3.12}$$

The researcher models of most interest are the completely unadjusted model, Equation (3.3); the researcher final model, Equation (3.6), or the model the researchers ultimately chose; and Equation (3.7), the model containing all three post-hoc covariates.

Also note that although the post-hoc covariates of lamb number, genotype, and initial BCS are discrete (as in Appendix A), we follow the researcher practice of including them as continuous covariates in the linear models. To achieve this, the researchers ranked the nominal genotype levels, ordering the three genotypes from least to most resilient to the effects of infected fescue on sheep fetal development. Although this difference exaggerates the significance of the (partial) significance of each of these covariates (see Table 16), it does not change the overall result of the (researchers' final) analysis; a full discussion is contained in Appendix E.

3.1.3 Potential model(s)

Finally, the potential approach uses matching methods to group the observed experimental units in different ways but still using information from the lamb number, genotype, and initial BCS of each ewe. The result of each the matching method is then modeled as

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(\cdot) + \varepsilon_i$$

where, similarly to the previous term BLK_i, the term $GRP_i(\cdot)$ is a series of indicator variables (see Appendix A) representing the "group effect," or the effect of the group to which the ith ewe is assigned. Recall that "blocks" refers solely to the researchers' originally created groups while the more general "groups" refers to any other way of grouping the ewes. Here, the ten potential models each has a different group effect $GRP_i(A), \ldots, GRP_i(J)$ and each may also produce different numbers of groups for different datasets. Thus, the error term in the potential approach models represents response variation not explained by the linear model, but the groups may be made using the full joint covariate distribution. That is, the error term for some groups may not include all possible covariate interaction effects, as some of these interactions may be captured in the group information.

We evaluate the results of various ways of making new groups and highlight those generated by four matching methods. The matching methods were chosen to heuristically represent two types of matching methods. The first two, matching on propensity score (PS) distance with a greedy nearest neighbor algorithm, represent an introductory matching method, such as might be common from a first use of propensity scores. The third, matching on Gaussian kernel distance with an optimized assignment algorithm, represents a more complex category of the many available matching methods; our kernel-based might be employed by those with a strong understanding of matching theory who are still resource-bound to pre-written software packages.

3.1.3.1 Propensity score models

The first highlighted potential model (`psmatchH`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(H) + \varepsilon_i, \tag{3.13}$$

uses matching on a propensity score distance with a greedy nearest neighbor algorithm. The second highlighted potential model (`psstratG`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(G) + \varepsilon_i, \tag{3.14}$$

differs from the first only in that after matching, it attempts to create five subgroups from the result.

Recall that a propensity score is the probability that an experimental unit receives treatment, assuming that the experimental unit has fixed covariate values. In a randomized experiment, propensity scores are known probabilities, but after a binary treatment is assigned, propensity scores can be estimated as in an observational study. Although the literature suggests many refinements of propensity score procedures, popular practice does not eschew the most basic propensity score procedures [26].

We thus naively estimate the propensity score via logistic regression of observed treatments on all available covariates [26]. That is, the propensity score of each ewe is estimated by logistic regression of the of the third trimester indicator $z_{i,t3}$ on the sum of the three covariates of lamb number, genotype, and initial BCS. Note that limiting treatments to a single binary factor, the third trimester effect, allow propensity score estimation.

As a summary of covariate information, the propensity score can be included in general adjustment methods in multiple ways ([17, 26]). In two variations, we use a greedy nearest neighbor algorithm ([25]) to first attempt attempt pairwise matches. This has a natural interpretation in a sheep experiment. Second, we attempt to condense the matches into five subgroups, similar to methods that are sometimes popular ([9, 26]). Recall that a greedy algorithm creates matches without regard to matches already made ([17]). In the specific algorithm executed, the ewes are split into treated and untreated lists and sorted, smallest to largest, on their propensity score. Without restrictions on maximum allowable disparity between the propensity scores of grouped units (the package default), this approximates a pairwise matching.

That is, the treated ewe with the smallest propensity score is matched to the untreated ewe with the smallest propensity score. The randomized design of the original experiment implies that all of our observed experimental units are useful for estimating the third trimester effect. As such, note that although all unmatched observations are discarded, per the algorithm default, matching is executed without replacement to maximize the number of matched ewes. If there are tied potential matches, Holmes ([26]) indicates that these should be broken randomly ;this is important, as ties will be prevalent in the simulated datasets.

These two matching methods are executed with the 'MatchIt' package in R [24, 25]. The package input parameters were specified based on works [1, 10, 16, 29, 45, 47]; also, texts [17] and [26].

3.1.3.2 Kernel-based model

The third highlighted potential model (`kernmatchJ`),

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(J) + \varepsilon_i, \tag{3.15}$$

uses on a kernel distance with an optimized assignment algorithm.

Recall that a kernel function measure distances between each pairwise combination of treated and untreated experimental units. Literature shows that forming subgroups via simultaneous optimization of multiple covariates' observed moments can sometimes be more useful than matching on a single covariate, which Kallus demonstrates in [6]. We thus estimate Gaussian kernel distances [30], [31] between every pairwise combination of ewes that did and did eat the infected diet in the third trimester.

As kernel functions exist on continuous spaces, we must represent lamb number, genotype, and initial BCS are continuous covariates, as the researchers chose in the researcher approach. Before calculating the pairwise kernel distances, we first standardize the observed covariates [30]. To set the Gaussian kernel bandwidth, we use a plug-in estimator; then, we calculate the pairwise kernel values between each treated and untreated ewes [22].

Recall that an optimized matching minimizes the overall distance between the units across all groups, typically by using a network flow algorithm [26]. Recall also that this specific optimized assignment algorithm [20] will create a full match, or a set of groups that optimizes the ratio of treated and untreated units within each group. Again, under the randomized treatment assignment mechanism of the original experiment, we should match every experimental unit, if possible. Optimal matching can use every ewe. This specific matching method does so since several untreated units may be matched to one treated unit and vice versa in full matching; full matching also explains how ties should generally be treated. Optimized matching is executed with the 'optmatch' package in R [19].

3.1.4 Other models

There are seven (7) more models to define from the potential approach. Using the eight blocks originally assigned by the experiment as groups, (`blocksA`) produces the model

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(A) + \varepsilon_i. \tag{3.16}$$

Using no groups (`nogroupB`), where the estimates of mean responses are identical to those from Equation (3.3), produces the model

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(B) + \varepsilon_i. \tag{3.17}$$

Using five groups created from stratifying on the continuous version of lamb number (`lmnostratC`) produces the model

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(C) + \varepsilon_i. \tag{3.18}$$

Using five groups created from stratifying on the continuous version of genotype (`genostratD`) produces the model

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(D) + \varepsilon_i. \tag{3.19}$$

Using five groups created from stratifying on the continuous version of initial BCS (`ibcsstratE`) produces the model

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(E) + \varepsilon_i. \tag{3.20}$$

Randomly assigning each experimental unit to one of five groups (`randstratF`) similar to the creation of the simulated datasets, produces the model

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(F) + \varepsilon_i. \tag{3.21}$$

Finally, using groups based on optimal matching of experimental units using a propensity score distance (`psoptmatchI`) produces the model

$$y_i = \beta_0 + \beta_{t3} z_{i,t3} + GRP_i(I) + \varepsilon_i. \tag{3.22}$$

3.2 Datasets

We have two types of datasets, the original dataset from the sheep researcher and simulated datasets (nsims=500). We use the original sheep dataset to recreate the researcher's analysis and to consider how we might use post-hoc covariate information beyond adding the covariates directly to a linear covariate adjustment. From the original sheep dataset, we simulate other sheep datasets, as though the researchers repeated their experiment many times.

If the researchers repeated experiment many times, each repetition's ewes will have the same covariate-response relationships and the same relationships between covariates at the population level. Also, each repetition's ewes would be impregnated in blocks created as a function of their initial BCS, so the planned randomization of treatment assignment in complete blocks would continually be challenged. In the actual experiment, the researchers believed that the end result of the complicated ewe impregnation scheme and the possible redistribution of ewes to other blocks different from their impregnation group resulted in an approximation of randomly assigning treatments within blocks.

The simulated datasets thus represent multiple potential "realizations" of how the actual experiment's results, so we can analyze each new dataset using the same three analysis approaches established above. We estimate every model on every dataset. Our specific interest is in the performance of these models, relative to one another, to estimate the third trimester effect.

3.2.1 Original dataset

After gestation, several weights recorded on each pregnant ewe measured fetal development; these were the researchers' response variables. After dropping five ewes and changing the block assignment of four from their initial impregnation group, the original dataset had the same structure as a randomized complete block design (RCBD) in which (nobs=32) were sorted into groups of size four (nblocks=8). Each ewes was approximately randomly assigned to one of four treatments within each block, resulting in an experiment with a randomized complete block design. The four treatment levels resulted from the experiment's two binary factors, the presence or absence of endophyte infested fescue seed in each ewes' second and third pregnancy trimesters. Additional covariate information of lamb number, genotype, and initial BCS was also recorded on each ewe.

Again, recall the following information about the covariates chosen by the researchers. First, lamb number is the number of fetuses each ewe gestates. To add each ewe's lamb number to the linear model, lamb number is a discrete integer covariate with three levels. In the sheep study, lamb number is a post-hoc covariate because it is not fully detectable in early sheep pregnancy and so is unavailable until after treatment assignment. Second, genotype refers to the value of a genetic marker that has been found in recent livestock literature to be associated with resistance to the detrimental effects of endophyte infected fescue. To add each ewe's genotype to the linear model, genotype is a nominal covariate with three levels. In the sheep study, genotype is a post-hoc covariate because physical constraints prevented the researchers from genotyping the ewes in advance of treatment assignment. Third, BCS is a livestock measurement in which healthier ewes tend to score more highly (i.e., higher values correspond to sheep more likely to have a healthy pregnancy). To add each ewe's initial BCS to the linear model, initial BCS is an integer covariate with levels from one to four.

Finally, recall that although lamb number is an integer, genotype is nominal , and initial BCS is an integer, the researchers chose to approximate these three covariates with continuous versions. Otherwise, they would be included as a series of indicator variables, as in Appendix A. Unlike the covariates' effects, discrete block group effects are represented as a series of indicator variables, as established above.

3.2.2 Simulated datasets

The sheep experiment ended with a design structure of 32 observations initially arranged in eight (8) blocks, each with four (4) observations. The initial sheep dataset is, however, merely one dataset, and as a single sample provides only a snapshot of a more general sheep population. We use the initial sheep dataset to simulate some other sheep datasets, as though the researchers had been able to repeat their experiment many times. The overall simulation approach is bootstrap resampling of the original sheep dataset. Each new dataset has 32 observations to mimic the initial sample, and this process is repeated several times (nsims=500).

If the researchers repeated experiment many times, each repetition's ewes will have the same covariate-response relationships and the same relationships between covariates at the population level. In basic bootstrap resampling, each observation from the initial sample has equal probability of being re-drawn, with replacement, as an experimental unit in the new dataset. This imitates choosing a new sheep of the same "type," or joint covariate distribution with similar covariate relationships to the measured response, but in different proportion to the protocol sample (as a subsequent sample could be). That is, we want to express the idea of preserving the relationships between the response and covariate and between covariates on a population level.

In the actual experiment, the researchers believed that the end result of the complicated scheme of ewe impregnation and the possible redistribution of ewes to other blocks different from their impregnation group resulted in an approximate of randomly assigning treatments to blocks. After resampling, eight (blocks) are completely randomly assigned to reflect complications in execution of the experiment that generated the initial dataset. In each simulated dataset, we use random permutation of the sequence $\{1, 2, \ldots, 8\}$, repeated four (4) times to simulate the complicated impregnation process. The result is that in each new dataset, the originally assigned treatments remain with each ewe but the blocks are reassigned.

Chapter 4

Results

To quantify detriments to fetal development in sheep, the sheep researchers recorded weights of a variety of reproductive organs. The response variables we consider are these weights. The analyses were conducted for every response variable. However, each ewe's total uterine weight was used as the main illustration of the results.

We give the results in two sections. First, using the original sheep dataset, we show three ANOVA tables for the ewe's uterine weight, based on models from from the protocol or researcher approaches. Appendix B provides these three tables using the original sheep dataset for every response variable. Second, for both the original sheep dataset and for every simulated dataset, we estimate the third trimester effect, or difference in means for the ewes with and without the infected third trimester diet. Boxplots summarize the estimated third trimester effect for every model across all available datasets (ndatasets=501). Appendix C provides the boxplots for every response variable. For quick reference, the names of the models are summarized in Appendix D.

Each boxplot shows the observed distribution of the estimated difference in means for the third trimester indicator Z_{t3}, or the estimated third trimester effect, along with the observed standard error of every estimate, or the standard deviation of the difference in means. Estimates and their standard errors are recorded for every model and on every simulated dataset in additional the original sheep dataset. Boxplots label each model type along the horizontal axis, and each figure shows the results for a difference response variable. In each figure, the bottom shows the estimate of difference in means (`EST`) while the top shows standard error of the difference in means (`SE`) for each response.

The boxplots thus compare the observed distributions of the estimate and its standard error between models within each response. Within every response, we overwhelmingly observe that the models' behavior relative to one another is very similar for these measures of sheep fetal development. Through the results, we use ewes' uterine weight ("yUteWtg") as an illustrative example. Uterine weight displays similarities and differences that are consistent across the other response variables.

4.1 Case study results

The methods section highlighted several models from the researcher approach, among which were the originally specified model, Equation (3.1), initially specified by the researchers experimental design; the completely unadjusted model, Equation (3.3), considered by the researchers after seeing the ANOVA results for the originally specified model; and the three covariate model, Equation (3.7), containing the covariates of lamb number, genotype, and initial BCS, which the researchers examined before adopting the final researcher model. Tables 4.1, 4.2, and 4.3, reproduced from Appendix B, provide the partial (Type III) ANOVA results Equations (3.1), (3.3), and (3.7), respectively, estimated on the original sheep dataset for the ewe's average uterine weight ("yUteWtg"). Note again that uterine weight is used as an illustrative example, as each of the responses show the same model trends for the original sheep dataset. Appendix B provides these three table for every response variable.

Table 4.1: ANOVA table for uterine weight from Equation (3.1)

yUteWtg: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	980652	980652	4.15	0.0544
block	7	1106896	158128	0.67	0.6953
tri3=toxicc	1	184030	184030	0.78	0.3873
tri3*tri2	1	67912	67912	0.29	0.5974
residuals	21	4958680	236128	NA	NA

Table 4.2: ANOVA table for uterine weight from Equation (3.3)

yUteWtg: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	980652	980652	4.66	0.0391
residuals	30	6317518	210584	NA	NA

Table 4.3: ANOVA table for uterine weight from Equation (3.7)

yUteWtg: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	1187805	1187805	10.90	0.0027
lamb num	1	3078709	3078709	28.24	0.0000
genotype	1	12419	12419	0.11	0.7383
initial BCS	1	2984	2984	0.03	0.8698
residuals	27	2943285	109011	NA	NA

Table 4.1 provides the results the researchers observed when using the originally specified model, Equation (3.1). From these results, they decided to omit the original block effect and the second trimester and interaction effects. The result was the completely unadjusted model, Equation (3.3), provided in Table 4.2. In Table 4.2, the researchers observed that moving from the originally specified model, Equation (3.1), to the completely unadjusted model, Equation (3.3), increases the "significance" of the third trimester effect. This does not consider whether the significance is better characterized by a more precise third trimester effect, which reduces the "noise" by ignoring terms from Equation (3.1), or whether the significance increase is better characterized by the adjusted difference degrees of freedom use in the p-value calculations.

That is, the researcher decision to ignore all but the third trimester effect is an implicit form of model building or variable selection, one often encountered in small sample livestock experiments. In this method, the researchers actually performed an informal preliminary test for ANOVA error pooling. Note that in such a situation, exact adjustments to the overall Type I error rate are described in literature concerning preliminary tests. Although formal preliminary tests for ANOVA error pooling are a diversion from the objective of evaluating the potential approach, we mention this connection in passing as it later becomes part of our final recommendation.

Returning to our discussion of Tables 4.1, 4.2, and 4.3, consider moving from Table 4.2 to Table 4.3. Recall that from the completely unadjusted model, Equation (3.3), the researchers wanted to use covariate information available for each ewe. From Table 4.2 to Table 4.3, adding covariates to the completely unadjusted model, increases the "significance" of the third trimester effect. Use of lamb number is highly significant, as expected, in predicting weights of ewes' reproductive organs. Of lamb number, genotype, and initial BCS, lamb number is by far the most important when each covariate is considered separately.

In their final researcher model, Equation (3.6), the researchers further chose to use genotype because suggestions from their literature indicate that there should be a genotype effect. Observe that genotype is a very "insignificant" as a post-hoc covariate. This could be due to the fact that genotype is numerically unbalanced across the third trimester treatments, or it could be due to the fact that that genotype is highly correlated with the other covariates.

Based on Table 4.3, the researchers did not include initial BCS, in their final researcher model, Equation (3.6). Rather, Table 4.3 confirmed for the researchers their previous decision to ignore the block effect in the originally specified model. As Initial BCS was the covariate on which

their blocks were based, the researchers were not surprised that it was not "significant," as they had already determined the block effect to be "not significant."

Overall, this sequence of three tables, Tables 4.1, 4.2, and 4.3, explains the researcher's perspective. The sequence catalogues the change in the third trimester effect and its "significance" as the researchers moved between covariate adjusted linear models. It also brings attention to the "problem" of the genotype effect, a real-life effect not at all detected in the original sheep dataset. Since there *should* be a genotype effect based on livestock science, perhaps the potential approach can offer an alternate way of including the genotype information.

4.2 Simulation results

Recall that Appendix C provides boxplots for every response variable and that Appendix D provides summaries of all models. Each boxplot figure compares model performance across all available datasets for a different response variable, with models along the horizontal axis. Model performance is reported as each dataset's third trimester effect (bottom, denoted "EST") and its corresponding standard error, or the difference in means and the standard deviation of the difference in means (top, denoted "SE"). Boxplots thus show the observed distribution of EST and SE. The boxplots for EST show connected means for each model within each response.

As with the tables from the original sheep dataset, uterine weight ("yUteWtg") is used as an illustrative example, as the same model trends persist across response variables. Figure 4.1 reproduces the boxplots for uterine weight from Appendix C; Figures 4.2 and 4.3 merely magnify parts of Figure 4.1. These figures reveal trends between specific models and between the protocol, researcher, and potential approaches.

Figure 4.1: Boxplots for uterine weight

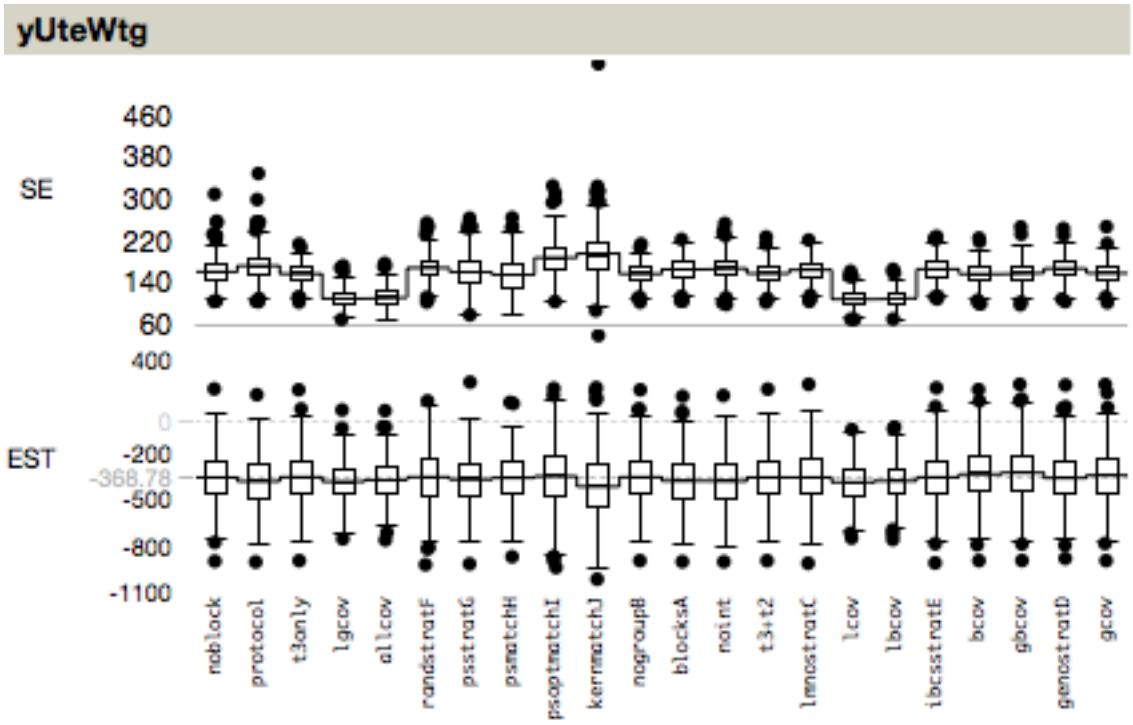


Figure 4.2: Boxplots for uterine weight, researcher approach

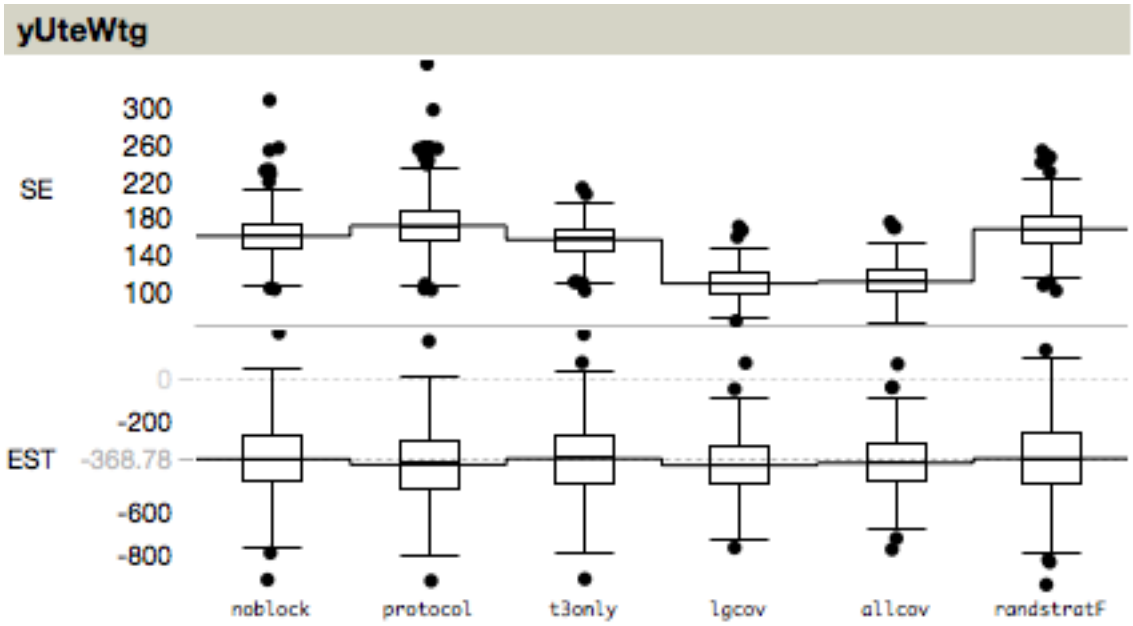
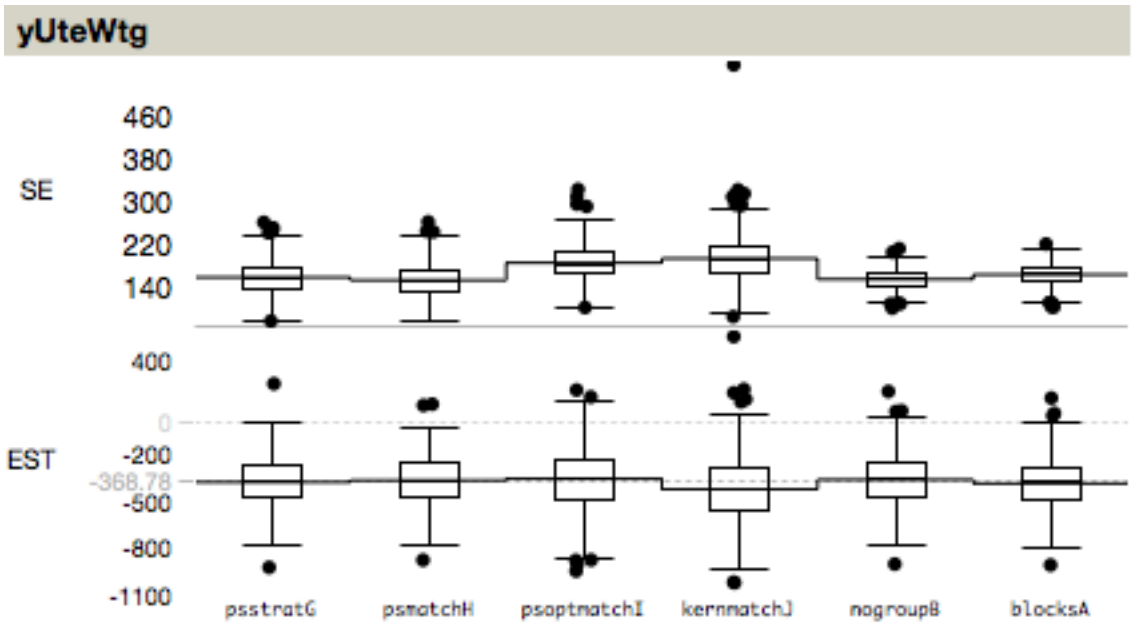


Figure 4.3: Boxplots for uterine weight, potential approach



From the protocol model, Equation (3.1), to the completely unadjusted model, Equation (3.3), the estimate remains the same. The boxplots suggest that the protocol model, Equation (3.1), is preferred over ignoring both the blocks and the second factor, the presence or absence of infected fescue in the second trimester. This is because the standard error of the completely unadjusted model, Equation (3.3), has a higher mean with wider range from a higher maximum and similar minimums. From the final researcher model, Equation (3.6), to the researcher unadjusted model, the estimate has the same mean but a narrower range from both a lower maximum and a higher minimum, i.e., both closer to the mean. The standard error of the final researcher model, Equation (3.6), has a notably lower mean and a wider range with a distinctly higher maximum. The boxplots suggest that the final researcher model, Equation (3.6), is an improvement over the completely unadjusted model, Equation (3.3).

From the final researcher model, Equation (3.6), to the three covariate model, Equation (3.7), the estimate has the same mean but a narrower range from both a lower maximum and a higher minimum, i.e., both closer to the mean. The standard error of the three covariate model, Equation (3.7), has a higher mean and a wider range with a distinctly higher maximum and lower minimum, i.e., both are further from the mean. There is also a pronounced upper outlier in the three covariate model. The boxplots suggest that further including the initial BCS, a covariate that may not be strongly related to the response, in the linear model is not recommended.

From the lamb number only to the final researcher model, Equation (3.6), shows a similar inefficiency. However, livestock literature indicates that genotype is useful in predicting the effects of endophyte infested fescue on fetal development. We thus want to compare the final researcher model, Equation (3.6), to models using the groups from matching methods.

Of the highlighted matching methods from the potential approach models, recall that we have Equation (3.14), made with a greedy nearest neighbor algorithm and then attempts to make five subgroups of the matches; Equation (3.13), made with the same greedy nearest neighbor algorithm but does not attempt to make five subgroups of the matches; and Equation (3.15), made with an optimal assignment algorithm using a Gaussian kernel distance distance. In this section, we additionally highlight a fourth model from the potential approach, Equation (3.22), which was made with an optimal assignment algorithm using a propensity score distance; this last model shows interesting results.

The nearest neighbor greedy matching methods, Equation (3.13) and Equation (3.22), have

overall smaller standard error than the optimal methods, Equation (3.22) and Equation (3.15). This is as expected, as the optimal methods match all the experimental units, while the greedy methods discard ewes that cannot be matched. Because the greedy matching methods, Equation (3.13) and Equation (3.22), can change the overall sample size of the experiment, they are not preferred. Instead, the optimal assignment methods are preferred in this setting, despite their overall higher standard error. That is, the gains in precision from the greedy methods to the optimal methods are meager compared to the loss of experimental units. Loss of experimental units is an important issue since experiments tend toward smaller sample sizes compared to the large observational studies under which matching techniques are developed.

Consider that the greedy nearest neighbor matching and how it discards observations that are not enough "overlap" in the covariate space. These matching methods target creation of two treatment groups with the focus of reducing bias; they make the two treatment groups as similar as possible in the covariate space. As only a finite amount of information is available in a dataset, these methods achieve the bias reduction by allowing larger variance among the retained experimental units. This is not always immediately relevant, as seen in the boxplots, as the overall variance may decrease from reduced sample sizes. We see this in the greedy methods. They look like they have better variance than the optimal methods but recall that the optimal methods preserve the experiment's original number of ewes. So all the matching methods have overall higher variance.

Among the four models using the groups from matching methods, Equations (3.14), (3.13), (3.22), and (3.15), between the two optimal matching methods, we see the following. The optimal matching with a kernel distance, Equation (3.15), has a distinctly biased estimate compared to optimal matching with a propensity score distance, Equation (3.22). Further, within optimal matching, the kernel distance method has an overall higher standard error compared to the propensity score distance method. Note also that for the kernel distance function requires that covariates be continuous. From the greedy methods, Equation (3.14) is biased for a number of the responses, and of the optimal methods, Equation (3.15) is biased for almost all of the responses. The matching methods assume there is response bias, so they correct for response bias. In an experimental setting, if we assume that there is no response bias across treatment groups due to a randomization mechanism, then matching methods may "overcorrect" or introduce bias in estimates of the mean response.

Because the kernel-based distance method presents a distinctly lower estimate, it is not preferred among the optimal methods. Further, the kernel-based distance method has a larger (both

higher and wider in range) standard error compared to the propensity scored distance method, so we would prefer the propensity scored distance, in this situation, to the kernel distance. However, neither show a standard error as small as the completely unadjusted model, Equation (3.3), Equation (3.17), which is equivalent to.

Overall, the boxplot results show three patterns. First, moving from the originally specified model Equation (3.1) to the completely unadjusted model, Equation (3.3), Equation (3.3), in the first part of the researcher approach and implicitly considering the Equation (3.2), Equation (3.4), and Equation (3.5) models as well, the estimate remains approximately the same, but the standard error is higher and slightly wider, with the same minimum and a higher maximum. The suggests that discarding all but the third trimester effect may not be an unreasonable approach.

Second, moving from the completely unadjusted model, Equation (3.3), Equation (3.3), model to the final researcher model, Equation (3.6), in the second part of the researcher approach and implicitly considering Equations (3.11), (3.10), and (3.12) as well, the estimate in means remains the approximately same, but the standard errors have a distinctly lower center. Thus, it is primarily lamb number that is improving the estimates' precision. So models that include lamb number are preferred.

We see that between the models with and without the initial BCS (from Equation (3.11) to quation (3.8); from Equation (3.12) to Equation (3.9) , and from Equation (3.7) to Equation (3.6)), the estimate maintains approximate the same center and presents only slightly higher and wider standard errors. The suggests that direct linear covariate adjustment is not unreasonable in this situation, so long as lamb number is included. Heuristically, the number of fetuses in a pregnant ewe is the critical covariate in this situation, the most predictive of the ewe's reproductive tissue weights after gestation, but initial BCS does not appear to be predictive of uterine weight.

Third, direct covariate adjustment gives more precise estimates than any of the potential approach models. Among the four matching methods of the potential approach, all have larger standard errors compared to Equation (3.21), in which five groups are randomly assigned to each ewe, and compared to Equation (3.16), adjustment only by each experiment's original blocks. Many matching methods, were initially developed to reduce the (response) bias in observational studies, so they are not always useful in increasing estimation precision. As seen in the final researcher model, Equation (3.6), even linear covariate adjustment using covariates insignificant in a partial ANOVA test is preferred over the matching methods.

Chapter 5

Summary and Discussion

Our case study began with an experiment that intended a randomized complete block design and planned an analysis with two binary factors, their interaction, and the intended blocks. Following the analysis of Clemson sheep researchers, we ignored all terms as predictors of detriments to fetal development in sheep, save the effect of a toxic endophyte infested fescue seed diet in pregnant ewes. We then used the post-hoc covariates of ewe's lamb number and genotype to further adjust our estimate for the remaining third trimester effect.

The boxplots based on simulated data show three overall results. The first is that removing the "not signficant" terms from the protocol model results in a third trimester treatment effect whose estimate is approximately unchanged but whose the standard error is higher and slightly wider, with the same minimum and a higher maximum. The suggests that discarding all but the third trimester effect may not be an unreasonable approach.

The second result is that adding covariates results a third trimester treatment effect whose estimate is approximately the same but whose standard errors have a distinctly lower center. Thus, it is primarily lamb number that is improving the estimates' precision. So models that include lamb number are preferred. We also see that between the models with and without the initial BCS that the third trimester treatment effect whose estimates are again approximately unchanged but whose standard errors are only slightly higher and wider. The suggests that direct linear covariate adjustment is not unreasonable in this situation, so long as lamb number is included. Heuristically, the number of fetuses in a pregnant ewe is the critical covariate in this situation, the most predictive of the ewe's reproductive tissue weights after gestation, but initial BCS does not appear to be

predictive of uterine weight.

Third, direct covariate adjustment gives more precise estimates than any of the matching methods of the potential approach. Among the four matching methods, all results have larger standard errors. Many matching methods, were initially developed to reduce the (response) bias in observational studies, so they are not always useful in increasing estimation precision. As seen in the final researcher model, Equation (3.6), even linear covariate adjustment using covariates insignificant in a partial ANOVA test is preferred over the matching methods.

Literature suggests that matching methods could be at least as useful in the creation of "better" (more estimation precision) blocks than covariate adjustment by lamb number and genotype; this is especially true of kernel-based matching methods ([32]). That is, a "better" matching method for this particular situation is one that increases the precision of our estimate of the difference in means contrast and minimizes our standard error of the difference in means. We also note also that although researchers might prefer to maximize the magnitude of a difference in means contrast, this alone is not a valid goal when comparing statistical methods.

Although such a "better" matching method may exist in theory, discovering and verifying such a matching method requires extensive effort. Specifically, every matching method requires careful compilation of distance, structure, and algorithm choices from a myriad available. At present, matching methods that have been compared for datasets of varying characteristics is far from exhaustive.

5.1 Discussion

In the case study of the original sheep dataset, sheep researchers wanted to ignore the blocks originally created for their randomized complete block design. This is because they acknowledged there were unforeseen physical challenges in the study execution that inarguably caused the randomization of treatments to be different from the study protocol. The sheep researchers justified ignoring their blocks because they believe the actual randomization was "close enough" to a complete randomized design (i.e., although their planned treatment randomization failed, the believed that it did not fail drastically). The researchers also observed the p-value of the of the "block effect" as insignificant in the (partial) ANOVA tests and noted that removal of the "block effect" from the ANOVA tests increased the significance of the p-value for the "third trimester effect," or the observed difference in means of the presence or absence of infected fescue in the sheep's third trimester diet. After exploring their results, the sheep researchers were highly interested in reporting the third trimester effect, which was significant in their ANOVA tests.

As in the original sheep dataset, difficulties in study execution that disrupt execution of randomized (complete) block designs are not uncommon in certain sciences. We offer two possible solutions to researchers' general desire of discarding blocks when planned randomization fails. The first, matching, represents a multitude of alternative methods to create a new set of blocks, again from a researcher-selected set of covariates. The second, post-hoc covariate adjustment, proceeds after discarding the original blocks in favor of post-blocking on select covariates. In both cases, the researchers discard their original blocks in favor of a new set of analysis groups, or groupings of the observed experimental units on available covariate values.

As (linear) post-hoc covariate adjustment assumes randomized treatment assignment, inference based on such models inference depends on researchers' relative belief that the failed randomization is acceptably close to the desired randomization. In absence of this belief, researchers should ideally appeal to matching methods, which were developed under situations that deliberately lack randomized treatment assignment (i.e., observational studies). Matching methods, however, are myriad and extremely complex, each defined by careful selection of distance, structure, and algorithm (d-s-a) combinations [16]. Because of the many choices, further development is necessary to identifying definitive dataset characteristics to indicate which d-s-a combinations are most efficient in certain data situations. Exploration of such factors, or dataset characteristics, is limited to

simulated comparisons of limited matching methods, typically on datasets with certain fixed characteristics [1, 23, 29], and Kallus [31] provides some early theory comparing matching methods that may ultimately increase our understanding of which methods work best for various datasets.

Thus, although an efficient matching method will ultimately be discovered for small-size science studies with failed block randomization of treatments, investigation of exact d-s-a selections will be required until we can identify how to construct an ideal d-s-a combination for this type of dataset. Our exploration of the original sheep dataset indicates that in readily executable matching methods, precision in naive contrast estimation for linear models is not enough improved over well-developed theory of preliminary testing combined with post-stratification to be worth the higher resource cost (e.g. researchers' time investment in analysis methods atypical for their science). Until theory indicating the relative advantages of various matching methods in certain types of datasets is better understood and until more matching methods are easily accessible, we recommend post-stratification combined with preliminary testing. This acts as a formalized version of the sheep researchers' automatic removal of their original blocks (preliminary testing) followed by post-hoc covariate adjustment (post-stratification) in this situation.

5.2 Recommendation

Our full recommendation for researchers who would automatically ignore failed block randomization and estimate treatment effects using linear covariate adjustment, instead combine formal preliminary testing to determine whether or not blocks should be ignored followed by post-stratification to estimate treatment effects and their relative standard errors. The bodies of literature for preliminary testing and for post-stratification are well-developed which we find quite exciting. Their results are immediately useable and, in the case of preliminary testing, extensively catalogued and extremely well-summarized. Bancroft published his original work describing the basic theory of error pooling [4] in 1944, followed by a series of papers further developing the idea [2]. In 1980 he jointly co-authored a user-friendly guide to the entire body of literature to date in [3].

There appears to be three reasons why these ideas are not often used in practice. First, the concept is only briefly mentioned in statistics methods classes and is not obviously executable in popular software packages, i.e. one-click button or option of "$\texttt{preliminary}_{\texttt{t}}\texttt{est} == \texttt{TRUE}$." Second, there is controversy in the statistical literature about the validity of a block test and the actual degrees of freedom that results from pooling. The final reason is that it only makes a difference in a few cases, as evidence by [5] and [36].

5.2.1 Preliminary test

Error pooling for ANOVAs is an application of conditionally specified models, which have been extensively studied in the literature ([2]); preliminary testing is a large part of this work. Generally, a statistical analysis has two phases, a preliminary testing phase and a primary testing phase. That is, preliminary testing is the process of "model building," wherein we compare potential predictors and forms of their combination, and primary testing is the more familiar statistical analysis, wherein we estimate, predict, and infer an average measured response. Use of preliminary testing, though not always explicitly declared, is omnipresent.

A common preliminary test determines whether or not to pool ANOVA sources of variation into error. This form of preliminary testing was originally presented in [4] and further developed in [7]. Bancroft and Han ([3]) describes this testing with the following example. Consider the the ANOVA for a two-factor factorial fixed-effects model without interaction. The significance of one of the factors may be in question if the factor's contribution to the overall variance of the response is minimal, i.e, if the factor's F-statistic (in a sequential ANOVA table) is relatively small. Since the objective is to determine the significance of an overall treatment effect (from either of the two factors), the preliminary test (of whether or not to include the factor in question in the primary testing phase) must employ a higher significance level than the original experiment in order to maintain the significance level of the overall treatment effect. In specific situations, comparing plotted size curves can determine the optimal significance for such a preliminary test, though commonly a 25% significance level will be acceptable [3].

The resulting primary analysis, based on a final analysis model, includes only the remaining factor if the preliminary test discards the other factor. Discarding a factor from the model 'pools' its associated variation with the ANOVA residual error. This type of preliminary testing is sometimes called a "sometimes-pooled test." Pooling error variation from factors with little contribution to the overall variation can increases the power available to test the effects of the remaining factors. Bozivich, et al [7] gives guidelines for fixed, random, and mixed ANOVA models, and we can investigate the potential advantages of pooling by comparing the power curves of ANOVA tests executed with and without pooling [3].

Ignoring the effects of blocks in an analysis based on a randomized complete block experiment design is an example of a sometimes pool test, or an error pooling preliminary test. For

example, the test would choose to use the originally specified model, Equation (3.1), if the block effect is not significant in the preliminary test; the same model without the block effect, Equation (3.2), otherwise. In the researcher approach, we also used error pooling to simultaneously remove the second trimester and interaction effects from the ANOVA model because although the researchers were initially interested in investigating which trimester (if either) produced stronger evidence of fescue detriments to fetal development, partial sums of squares F-tests of factor coefficients suggested that only in the third trimester does fescue have a significant effect on the measured response(s). Thus, the researchers determined to remove from their model the estimates second trimester effect and fescue-trimester interaction in order to increase the power of their model for detecting a fescue effect in only the third trimester; the model because Equation (3.3). The formal preliminary test, as described above and specifically detailed in [3] and [4], would not have discarded the original blocks at a 25% significance level [3].

5.2.2 Post-stratification

As previously mentioned, many researchers (in the sciences) carefully differentiate between the use of covariate information before treatment assignment (usually in an ANOVA with blocking) versus the use of covariate information in the model and analysis after treatment assignment (usually in an ANCOVA). In ANOVA, the mechanism of randomization that hopefully accompanies an experiment ensures (theoretical) orthogonality of covariate and treatment spaces; informally, randomization means that covariates should not contain overlapping information, on average, in an experimental setting. In ANCOVA, the idea that covariates' explanatory capacities towards a measured response are mutually exclusive is simulated through orthogonal projection. Additionally, in ANOVA use allows the number of observations within each covariate category or block to be numerically balanced, while in ANCOVA, treatment levels contain varying proportions of observations across covariate categories, so that the number of observations within each category could be unbalanced.

These two approaches motivated extensive investigation of the relative advantages and disadvantages of each case of covariate adjustment. Most recommendations are primarily based on a covariate's relative strength of association to the measured response (typically, their observed correlation), while consideration is also given to the size of a covariate's variance (relative to the overall error in the response) and to the underlying model complexity (whether or not a strictly linear model is truly appropriate). Detailed findings are presented in [9, 12, 37], and, most recently, [46].

In addition to the decision to whether or not to use covariates in the analysis before or after treatment assignment, further distinction can be made depending on whether covariates are measured before treatment application ("pre-experimental") or after treatment application ("post-experimental" or "post-hoc"). This distinction leads to the subtly different assumptions of ANOVA and post-stratification, respectively. Post-stratification creates a dataset with similar structure to that of an unbalanced blocking ANOVA; the primary difference, as detailed by [38], is that in post-stratification, the number of observations available to each treatment-covariate combination can be considered a random variable. That is, it is a random quantity as opposed to a fixed quantity as when blocking is done before treatment assignment. Measuring covariates after applying treatment may also generate concerns of whether post-hoc covariates are affected by the treatment [39].

Thus, when we use post-stratification to include new covariates, the overall structure of

experimental units in post-stratification will be identical to the structure of an (unbalanced) analysis of variance; the only difference is the treatment assignment mechanism. As the resulting dataset structure is the same as that of an incomplete block design, the same regression adjustment used in an incomplete block design is used in post-stratification. However, because the number of experimental units assigned to each treatment is random within strata, the variance (standard error) estimate of post-hoc covariates is larger than in a corresponding analysis of variance model [38].

5.3 Future work

Future work has two parts. The overarching goal is identifying specific dataset characteristics, or factors, indicating which matching methods choices (distance-structure-algorithm combinations) are most efficient for certain datasets. Considering the number of distance, structure, and algorithm chocies already proposed, our next step is to extend our comparison to many other matching methods by focusing on different combinations of distances, structures, and algorithms that are not directly available in packages. From here, there are several directions, the choice of which would be based on which methods are most promising for the broader goal.

5.3.1 Many directions

One direction designs a simulation in which a dataset's collinearity between covariates, correlation between response and covariates, and correlation between response and true underlying error are all precisely controlled. Such a simulation could directly investigate the relationship between the linear association between a response and covariates. To date, we have made strides in this direction, but we are still determining how to clearly characterize dataset characteristics that differentiate between dataset "types" and how to define an underlying "truth" model to which we will compare simulated results.

A second direction is algebraic comparison of the newly proposed distance metrics for matching methods, ultimately allowing an algebraic comparison of different distance metrics under different algorithms. Some distance metrics have been compared, as in [13], and this question was extremely well investigated in [34] in the mid-2000s, but since that time those results have not been updated with the many distance metrics since proposed.

A third direction evaluates matching methods for non-binary treatments, including of using matching estimators (designed for observational study situations) under the assumptions of randomized experiments. Developments ([28, 31]) in this direction are relatively new.

5.3.2 Broader perspective

For datasets with certain characteristics, we want to consider which characteristics, or factors, determine which matching methods work well. Literature on matching methods and related topics obliquely allude to the following relationships which we have intuited as potential factors for

important dataset markers based on repeated references.

We identified three primary areas for investigation that implicitly capture the underlying model "truth," or complexity; we call them "prognostic association," the joint relationship between a measure responses and covariates, "covariate overlap," the joint relationship between all available covariates and functions of covariates (including "treatments"), and "error complexity," the joint relationship between the response variation explained and not explained by a model. Note that model complexity is often characterized in terms of bias from selection and/or unknown treatment assignment mechanism, heteroscedasticity of error variance (sometimes, "endogeneity"), measurement error (potentially on every model term), bias from "omitted" covariates or higher-order terms, heterogeneous treatment effect. Similarly, we may additionally consider error from coding choices and datatype assumptions made for convenience.

Prognostic association refers to the combined strength of (association of) each covariate to response. Interest in prognostic association originates from original ANOVA vs Regression literature [9, 37, 12] in which it is measured naively as correlation or the coefficient of determination. Here, linear association (correlation) is far from the only measure of dependence between two random vectors [44], and literature suggests Hilbert-Schmidt norms ([15]) as a first further measure of association.

Error complexity refers to the relative size of the error variance and overall response variance. Interest in error complexity originates from use of F-test in ANOVA to assess variance ratios. A similar idea is seen in the use of variance ratios and degrees of freedom differences to set cases for preliminary tests (for ANOVA sums of squares).

Covariate overlap refers to the joint distribution of the "treatments," measured covariates, and functions of both. In linear models, the relationship between measured covariates is often characterized as multicollinearity, which could potentially be extended in a manner similar to extending linear correlation to the more general prognostic association. Further, ANOVA tests depend on the relative degrees of freedom of their sums of squares decomposition; more heuristically, relative degrees of freedom sizes is a way to characterize the relative number of observations in each treatment-block "cell."

Literature towards the idea of determining such factors is only recently coming under suggestions of investigation. Some examples are simulations on specific datasets [11], the development of coarsened exact matching [27] and the formalized theory connecting Gu and Rosenbaum's the distance, structure, and algorithm choices for matching methods [31].

Appendices

Appendix A Indicator variables

A.1 Protocol (Potential) method

A.1.1 Block (Group) effect

In Equation (3.1), we see the shorthand notation

$$Y = \beta_0 + \beta_{t3} z_{i,t3} + \beta_{t2} z_{i,t2} + \beta_{int} z_{i,int} + BLK_i + \varepsilon_i$$

for $i = 1, 2, \ldots, 32$ experimental units. Note that $s = 8$ in the original sheep dataset. Note also that in the *protocol* approach, we refer to the originally created groups as "blocks," whereas in the *potential* approach, we use the more general term, "groups," to reflect that later subgroups are not required to be either numerically balanced or mutually exclusive.

As each set of block (groups), take values $x_{i,sub} \in \{1, 2, \ldots, s\}$, where s represents the number of subgroups used for a particular dataset in a particular model, the shorthand BLK_i (GRP_i) actually represents

$$BLK_i = \beta_{sub} x_{i,sub} \quad \text{for} \quad sub = 1, 2, \ldots, s$$

where

$$BLK_i = \beta_{sub} x_{i,sub} = \beta_{sub=2} z_{i,sub=2} + \beta_{sub=3} z_{i,sub=3} + \cdots + \beta_{sub=s} z_{i,sub=s}$$

with

$$\begin{aligned}
x_{i,sub} = 1 &\iff z_{i,sub=2} = 0, z_{i,sub=3} = 0, \ldots, z_{i,sub=s} = 0 \\
x_{i,sub} = 2 &\iff z_{i,sub=2} = 1, z_{i,sub=3} = 0, \ldots, z_{i,sub=s} = 0 \\
x_{i,sub} = 3 &\iff z_{i,sub=2} = 0, z_{i,sub=3} = 1, \ldots, z_{i,sub=s} = 0 \\
&\vdots \\
x_{i,sub} = s &\iff z_{i,sub=2} = 0, z_{i,sub=3} = 0, \ldots, z_{i,sub=s} = 1
\end{aligned}$$

Here, $z_{i,sub=2}, z_{i,sub=3}, \ldots,$ and $z_{i,sub=s}$ are all binary indicators taking values in $\{0, 1\}$ so that

$\beta_{sub=2}$, $\beta_{sub=3}$, and $\beta_{sub=s}$ respectively represent the shift the mean response to account for the difference between $x_{i,sub} = 1$ to $x_{i,sub} = 2$, for the difference between $x_{i,sub} = 2$ to $x_{i,sub} = 3, \ldots,$ and for the difference between $x_{i,sub} = s - 1$ to $x_{i,sub} = s$. Thus, the shorthand BLK_i (GRP_i) can be considered a block (group) effect.

A.2 Researcher method

In Equation (3.7), we see the shorthand notation

$$Y = \beta_0 + \beta_T z_{i,T} + \beta_{ibcs} x_{i,ibcs} + \beta_{lmno} x_{i,lmno} + \beta_{geno} x_{i,geno} + \varepsilon_i$$

A.2.1 Lamb number

As the lamb number takes values $x_{i,lmno} \in \{1, 2, 3\}$, the shorthand $\beta_{lmno} x_{i,lmno}$ actually represents

$$\beta_{lmno} x_{i,lmno} = \beta_{lmno=2} z_{i,lmno=2} + \beta_{lmno=3} z_{i,lmno=3}$$

with

$$\begin{aligned}
x_{i,lmno} = 1 &\iff z_{i,lmno=2} = 0, z_{i,lmno=3} = 0 \\
x_{i,lmno} = 2 &\iff z_{i,lmno=2} = 1, z_{i,lmno=3} = 0 \\
x_{i,lmno} = 3 &\iff z_{i,lmno=2} = 0, z_{i,lmno=3} = 1
\end{aligned}$$

Here, $z_{i,lmno=2}$ and $z_{i,lmno=3}$ are each binary indicators taking values in $\{0, 1\}$ so that $\beta_{lmno=2}$ and $\beta_{lmno=3}$ respectively represent the shift the mean response to account for the difference between $x_{i,lmno} = 1$ to $x_{i,lmno} = 2$ and the difference between $x_{i,lmno} = 2$ to $x_{i,lmno} = 3$. Thus, the shorthand β_{lmno} can be considered a lamb number effect.

A.2.2 Genotype

As the genotype takes values $x_{i,geno} \in \{AA, AG, GG\}$, the shorthand $\beta_{geno} x_{i,geno}$ actually represents

$$\beta_{geno} x_{i,geno} = \beta_{AG} z_{i,AG} + \beta_{GG} z_{i,GG}$$

with

$$x_{i,geno} = AA \iff z_{i,AG} = 0, z_{i,GG} = 0$$
$$x_{i,geno} = AG \iff z_{i,AG} = 1, z_{i,GG} = 0$$
$$x_{i,geno} = GG \iff z_{i,AG} = 0, z_{i,GG} = 1$$

Here, $z_{i,AG}$ and $z_{i,GG}$ are each binary indicators taking values in $\{0, 1\}$ so that β_{AG} and β_{GG} respectively represent the shift the mean response to account for the difference between $x_{i,geno} = AA$ to $x_{i,geno} = AG$ and for the difference between $x_{i,geno} = AG$ to $x_{i,geno} = GG$. Thus, the shorthand β_{geno} can be considered a genotype effect.

A.2.3 Body condition score

As the initial BCS takes values $x_{i,ibcs} \in \{1, 2, 3, 4\}$, the shorthand $\beta_{ibcs}x_{i,ibcs}$ actually represents

$$\beta_{ibcs}x_{i,ibcs} = \beta_{ibcs=2}z_{i,ibcs=2} + \beta_{ibcs=3}z_{i,ibcs=3} + \beta_{ibcs=4}z_{i,ibcs=4}$$

with

$$x_{i,ibcs} = 1 \iff z_{i,ibcs=2} = 0, z_{i,ibcs=3} = 0, z_{i,ibcs=4} = 0$$
$$x_{i,ibcs} = 2 \iff z_{i,ibcs=2} = 2, z_{i,ibcs=3} = 0, z_{i,ibcs=4} = 0$$
$$x_{i,ibcs} = 3 \iff z_{i,ibcs=2} = 0, z_{i,ibcs=3} = 1, z_{i,ibcs=4} = 0$$
$$x_{i,ibcs} = 4 \iff z_{i,ibcs=2} = 0, z_{i,ibcs=3} = 0, z_{i,ibcs=4} = 1$$

Here, $z_{i,ibcs=2}$, $z_{i,ibcs=3}$, and $z_{i,ibcs=4}$ are each binary indicators taking values in $\{0, 1\}$ so that $\beta_{ibcs=2}$, $\beta_{ibcs=3}$, and $\beta_{ibcs=4}$ respectively represent the shift the mean response to account for the difference between $x_{i,ibcs} = 1$ to $x_{i,ibcs} = 2$, for the difference between $x_{i,ibcs} = 2$ to $x_{i,ibcs} = 3$, and for the difference between $x_{i,ibcs} = 3$ to $x_{i,ibcs} = 4$. Thus, the shorthand β_{ibcs} can be considered an initial BCS effect.

Appendix B ANOVA tables

Table 1: ANOVA tables for uterine weight, yUteWtg

These ANOVA tables provide partial (Type III) ANOVA results for the ewe response of uterine weight estimated on the original sheep dataset ("yUteWtg") for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yUteWtg: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	980652	980652	4.15	0.0544
block	7	1106896	158128	0.67	0.6953
tri2=toxic	1	184030	184030	0.78	0.3873
tri3*tri2	1	67912	67912	0.29	0.5974
Residuals	21	4958680	236128	NA	NA

yUteWtg: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	980652	980652	4.66	0.0391
Residuals	30	6317518	210584	NA	NA

yUteWtg: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	1187805	1187805	10.90	0.0027
lamb num	1	3078709	3078709	28.24	0.000
genotype	1	12419	12419	0.11	0.7383
initial BCS	1	2984	2984	0.03	0.8698
Residuals	27	2943285	109011	NA	NA

Table 2: ANOVA tables for yACarTotWT

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yACarTotWT: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	2485	2485	1.11	0.3045
block	7	31349	4478	2.00	0.1041
tri2=toxic	1	1613	1613	0.72	0.4060
tri3*tri2	1	4783	4783	2.13	0.1590
Residuals	21	47097	2243	NA	NA

yACarTotWT: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	2485	2485	0.88	0.3560
Residuals	30	84842	2828	NA	NA

yACarTotWT: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	3203	3203	1.16	0.2902
lambnum	1	4239	4239	1.54	0.2252
genotype	1	627	627	0.23	0.6369
initialBCS	1	7693	7693	2.80	0.1061
Residuals	27	74298	2752	NA	NA

Table 3: ANOVA tables for yACotTotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yACotTotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	10552	10552	0.86	0.3634
block	7	138101	19729	1.61	0.1862
tri2=toxic	1	3979	3979	0.33	0.5744
tri3*tri2	1	16858	16858	1.38	0.2534
Residuals	21	256735	12225	NA	NA

yACotTotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	10552	10552	0.76	0.3898
Residuals	30	415672	13856	NA	NA

yACotTotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	13981	13981	1.05	0.3148
lamb num	1	11947	11947	0.90	0.3521
genotype	1	595	595	0.04	0.8343
initial BCS	1	51698	51698	3.88	0.0592
Residuals	27	359816	13327	NA	NA

Table 4: ANOVA tables for yATotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yATotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	23279	23279	0.95	0.3419
block	7	296406	42344	1.72	0.1585
tri2=toxic	1	10658	10658	0.43	0.5177
tri3*tri2	1	39600	39600	1.61	0.2186
Residuals	21	517019	24620	NA	NA

yATotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	23279	23279	0.81	0.3757
Residuals	30	863683	28789	NA	NA

yATotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	30568	30568	1.10	0.3037
lamb num	1	30420	30420	1.09	0.3048
genotype	1	2443	2443	0.09	0.7692
initial BCS	1	99277	99277	3.57	0.0696
Residuals	27	750724	27805	NA	NA

Table 5: ANOVA tables for yBCarTotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yBCarTotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	11207	11207	5.36	0.0308
block	7	41788	5970	2.86	0.0293
tri2=toxic	1	3	3	0.00	0.9697
tri3*tri2	1	1231	1231	0.59	0.4514
Residuals	21	43902	2091	NA	NA

yBCarTotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	11207	11207	3.87	0.0585
Residuals	30	86924	2897	NA	NA

yBCarTotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	11084	11084	3.60	0.0685
lamb num	1	992	992	0.32	0.5749
genotype	1	12	12	0.00	0.9510
initial BCS	1	1941	1941	0.63	0.4341
Residuals	27	83104	3078	NA	NA

Table 6: ANOVA tables for yBCotTotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yBCotTotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	130959	130959	6.15	0.0217
block	7	379790	54256	2.55	0.0456
tri2=toxic	1	1123	1123	0.05	0.8205
tri3*tri2	1	8470	8470	0.40	0.5349
Residuals	21	446823	21277	NA	NA

yBCotTotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	130959	130959	4.70	0.0383
Residuals	30	836206	27874	NA	NA

yBCotTotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	128929	128929	4.58	0.0415
lamb num	1	17097	17097	0.61	0.4425
genotype	1	48	48	0.00	0.9673
initial BCS	1	42797	42797	1.52	0.2281
Residuals	27	759689	28137	NA	NA

Table 7: ANOVA tables for yBTotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yBTotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	218786	218786	6.23	0.0210
block	7	664741	94963	2.70	0.0365
tri2=toxic	1	1008	1008	0.03	0.8671
tri3*tri2	1	16159	16159	0.46	0.5050
Residuals	21	737687	35128	NA	NA

yBTotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	218786	218786	4.62	0.0397
Residuals	30	1419596	47320	NA	NA

yBTotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	215620	215620	4.46	0.0441
lamb num	1	26327	26327	0.54	0.4669
genotype	1	108	108	0.00	0.9627
initial BCS	1	62965	62965	1.30	0.2638
Residuals	27	1305123	48338	NA	NA

Table 8: ANOVA tables for yCCarTotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yCCarTotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	837	837	0.35	0.5584
block	7	10788	1541	0.65	0.7097
tri2=toxic	1	220	220	0.09	0.7633
tri3*tri2	1	4652	4652	1.97	0.1756
Residuals	21	49716	2367	NA	NA

yCCarTotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	837	837	0.38	0.5400
Residuals	30	65376	2179	NA	NA

yCCarTotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	1123	1123	0.51	0.4831
lamb num	1	1181	1181	0.53	0.4721
genotype	1	198	198	0.09	0.7675
initial BCS	1	4915	4915	2.21	0.1484
Residuals	27	59938	2220	NA	NA

Table 9: ANOVA tables for yCCotTotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yCCotTotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	2510	2510	0.07	0.7961
block	7	141203	20172	0.55	0.7868
tri2=toxic	1	153	153	0.00	0.9491
tri3*tri2	1	78418	78418	2.14	0.1583
Residuals	21	769628	36649	NA	NA

yCCotTotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	2510	2510	0.08	0.7846
Residuals	30	989402	32980	NA	NA

yCCotTotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	5181	5181	0.16	0.6951
lamb num	1	26357	26357	0.80	0.3795
genotype	1	911	911	0.03	0.8693
initial BCS	1	87163	87163	2.64	0.1158
Residuals	27	891337	33012	NA	NA

Table 10: ANOVA tables for yCTotWt

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yCTotWt: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	6246	6246	0.11	0.7441
block	7	227278	32468	0.57	0.7729
tri2=toxic	1	6	6	0.00	0.9918
tri3*tri2	1	121271	121271	2.12	0.1598
Residuals	21	1198806	57086	NA	NA

yCTotWt: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	6246	6246	0.12	0.7303
Residuals	30	1547361	51579	NA	NA

yCTotWt: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	11126	11126	0.21	0.6467
lamb num	1	38695	38695	0.75	0.3949
genotype	1	1959	1959	0.04	0.8472
initial BCS	1	133473	133473	2.58	0.1200
Residuals	27	1397925	51775	NA	NA

Table 11: ANOVA tables for yCarTot

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yCarTot: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	13088	13088	14.96	0.0009
block	7	9975	1425	1.63	0.1819
tri2=toxic	1	3283	3283	3.75	0.0663
tri3*tri2	1	634	634	0.72	0.4042
Residuals	21	18369	875	NA	NA

yCarTot: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	13088	13088	12.17	0.0015
Residuals	30	32261	1075	NA	NA

yCarTot: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	13203	13203	11.64	0.0020
lamb num	1	910	910	0.80	0.3784
genotype	1	1	1	0.00	0.9809
initial BCS	1	371	371	0.33	0.5722
Residuals	27	30624	1134	NA	NA

Table 12: ANOVA tables for yCotTot

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yCotTot: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	195349	195349	8.81	0.0073
block	7	45534	6505	0.29	0.9490
tri2=toxic	1	5058	5058	0.23	0.6379
tri3*tri2	1	16657	16657	0.75	0.3959
Residuals	21	465726	22177	NA	NA

yCotTot: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	195349	195349	11.00	0.0024
Residuals	30	532974	17766	NA	NA

yCotTot: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	209034	209034	13.69	0.0010
lamb num	1	99966	99966	6.55	0.0164
genotype	1	10958	10958	0.72	0.4044
initial BCS	1	1209	1209	0.08	0.7806
Residuals	27	412328	15271	NA	NA

Table 13: ANOVA tables for yPlacTot

These ANOVA tables provide partial (Type III) ANOVA results for the models Equation (3.1) (`protocol`), Equation (3.3) (`t3only`), and Equation (3.7) (`allcov`), respectively.

yPlacTot: `protocol`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	309564	309564	12.54	0.0019
block	7	65699	9386	0.38	0.9036
tri2=toxic	1	16491	16491	0.67	0.4229
tri3*tri2	1	10792	10792	0.44	0.5157
Residuals	21	518387	24685	NA	NA

yPlacTot: `t3only`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	309564	309564	15.19	0.0005
Residuals	30	611368	20379	NA	NA

yPlacTot: `allcov`

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	327307	327307	18.94	0.0002
lamb num	1	119949	119949	6.94	0.0138
genotype	1	11129	11129	0.64	0.4292
initial BCS	1	240	240	0.01	0.9070
Residuals	27	466502	17278	NA	NA

Appendix C Boxplots

Each boxplot contains the the observed distribution of the estimated difference in means for the third trimester indicator Z_{t3}, or the third trimester effect, along with the observed standard error of every term. Estimates and their standard errors are recorded for every model and on every simulated dataset in additional the original sheep dataset. Boxplots show the observed distribution of each quantity for each model type, labeled along the horizontal axis.

Each boxplot figure compares model performance across all available datasets for a different response variable, with models along the horizontal axis. Model performance is reported as each dataset's third trimester effect (bottom, denoted "EST") and its corresponding standard error, or the difference in means and the standard deviation of the difference in means (top, denoted "SE"). The boxplots for EST show connected means for each model within each response.

We can thus use the boxplots to compare the observed distributions of the estimate and standard error between models within each response. Within every response, we overwhelmingly observe that the relative model behavior is similar for these measures of sheep fetal development. Through the above sections, each ewe's uterine weight, or "yUteWtg',' is used as an illustrative example, as the same model trends persist across response variables.

Figure 1: Boxplots for uterine weight, yUteWtg

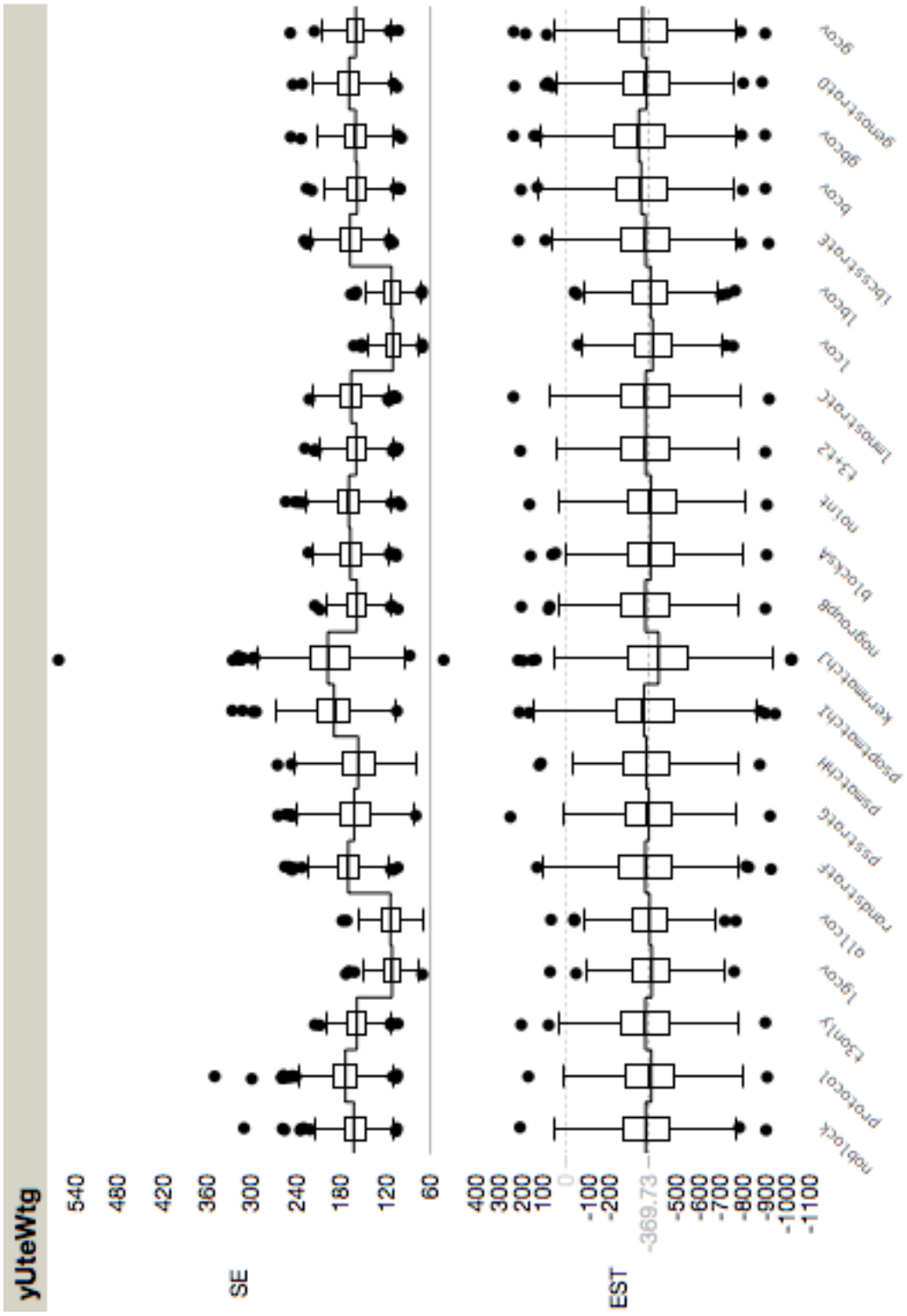


Figure 2: Boxplots for yACarTotWT

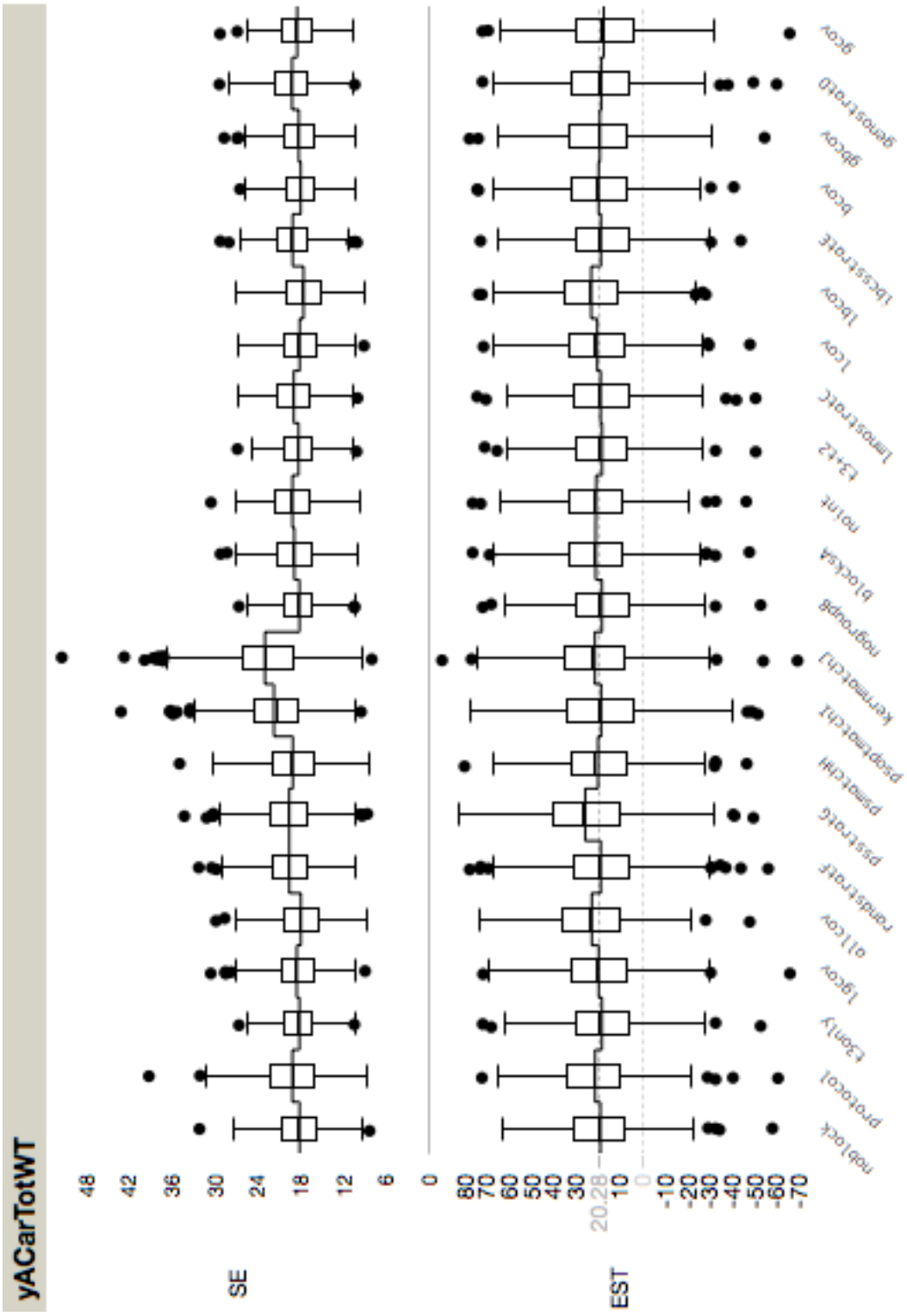


Figure 3: Boxplots for yACotTotWT

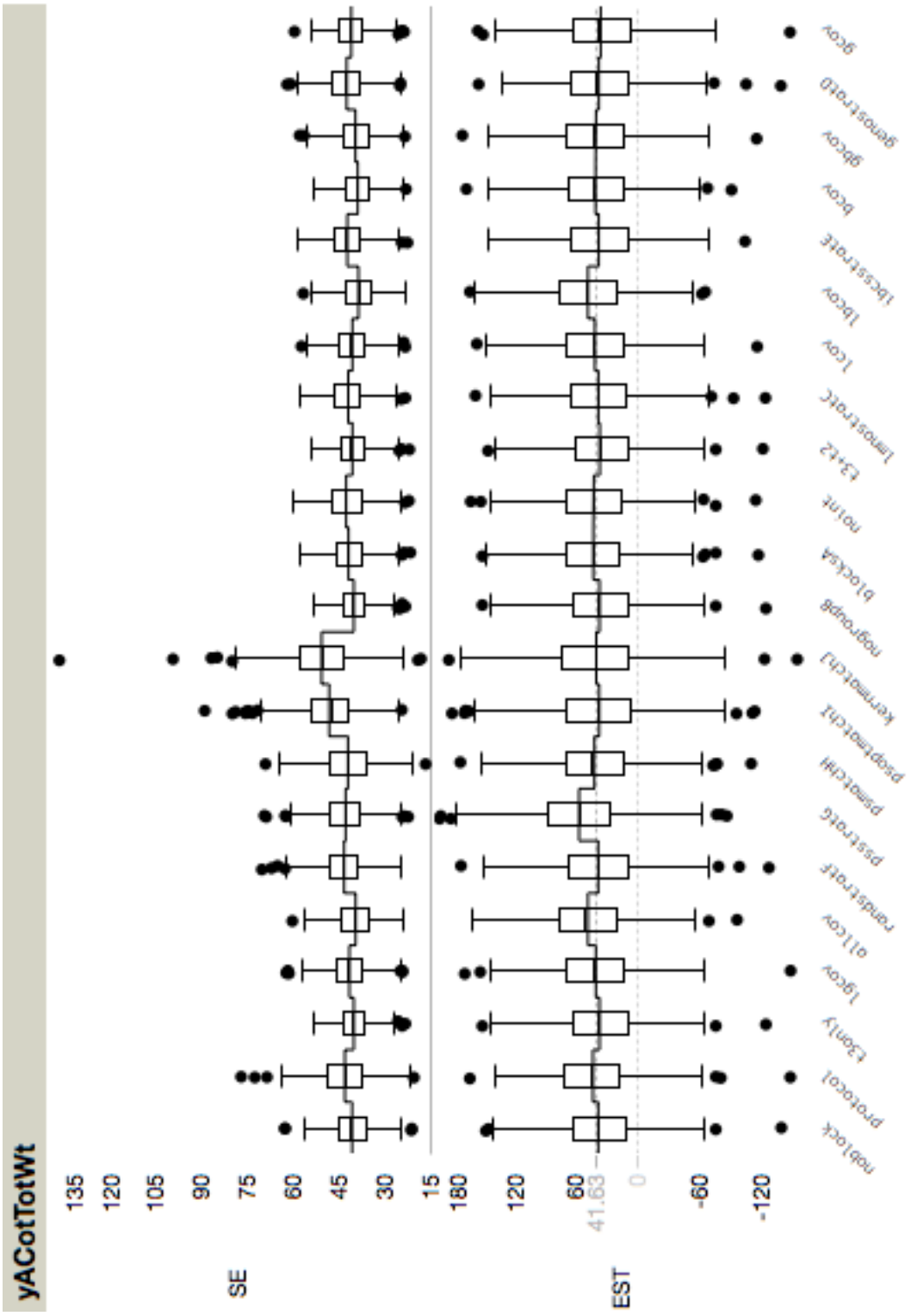


Figure 4: Boxplots for yATotWt

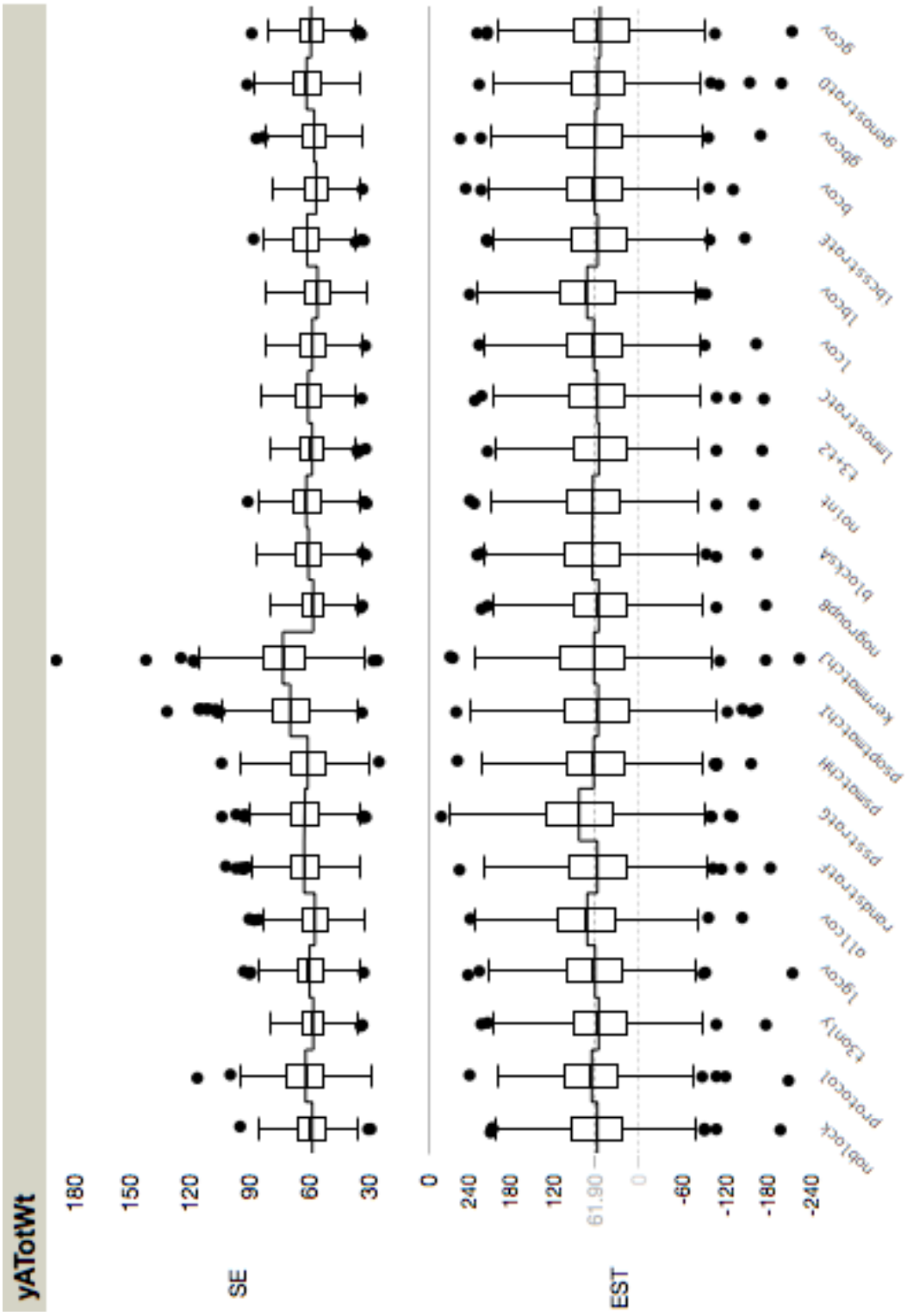


Figure 5: Boxplots for yBCarTotWT

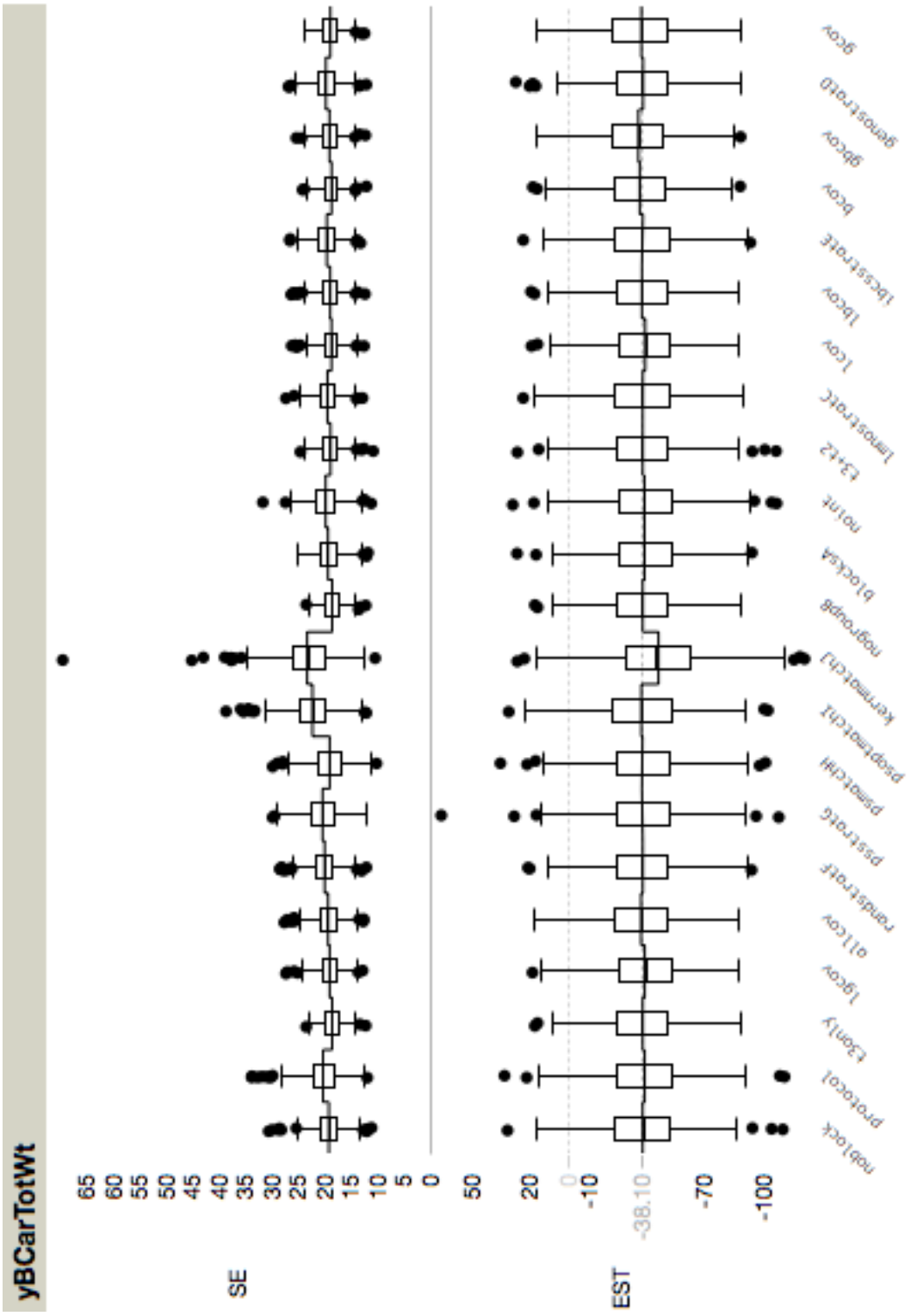


Figure 6: Boxplots for yUteWtg

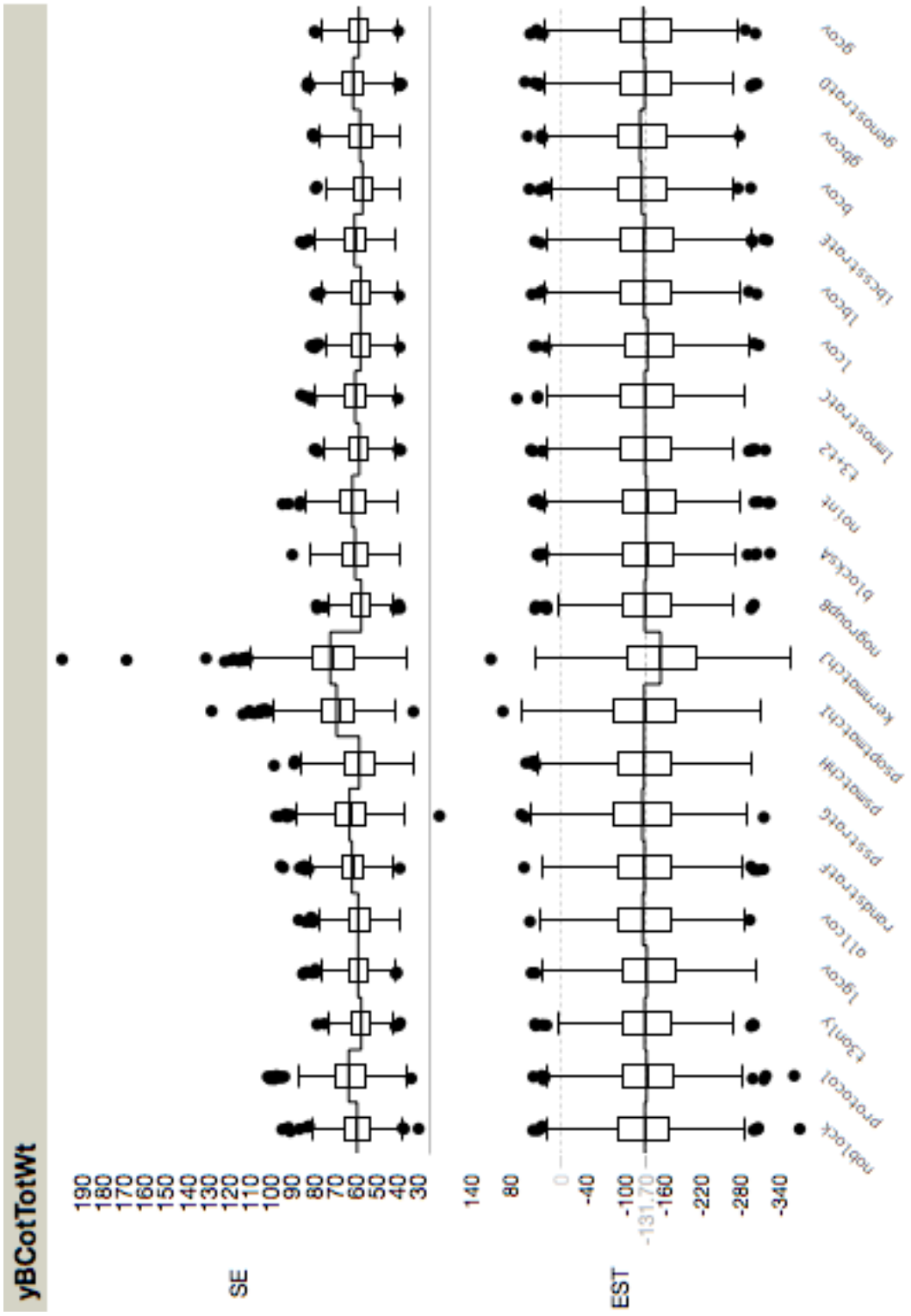


Figure 7: Boxplots for yBTotWt

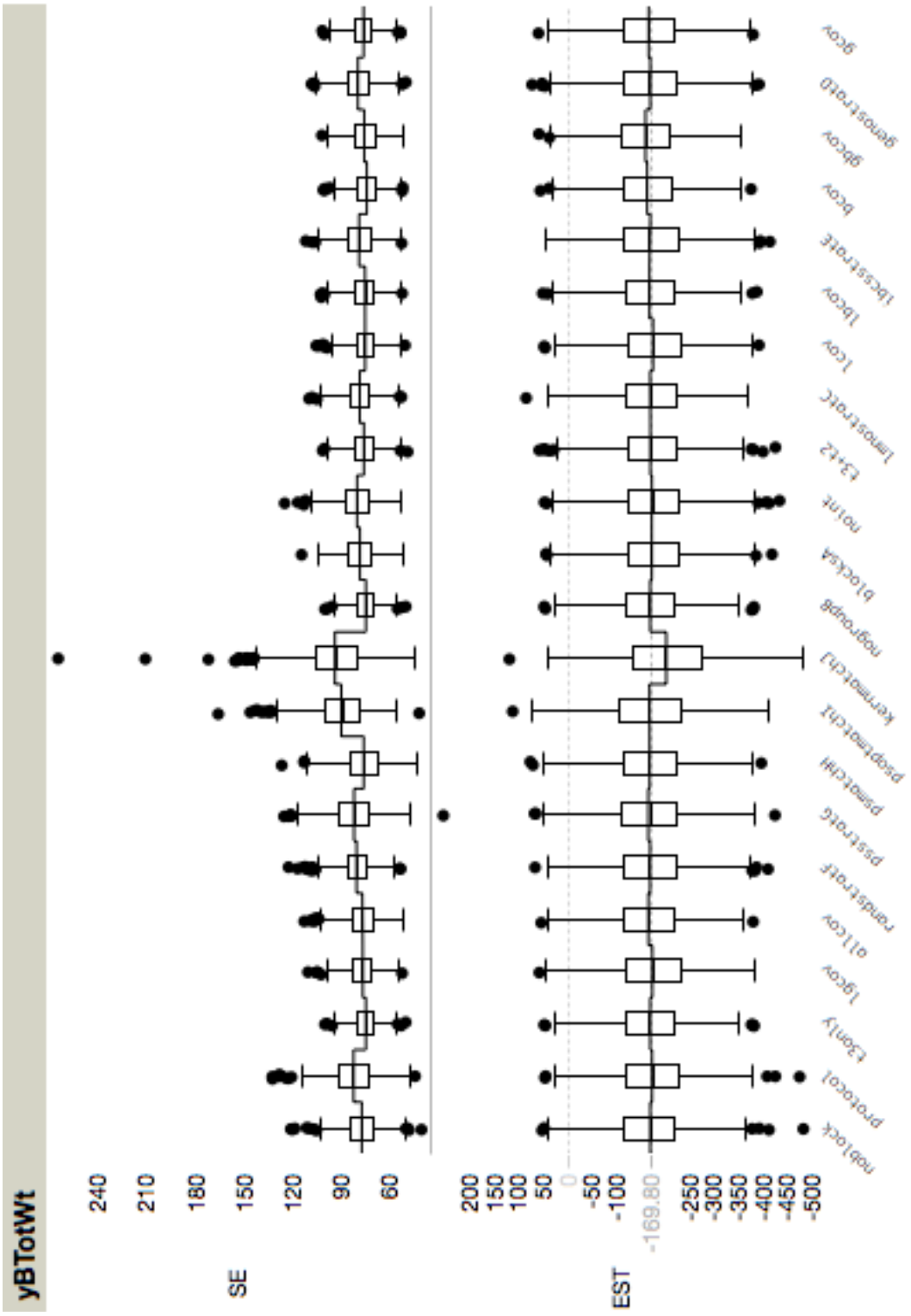


Figure 8: Boxplots for yCCarTotWT

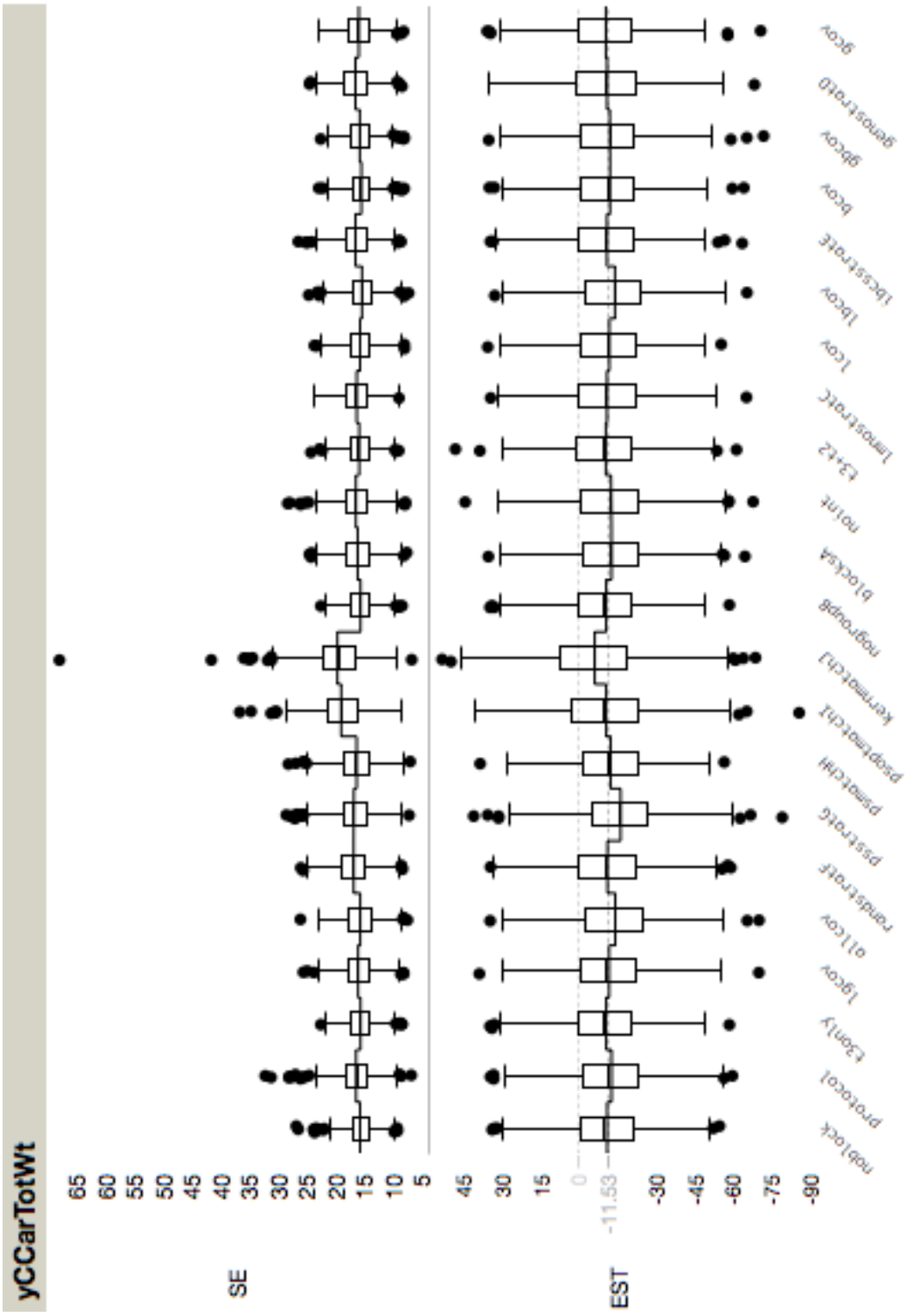


Figure 9: Boxplots for yUteWtg

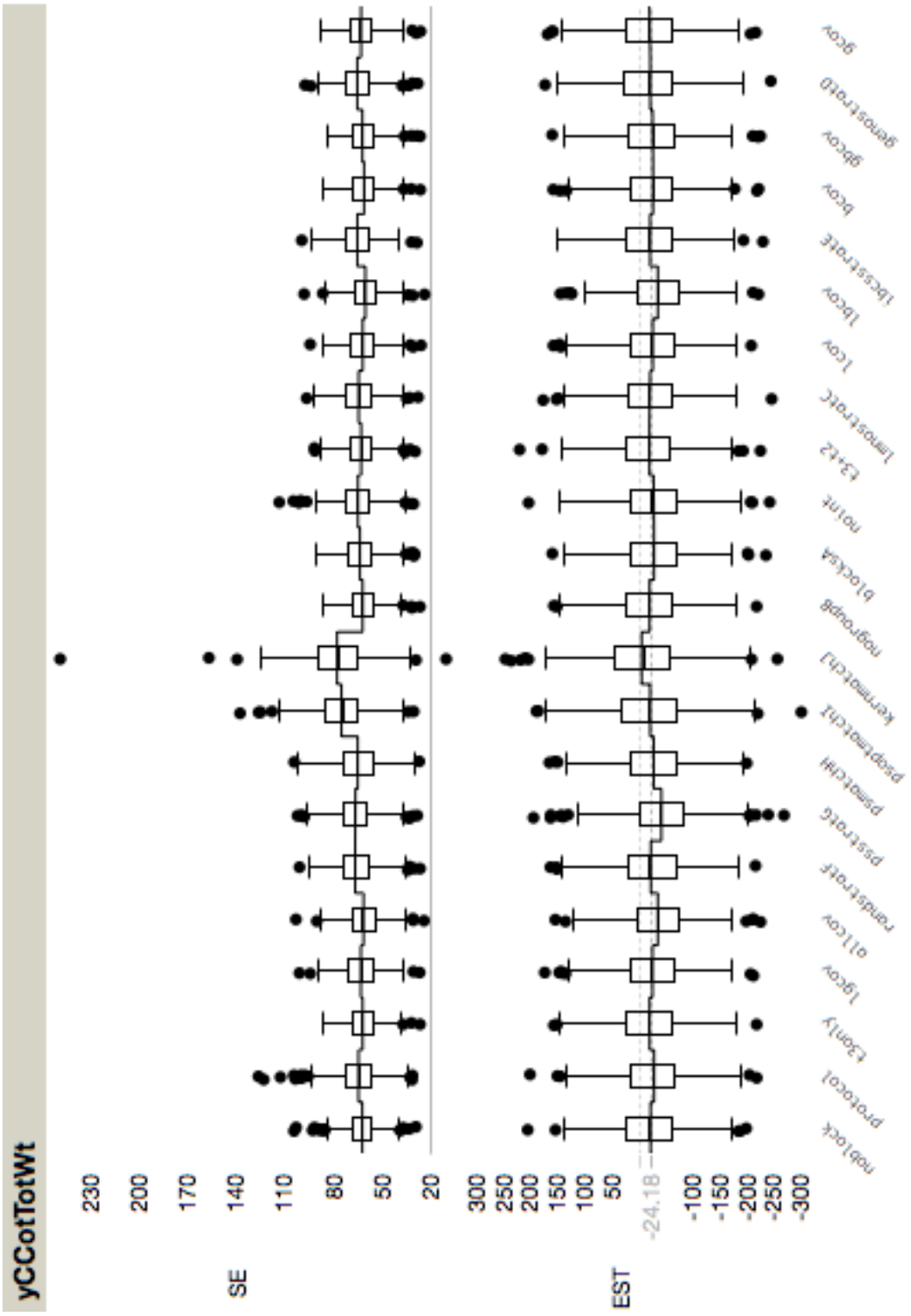


Figure 10: Boxplots for yCTotWt

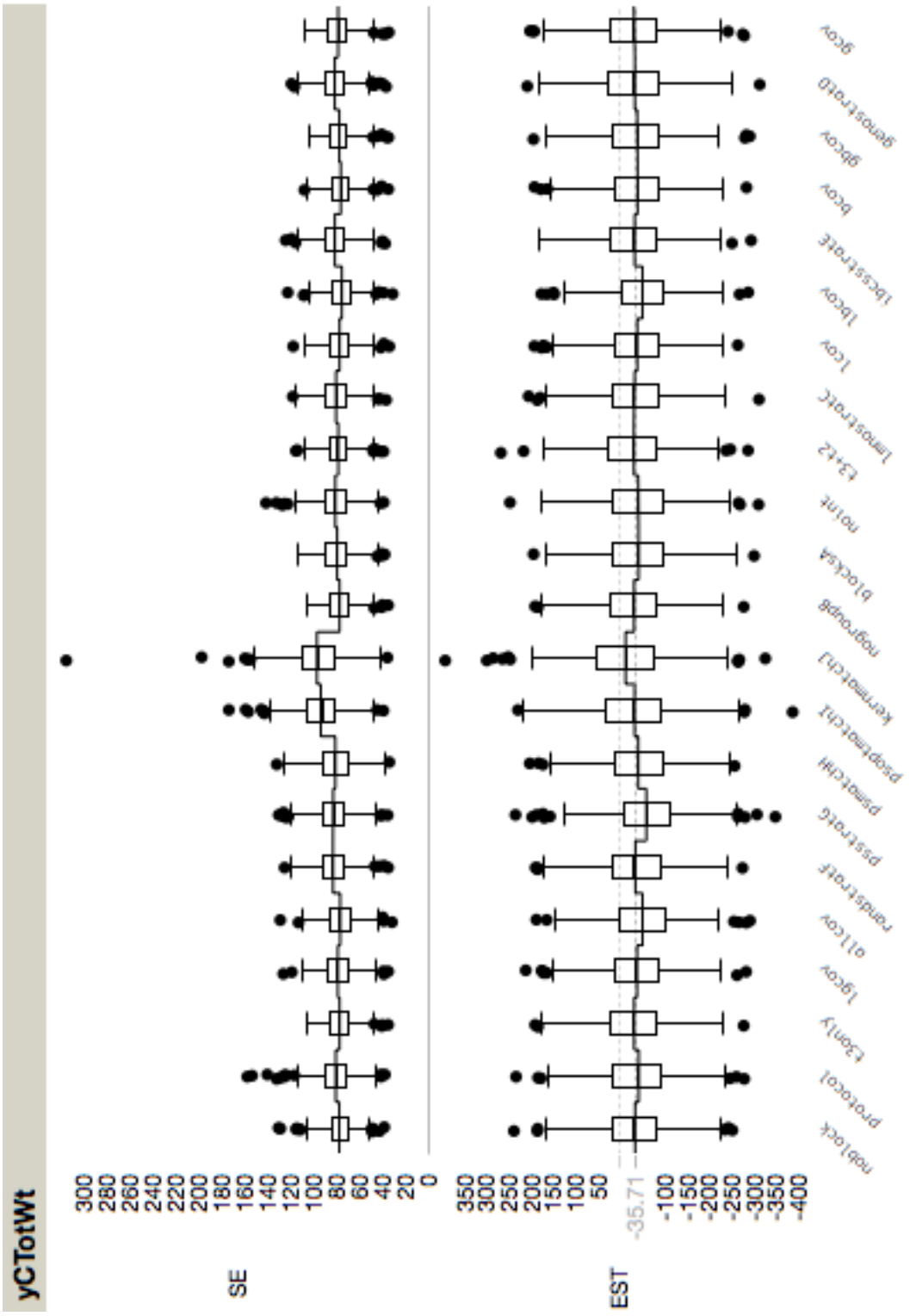


Figure 11: Boxplots for yCarTot

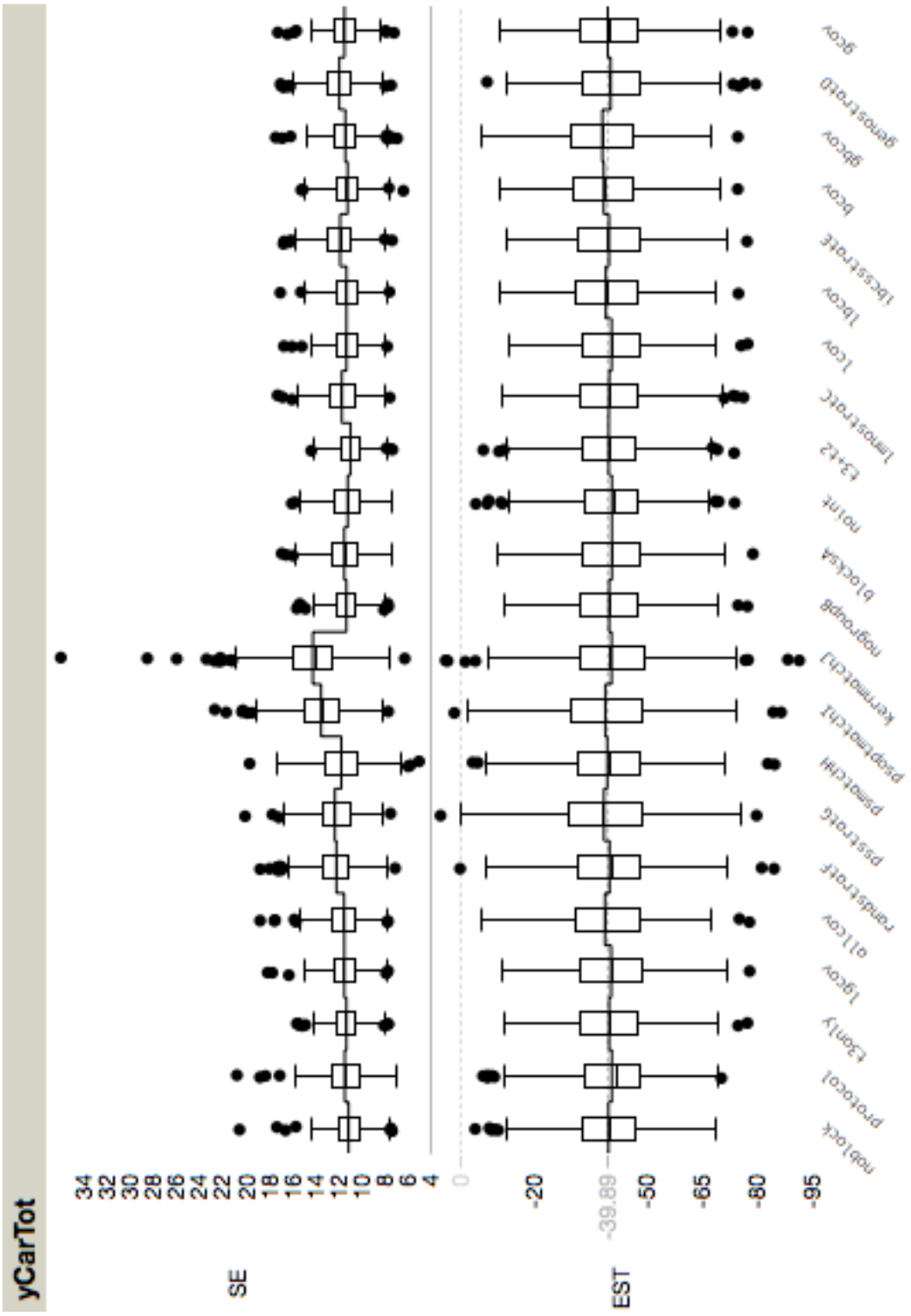


Figure 12: Boxplots for yCotTot

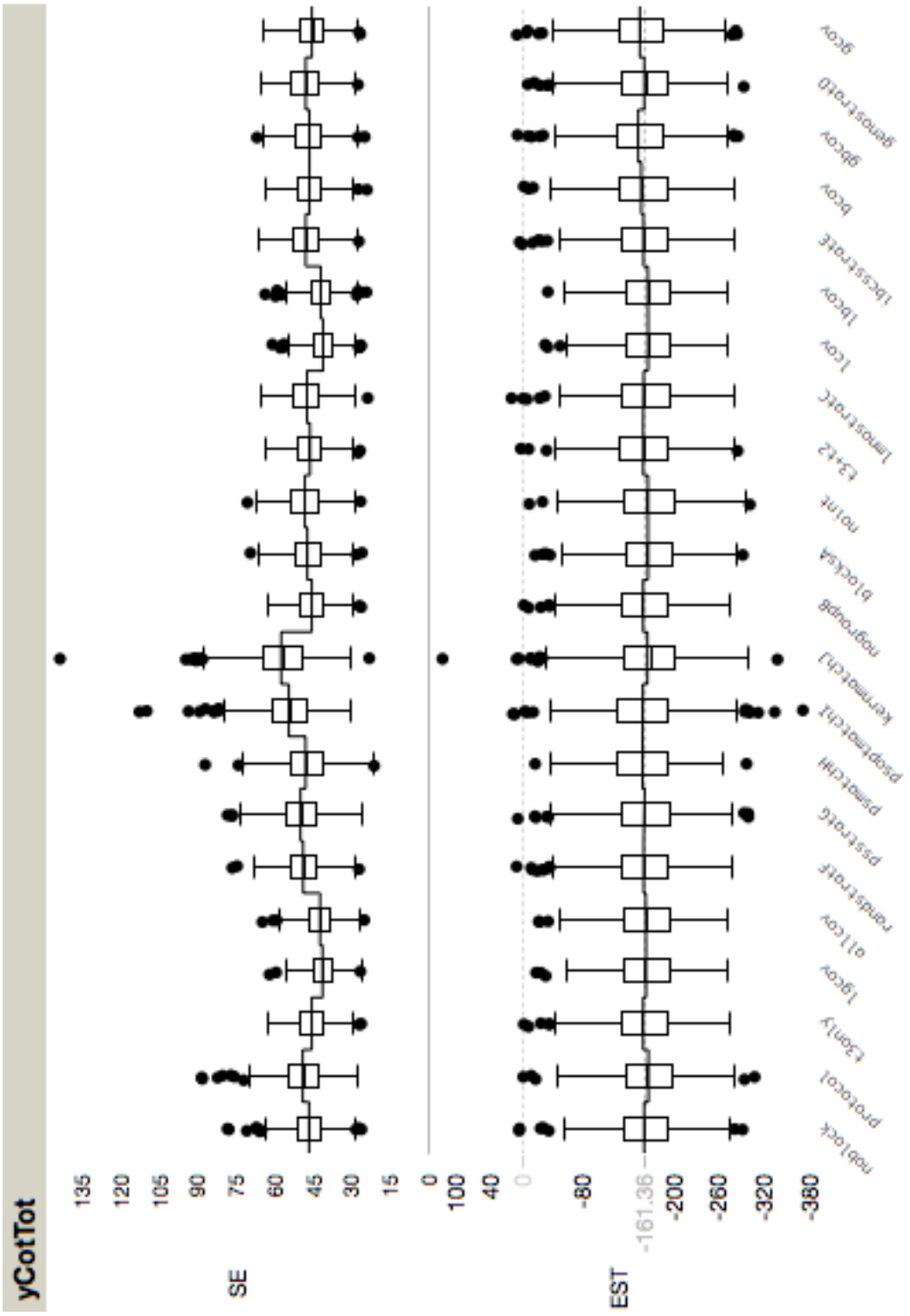
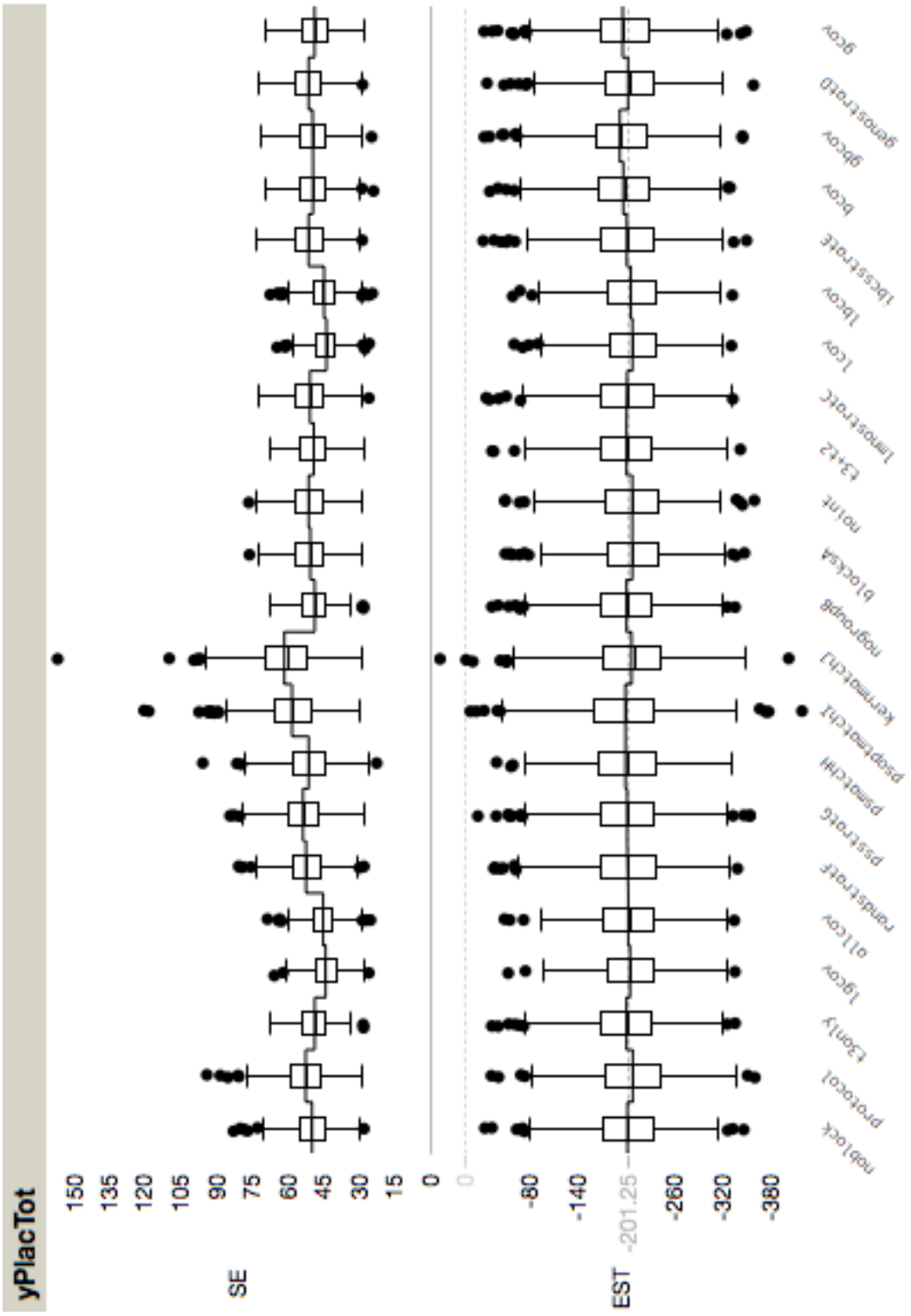


Figure 13: Boxplots for yPlacTot



Appendix D Models summary

This appendix summarizes the intent of all twenty two (22) linear models considered. The models highlighted in the methods are in listed in **bold**.
For the protocol approach, there is only the one model.

1. In Equation (3.1) (`protocol`), the researcher includes the third trimester effect, the second trimester effect, the interaction effect, and the block effect.

For the researcher approach, there are eleven (11) models, of which we highlighted three (3) in the methods. Recall that lamb number, genotype, and initial BCS are included as continuous covariates, per the researchers' choice.

1. In Equation (3.2) (`noblock`), the researcher includes the third trimester effect, the second trimester effect, and the interaction effect, dropping only the block effect from the `protocol` model.

2. In Equation (3.4) (`noint`), the researcher includes the third trimester effect, the second trimester effect, and the block effect, dropping only the interaction effect from the `protocol` model.

3. In Equation (3.5) (`t3+t2`), the researcher includes the third trimester effect and the second trimester effect, dropping the block and interaction effects from the `protocol` model.

4. **In Equation** (3.3) **(`t3only`), the researcher includes only the third trimester effect, dropping the block, second trimester, and interaction effects from the** `protocol` **model.**

5. In Equation (3.8) (`lcov`), the researcher includes only the lamb number, which is discrete and has three levels, as a continuous covariate.

6. In Equation (3.9) (`gcov`), the researcher includes only the genotype, which is nominal and has three levels, as a continuous covariate.

7. In Equation (3.10) (`bcov`), the researcher includes only the initial BCS, which is discrete and has four levels, as a continuous covariate.

8. **In Equation** (3.6) **(`lgcov`), the researcher includes lamb number and genotype; this is the final researcher model.**

9. In Equation (3.11) (`lbcov`), the researcher includes lamb number and initial BCS.

10. In Equation (3.12) (`gbcov`), the researcher includes genotype and initial BCS.

11. **In Equation** (3.7) **(`allcov`), the researcher includes lamb number, genotype, and initial BCS.**

For the potential approach, there are ten (10) models, of which we highlighted three (3) in the methods.

1. In Equation (3.16) (`blocksA`) we use eight blocks assigned by the experiment as groups.

2. In Equation (3.17) (`nogroupB`) we use no groups, so the estimate of the mean responses are identical to the (`t3only`) model.

3. In Equation (3.18) (`lmnostratC`) we use five groups created from stratifying on the continuous version of lamb number.

4. In Equation (3.19) (`genostratD`) we use five groups created from stratifying on the continuous version of genotype.

5. In Equation (3.20) (`ibcsstratE`) we use five groups created from stratifying on the continuous version of initial BCS.

6. In Equation (3.21) (`randstratF`) we randomly assign each experimental unit to one of five groups; this similar to the creation of the simulated datasets.

7. **In Equation** (3.14) **(`psstratG`) we use groups based on optimal matching of experimental units using a propensity score distance.**

8. **In Equation** (3.13) **(`psmatchH`) we form groups by attempting to collapse the groups from `psstratG` into five subgroups.**

9. In Equation (3.22) (`psoptmatchI`) we use groups based on optimal matching of experimental units using a propensity score distance.

10. **In Equation** (3.15) **(**`kernmatchJ`**) we use groups based on optimal matching of experimental units using a kernel distance.**

For completeness, all twenty two equation numbers and model names are summarized in Table 14, sorted by equation number and in Table 15, sorted by the boxplot display order.

Table 14: Models list by equation number

Equation (3.1)	(`protocol`)
Equation (3.2)	(`noblock`)
Equation (3.3)	(`t3only`)
Equation (3.4)	(`noint`)
Equation (3.5)	(`t3+t2`)
Equation (3.6)	(`lgcov`)
Equation (3.7)	(`allcov`)
Equation (3.8)	(`lcov`)
Equation (3.9)	(`gcov`)
Equation (3.10)	(`bcov`)
Equation (3.11)	(`lbcov`)
Equation (3.12)	(`gbcov`)
Equation (3.13)	(`psmatchH`)
Equation (3.14)	(`psstratG`)
Equation (3.15)	(`kernmatchJ`)
Equation (3.16)	(`blocksA`)
Equation (3.17)	(`nogroupB`)
Equation (3.18)	(`lmnostratC`)
Equation (3.19)	(`genostratD`)
Equation (3.20)	(`ibcsstratE`)
Equation (3.21)	(`randstratF`)
Equation (3.22)	(`psoptmatchI`)

Table 15: Models list by boxplot display order

Equation (3.2)	(`noblock`)
Equation (3.1)	(`protocol`)
Equation (3.3)	(`t3only`)
Equation (3.6)	(`lgcov`)
Equation (3.7)	(`allcov`)
Equation (3.21)	(`randstratF`)
Equation (3.14)	(`psstratG`)
Equation (3.13)	(`psmatchH`)
Equation (3.22)	(`psoptmatchI`)
Equation (3.15)	(`kernmatchJ`)
Equation (3.17)	(`nogroupB`)
Equation (3.16)	(`blocksA`)
Equation (3.4)	(`noint`)
Equation (3.5)	(`t3+t2`)
Equation (3.18)	(`lmnostratC`)
Equation (3.8)	(`lcov`)
Equation (3.11)	(`lbcov`)
Equation (3.20)	(`ibcsstratE`)
Equation (3.10)	(`bcov`)
Equation (3.12)	(`gbcov`)
Equation (3.19)	(`genostratD`)
Equation (3.9)	(`gcov`)

Appendix E Covariate continuity

None of the post-hoc covariates of lamb number, genotype, or initial BCS are continuous. Lamb number is a integer taking values of 1, 2, or 3; genotype is a nominal label taking values of *AA, AG, or GG*; and initial BCS is a integer taking values of 1, 2, 3, or 4. To include discrete values in a linear covariate adjustment, we typically use a series of indicator variables, as in Appendix A.

Throughout, we follow the researcher practice of using the continuous version of these covariates. To achieve this, the researchers ranked the nominal genotype levels, ordering the three genotypes from least to most resilient to the effects of infected fescue on sheep fetal development. Consider , boxplots summarizing the difference in means and its standard error for uterine weight across all simulated datasets, estimated on the original dataset for all models, shown in Figure 14. Lamb number, genotype, and initial BCS are included as continuous in the researcher approach in the top part of the figure and are included as discrete in the bottom part of the figure.

We are confident that although consider the covariates as continuous random variables is an approximation, it does not obscure the methods' relative abilities to estimate the the third trimester effect and its corresponding standard error. Also, this practice has advantages in interpretation. The coefficients association with lamb number, genotype, and initial BCS effects are the average change in the mean as each covariate increases by one unit. Finally, this practice is convenient to compare our three approaches, as kernel functions, being defined only continuous spaces, must use continuous approximations of discrete variables. That is, for the logistic regression used to estimate each ewe's propensity score and the linear covariate adjustment of the researcher approach, using continuous covariates to approximate their exact, discrete versions allows better comparison to the kernel methods.

When discrete covariates are treated not as indicator variables, but treated as though they were a single, continuous term, this increases the degrees of freedom available for estimating a linear model's error term. Consider Table 16, the partial ANOVA tables for the three covariate model, Equation (3.7), estimated on each with lamb number, genotype, and initial BCS included as continuous and as discrete covariates, respectively. The approximation does not change the overall result of the (researchers' final) analysis; the third trimester effect is still "significant" in the partial ANOVA table, while each covariate is distinctly "not significant."

We do note that this approximation does reduce the p-value of the (partial ANOVA) hy-

pothesis tests of each of these covariates (see Table 16), which is a direct result of increasing the numerator degrees of freedom of the F-statistic. More heuristically, it is a result of decreasing the amount of information used to estimate the effect of each covariate because we are estimating average covariate effects in place of an effect for each level of the discrete covariates.

Table 16: Continuous versus discrete covariates

The partial ANOVA table for the three covariate model, Equation (3.7), using yUteWtg, estimated on the original dataset with lamb number, genotype, and initial BCS included as CONTINUOUS (top) and DISCRETE (bottom) covariates.

yUteWtg: `allcov`, covariates CONTINUOUS

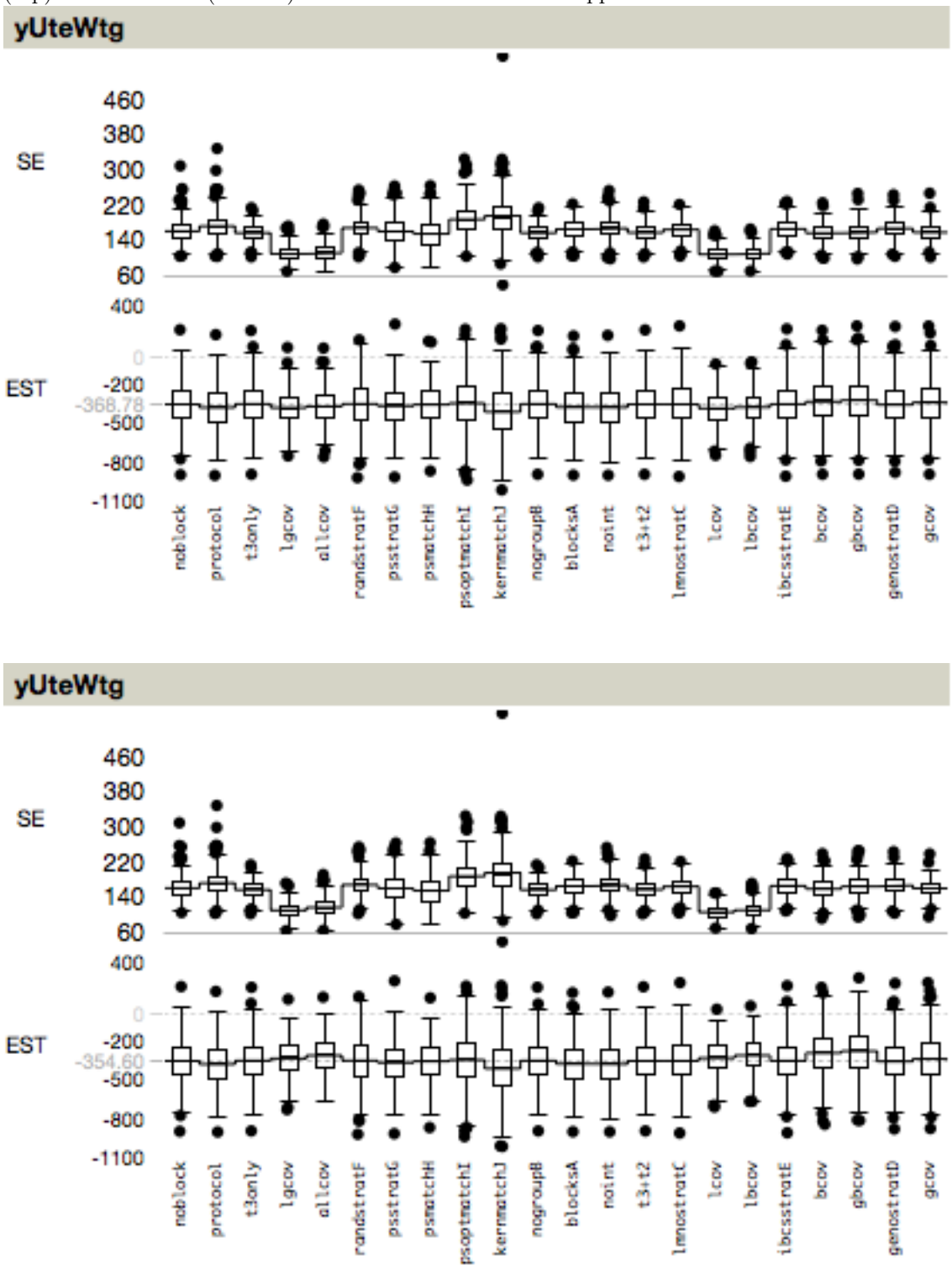
Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	1187805	1187805	10.90	0.0027
lamb num	1	3078709	3078709	28.24	0.0000
genotype	1	12419	12419	0.11	0.7383
initial BCS	1	2984	2984	0.03	0.8698
Residuals	27	2943285	109011	NA	NA

yUteWtg: `allcov`, covariates DISCRETE

Terms	Df	SumSq	MeanSq	Fstat	Pval
tri3=toxic	1	710358	710358	6.48	0.0181
lamb num	2	3113547	1556773	14.19	0.0001
genotype	2	152	76	0.00	0.9993
initial BCS	3	45829	15276	0.14	0.9355
Residuals	23	2523193	109704	NA	NA

Figure 14: Boxplots comparing continuous versus discrete covariate use

Boxplots summarizing the difference in means and its standard error using yUteWtg across all simulated datasets for all models. Lamb number, genotype, and initial BCS included as CONTINUOUS (top) and DISCRETE (bottom) covariates in the researcher approach.



Bibliography

[1] Peter C Austin. A comparison of 12 algorithms for matching on the propensity score. *Statistics in Medicine*, 33(6):1057–1069, 2014.

[2] TA Bancroft and Chien-Pai Han. Inference based on conditional specification: a note and a bibliography. *International Statistical Review/Revue Internationale de Statistique*, pages 117–127, 1977.

[3] TA Bancroft and Chien-Pai Han. 13 inference based on conditionally specified anova models incorporating preliminary testing. *Handbook of Statistics*, 1:407–441, 1980.

[4] Theodore Alfonso Bancroft. On biases in estimation due to the use of preliminary tests of significance. *The Annals of Mathematical Statistics*, 15(2):190–204, 1944.

[5] Theodore Alfonso Bancroft. Analysis and inference for incompletely specified models involving the use of preliminary test (s) of significance. *Biometrics*, 20(3):427–442, 1964.

[6] Dimitris Bertsimas, Mac Johnson, and Nathan Kallus. The power of optimization over randomization in designing experiments involving small samples. *Operations Research*, 63(4):868–876, 2015.

[7] Helen Bozivich and TA Bancroft. Analysis of variance: Preliminary tests, pooling, and linear models; tests of significance and pooling procedures for certain incompletely specified models. Technical report, Iowa State University Ames Statistical Lab, 1956.

[8] JL Britt, MF Miller Jr, BM Koch, NM Long, SL Pratt, JG Andrae, and SK Duckett. 059 effects of ergot alkaloids during mid-to-late gestation on uteroplacental sufficiency and fetal growth. *Journal of Animal Science*, 95(suppl_1):29–29, 2016.

[9] William G Cochran. *Planning and Analysis of Observational Studies*. John Wiley & Sons, 1983.

[10] María de los Angeles Resa and Jose R Zubizarreta. Evaluation of subset matching methods and forms of covariate balance. *Statistics in Medicine*, 35(27):4961–4979, 2016.

[11] Markus C Elze, John Gregson, Usman Baber, Elizabeth Williamson, Samantha Sartori, Roxana Mehran, Melissa Nichols, Gregg W Stone, and Stuart J Pocock. Comparison of propensity score methods and covariate adjustment: evaluation in 4 cardiovascular studies. *Journal of the American College of Cardiology*, 69(3):345–357, 2017.

[12] Leonard S Feldt. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 23(4):335–353, 1958.

[13] Jessica M Franklin, Jeremy A Rassen, Diana Ackermann, Dorothee B Bartels, and Sebastian Schneeweiss. Metrics for covariate balance in cohort studies of causal effects. *Statistics in Medicine*, 33(10):1685–1699, 2014.

[14] Arthur Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, page 16, 2013.

[15] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International Conference on Algorithmic Learning Theory*, pages 63–77. Springer, 2005.

[16] Xing Sam Gu and Paul R Rosenbaum. Comparison of multivariate matching methods: Structures, distances, and algorithms. *Journal of Computational and Graphical Statistics*, 2(4):405–420, 1993.

[17] Shenyang Guo and Mark W Fraser. *Propensity Score Analysis*, volume 12. Sage, 2014.

[18] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[19] BB Hansen. Optmatch (r package optmatch). *R News*, 7:18–24, 2007.

[20] Ben B Hansen and Stephanie Olsen Klopfer. Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics*, 15(3):609–627, 2006.

[21] Chad Hazlett. Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects. 2016.

[22] Chad Hazlett and Jens (Maintainer) Hainmueller. Package krls. 2017.

[23] Keisuke Hirano and Guido W Imbens. Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization. *Health Services and Outcomes Research Methodology*, 2(3-4):259–278, 2001.

[24] Daniel Ho, Kosuke Imai, Gary King, Elizabeth Stuart, and Alex Whitworth. Package matchit, 2018.

[25] Daniel E Ho, Kosuke Imai, Gary King, Elizabeth A Stuart, et al. Matchit: nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011.

[26] William M Holmes. *Using Propensity Scores in Quasi-experimental Designs*. Sage Publications, 2013.

[27] Stefano M Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

[28] Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):243–263, 2014.

[29] Jessica N Jacovidis. *Evaluating the performance of propensity score matching methods: a simulation study*. PhD thesis, James Madison University, 2017.

[30] Nathan Kallus. A framework for optimal matching for causal inference. *arXiv preprint arXiv:1606.05188*, 2016.

[31] Nathan Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

[32] Nathan Kallus. Optimal a priori balance in the design of controlled experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):85–112, 2018.

[33] Gary King, Richard Nielsen, Carter Coberley, James E Pope, and Aaron Wells. Comparative effectiveness of matching methods for causal inference. *Unpublished manuscript*, 15:41, 2011.

[34] Yevgeniya N Kleyman. *Testing for Covariate Balance in Comparative Studies.* PhD thesis, 2009.

[35] Zhenjun Ma and Feifang Hu. Balancing continuous covariates based on kernel densities. *Contemporary Clinical Trials*, 34(2):262–269, 2013.

[36] Nathan Mantel, R McHugh, and J Matts. Pre-stratification or post-stratification, 1984.

[37] Scott E Maxwell, Harold D Delaney, and Charles A Dill. Another look at ancova versus blocking. *Psychological Bulletin*, 95(1):136, 1984.

[38] Richard McHugh and John Matts. Post-stratification in the randomized clinical trial. *Biometrics*, pages 217–225, 1983.

[39] Luke W Miratrix, Jasjeet S Sekhon, and Bin Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(2):369–396, 2013.

[40] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyonds. *Stat*, 1050:31, 2016.

[41] Paul R Rosenbaum. Observational studies. In *Observational Studies*. Springer, 2002.

[42] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[43] Donald B Rubin. *Matched Sampling for Causal Effects*. Cambridge University Press, 2006.

[44] Mari Dominique Drouet Kotz Samuel, Dominique Drouet Mari, and Samuel Kotz. *Correlation and Dependence*. World Scientific, 2001.

[45] Elizabeth A Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1, 2010.

[46] Yi-Cheng Wu. Using a priori versus post-hoc assignment of a concomitant variable to achieve optimal power from anova, block, and ancova designs. *Research in the Schools*, 3(1):67–82, 1996.

[47] Zhong Zhao. Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence. *Review of Economics and Statistics*, 86(1):91–107, 2004.

[48] Yeying Zhu, Jennifer S Savage, and Debashis Ghosh. A kernel-based metric for balance assessment. *Journal of Causal Inference*.