2019

On Cluster Robust Models

José Bayoán Santiago Calderón
Claremont Graduate University

On Cluster Robust Models

by

José Bayoán Santiago Calderón

Claremont Graduate University

2019

APPROVAL OF THE DISSERTATION COMMITTEE

This dissertation has been duly read, reviewed, and critiqued by the Committee listed below, which hereby approves the manuscript of **José Bayoán Santiago Calderón** as fulfilling the scope and quality requirements for meriting the degree of Doctor of Philosophy in Economics.

Thomas J. Kniesner, PhD (Chair)

Claremont Graduate University

University Professor

C. Monica Capra, PhD

Claremont Graduate University

Professor of Economic Sciences

Joshua Tasoff, PhD

Claremont Graduate University

Associate Professor of Economic Sciences

Hisam Sabouni, PhD

Claremont Graduate University

Clinical Assistant Professor

On Cluster Robust Models

José Bayoán Santiago Calderón

Claremont Graduate University

2019

ABSTRACT

Cluster robust models are a kind of statistical models that attempt to estimate parameters considering potential heterogeneity in treatment effects. Absent heterogeneity in treatment effects, the partial and average treatment effect are the same. When heterogeneity in treatment effects occurs, the average treatment effect is a function of the various partial treatment effects and the composition of the population of interest. The first chapter explores the performance of common estimators as a function of the presence of heterogeneity in treatment effects and other characteristics that may influence their performance for estimating average treatment effects. The second chapter examines various approaches to evaluating and improving cluster structures as a way to obtain cluster-robust models. Both chapters are intended to be useful to practitioners as a how-to guide to examine and think about their applications and relevant factors. Empirical examples are provided to illustrate theoretical results, showcase potential tools, and communicate a suggested thought process.

The third chapter relates to an open-source statistical software package for the Julia language. The content includes a description for the software functionality and technical elements. In addition, it features a critique and suggestions for statistical software development and the Julia ecosystem. These comments come from my experience throughout the development process of the package and related activities as an open-source and professional software developer. One goal of the paper is to make econometrics more accessible not only through accessibility to functionality, but understanding of the code, mathematics, and transparency in implementations.

DEDICATION

To those that follow may your curiosity never be sated and to Puerto Rico may it prosper.

ACKNOWLEDGMENT

I would like to thank my family for having raised me with solidarity towards others. A long time ago I asked myself what I wanted to study as a mean to empower myself and be able to attain my goals. Although I am agnostic, I received a Catholic education and more specifically from Jesuits and Redemptorists. Their influence and values will forever be part of me. The Jesuits taught to be a man for others. I interpret the Ignatian value to mean that serving others is both a vital responsibility of being human as well as the greatest honor. The Redemptorists taught me the importance of striving for justice for all especially those marginalized and in need. The last component in my analysis was patriotism which my parents taught to me from an early age. The last one being of utter significance having been born, raised, and lived in a colony. My commitment is akin to the The Doctor's promise, "Never cowardly or cruel. Never give up, never give in", and so I promise to forever be relentless in pursuing those aspirations.

A great part that motivates me to work for the common good whether through scientific inquiry or other services is how grateful I am for the contributions of those that came before me. I regard those who have contributed so much for me to enjoy akin to how the dwarfs of Thedas regard their paragons. In my own research, I would like to thank Jeff Wooldridge for making it impossible to do research on many of my interests without reading his work constantly. I would also like to thank all the contributors to open-source software that I constantly use from operating systems, web browsers, numerical libraries, and especially to the R and Julia community. I will do my best to pay it back and forward!

My journey has been long, but I would like to highlight a few people that have contributed to my professional path. Those include my principles of economics, intermediate microeconomics, capstone, and undergraduate adviser Dirk Early at Southwestern University. Thank you for teaching me how to think like an economist and great advising. Another great influence in my professional career is Marie Mora who is an extraordinary person and shows as much through her work in the AEA mentoring program and current research on

issues related to Puerto Rico. I must thank all the community in the mentoring program for having supported me and introduced me to a great network for my benefit and for allowing me to give back to others. A shout out to Coursera and the people involved in the many courses I took such as the Johns Hopkins University Data Science and Rice University Fundamentals of Computing specializations. Another shout out goes to all the various institutions and people that have given me an opportunity to study or work with them through my tenure.

One person in particular deserves an utmost recognition for his great friendship. Hisam, it has been a pleasure to having gone through the program together both part of the cohort and now you being in my committee. We had the best of times from Econ316 problem sets, late night Denny's, morning hikes before macro, quals prep week, driving though California, and hanging out with your family. Very proud of your hard work and ethic and will most definite pay your back in the same currency.

Finally, I want to thank those who have made it all worth it. Noche, my dog, for being the cutest thing ever. TV series, manga, and anime for being the absolute best thing ever. Those are the things that make live truly worth living. I am very exited for the Game of Thrones final season shortly after my defense. I could go on about how having my daily dose of happiness from reading manga and watching TV kept me going, but instead, at least this time, I give you my dissertation instead. Reader, while you go through this work, I only hope that perhaps, you learn a thing or two or realize a thought that leads to helping you or others.

#EstamosBien #LaNuevaReligón

Contents

List of Figures

List of Tables

Chapter 1

On Cluster Robust Models

Regression analysis is commonly used to estimate parameters in a model for prediction or causal inference (i.e., $\hat{\beta}$ vs $\hat{y}$). In either case, the process consists in obtaining good estimates for model parameters which can help answer the following questions: What is the expected effect of a treatment on the outcome ceteris paribus? For a specific case, what is the expected effect of a treatment on its outcome ceteris paribus? For these purposes, the average treatment effect (ATE) is a common parameter of interest which is defined in equation 1.1

$$\beta^{\text{ATE}} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial y_i}{\partial x} \tag{1.1}$$

where i denotes an observation in the population of interest with m observations and $\frac{\partial y_i}{\partial x}$ denotes the marginal effect of x (a feature) on the outcome y. The definition of the ATE is then the average of the partial effects across every observation in the population of interest. When the partial effect is the same across every observation, the average treatment effect simplifies to just the partial effect as seen in equation 1.2.

$$\beta^{\text{ATE}} = \frac{\partial y}{\partial x} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial y_i}{\partial x} \tag{1.2}$$

When the partial effects vary across observations, the population exhibits heterogeneity in treatment effects (HTE). The ATE under HTE simplifies to equation 1.3

$$\beta^{\text{ATE}} = \frac{1}{m} \sum_{j=1}^{g} |G_j| \frac{\partial y_j}{\partial x} \tag{1.3}$$

where j is a cluster identifier that groups all observations that share the same partial effect and $\frac{|G_j|}{m}$ is the share of cluster j in the population of interest.

Estimators for the ATE ideally exhibit the same desirable traits regardless of the presence of HTE such as (1) consistency, (2) unbiasedness or low bias, (3) efficiency, and (4) provide an appropriate estimator for the variance of the estimates. This study examines the properties of estimators given relevant assumptions identified in previous research. However, unlike previous work, the present provide an analysis of these estimators under a set of joint assumptions that fully characterize an application rather than analyzing each condition in isolation. Fully characterizing a scenario allows practitioners to better identify the case for their application and choose among potential options taking into account a more comprehensive view.

A canonical example of HTE relates to the medical literature (e.g., clinical trials). Dettori et al. (2011) discusses the different questions a clinical trial seeks to answer; a clinical trial attempts to answer "is treatment A better than treatment B on average for a select population?" while most clinicians would like to know "is treatment A better than treatment B for this specific patient?" In the ideal scenario, the clinician would be able to correctly identify the patient's type (how it is likely to respond to treatments) and based on the patient's type, prescribe the optimal treatment (e.g., personalized medicine). A second best, is to use the distribution of partial effects to inform the optimal course, for example, using the ATE when screening is unfeasible.

HTE are also critical to fields such as discrimination and program evaluation. For example, the Blinder-Oaxaca decomposition (Blinder 1973; Oaxaca 1973) is one tool specifically

designed to address HTE in the context of discrimination which is often applied in the context of race/gender for wages and returns to education such as in Card and Krueger (1992). In the case of program evaluations, a critical component is external validity or what features influence whether the intervention effect generalizes. External validity may be invalid due to differences in explanatory variables (e.g., differences in implementations) or HTE, meaning that the estimated relation does not hold for cases different from the ones used in the analysis (Athey and Guido W. Imbens 2016).

Differences in the ATE computed from incorrect models do not provide correct estimates and in many cases these differ substantially from cluster robust models (Gibbons, Súarez Serrato, and Urbancic 2018). The importance of HTE is often ignored with grave consequences. One known debacle concerns breast screening guidelines for which non-white women are recommended to start screening decades later than what would be optimal for these groups resulting in serious under-screening (Stapleton et al. 2018). In this case, extrapolating from one subgroup to the population of interest was a mistake and disentangling heterogeneity between groups could have led to an improvement through group specific models.

The ATE under HTE is a function of the partial treatment effects and the population composition (see equation 1.3). For estimating an ATE with potential HTE, the literature has identified several factors that influence the performance common estimators: the sampling design (random or clustered), the treatment mechanism (for experimental studies), and the distribution of the treatment (Abadie et al. 2017). Other considerations are shared with estimators in general such as the sample size. The set of commonly used estimators include pooling, fixed effects (within), and interaction. In addition to these, I include the regression-weighted estimator (RWE) proposed in Gibbons, Súarez Serrato, and Urbancic (2018). This study presents the properties of each estimator under fully characterizing applications in order to assess when are these appropriate and which are optimal if any.

1.1 Assumptions and Conditions

The two conditions that affect the properties of commonly used estimators for ATE with HTE are the sampling design and the distribution of the treatment. The sampling designs can be either random or clustered. Under the random sampling design, every observation has an equal probability of being observed. Under the clustered sampling design, every observation has a probability of being observed that depends on the cluster. Equation 1.4 describes the sampling probability for each observation under each sampling design

$$\mathbb{P}\left(s_i | i \in G_j\right) = \begin{cases} s & \text{if random sampling} \\ \frac{|G_j|}{m} s_j & \text{if clustered sampling} \end{cases} \tag{1.4}$$

where $\mathbb{P}\left(s_i | i \in G_j\right)$ is the conditional probability of observation i, an observation that belongs to cluster j, being observed, s is a constant probability, and $\frac{|G_j|}{m} s_j$ expresses the product of the cluster's share in the population and a cluster specific sampling probability s_j.

The conditional distribution of treatment in observational studies is akin to the treatment mechanism in experimental studies. For this study, I have chosen a continuous treatment normally distributed which allows us to fully characterize its distributions by its first two moments. The four possible conditions considered are,

- $D \sim \mathcal{N}\left(\mu, \sigma^2\right)$
- $D \sim \mathcal{N}\left(\mu_j, \sigma^2\right)$
- $D \sim \mathcal{N}\left(\mu, \sigma_j^2\right)$
- $D \sim \mathcal{N}\left(\mu_j, \sigma_j^2\right)$

which allows for constant or cluster dependent moments for the first two moments (i.e., fully characterizes the distribution).

Considering both the sampling design and the possible treatment distribution, a necessary assumption is that any sample will have, in expectation, a representative distribution of explanatory variables by each cluster. An assumption I impose on the cluster-specific sampling probabilities ($\{s\}$) is that every element is sufficiently relatively high conditioned on the population composition such that that every cluster is represented in the sample.

1.2 Evaluation Metrics

Desirable characteristics of an estimator include consistency, unbiasedness or low bias, efficiency, and an estimator for the variance of the estimates. Approaches to evaluating estimators include deriving the probability limits (i.e., probability limit and convergence rate) and Monte Carlo simulation to observe finite-sample properties. This study uses both approaches by relying on probability limits derived in previous studies to explain and document finite-sample properties through Monte Carlo simulations.

The consistency, bias, and efficiency of estimators is assessed based on the distribution of estimates from Monte Carlo simulations. For evaluating the properties of the variance of the estimates, this study consider the 95% confidence intervals from various variance covariance estimators in order to assess their performance in the context of empirical coverage rates. The empirical coverage rates are simply the average rate at which statistically significance at the given confidence level is observed. The statistical significance refers to the trial estimate being statistically different from expected value (e.g., for consistent estimators the probability limit).

1.3 Estimators

The two major strategies to handle HTE in estimating ATE are: estimating the cluster-specific partial effects or using weights to correct the sample or intermediate estimates used by the estimator (Athey and Guido W. Imbens 2016). However, one should note that simply

using weights to make a sample representative does not provide a suitable model for ATE with possible HTE (Solon, Haider, and Wooldridge 2015). The estimators considered in this study include both strategies.

The most basic estimator is the pooling model which ignores the HTE and fits a single slope for all observations leading to equation 1.5

$$y = \beta_0 + X_1\beta_1 + X_2\beta_2 + u \tag{1.5}$$

where y is a continuous non-censored outcome, X_1 and X_2 are two explanatory features, β_1 and β_2 are the parameters for the partial effects, and u is the idiosyncratic error term. The pooling model makes the assumption that the partial effect $\hat{\beta}_1$ is a good estimate for the ATE with HTE for dimension X_1. A second estimator is the within estimator (fixed effects models) which uses equation 1.6

$$y = \beta_{G_j} + X_1\beta_1 + X_2\beta_2 + u \tag{1.6}$$

where β_1 is the ATE for X_1. The within estimator may be estimated through a series of indicators for cluster memberships or through the annihilated version of the linear predictor/response. A third estimator is the regression-weighted estimator (Gibbons, Súarez Serrato, and Urbancic 2018) which uses the proportional inverse of the variance of the annihilated treatment variable by cluster as the weights as in equation 1.7.

$$\hat{w}_i = \left[\hat{\mathbb{V}}\left(\tilde{X}_j\right)\right]^{-1/2} \tag{1.7}$$

The annihilated distribution of the treatment can be computed from the residuals of the model given by 1.8.

$$X_1 = \gamma_{G_j} + X_2 + u \tag{1.8}$$

One needs to obtain the annihilated version of the outcome, $\tilde{y}$, which can be obtained as the

residuals of the model given by 1.9.

$$y = \gamma_{G_j} + X_2 + u \tag{1.9}$$

One obtains the parameter estimates through the weighted least squares model as in equation 1.10

$$\hat{\beta}_{\text{RWE}} = \left(\tilde{X}^\top W \tilde{X}\right) W^\top \tilde{y} \tag{1.10}$$

Lastly, the interactions model which uses equation 1.11

$$y = \beta_0 + X_1 \left(\beta_1 + \gamma_g\right) + X_2 \beta_2 + u \tag{1.11}$$

where the ATE is constructed using a linear combination of the partial effects and estimates for the population composition. When the sample frequencies are used as the estimates for the population frequency I refer to the estimator as the interaction-weighted estimator (IWE). If some outside estimate is used for estimating the population composition I refer to the estimator as interaction-weighted estimator using population composition estimates (IWE POP).

We present a dummy example of a trial using a population to illustrate how the interaction estimator allows to obtain the ATE. The first step is to fit the model described in equation 1.11. Without loss of generality, the results may look something like Table 1.1. Next, I look at the sample composition (equation 1.12). Lastly, I can apply equation 1.13 for the ATE and equation 1.14 for the standard error estimates. When applying the interactions model to a sample the difference between IWE and IWE POP is how is $\hat{L}$ formed; through the sample frequencies or some outside estimate (e.g., gender ratios, age distributions, urban share of population, etc.)

$$L = \left(\frac{|G_1|}{m}, \frac{|G_2|}{m}, \frac{|G_3|}{m}, \frac{|G_4|}{m}, \frac{|G_5|}{m}\right) = (0.350, 0.385, 0.071, 0.075, 0.119) \tag{1.12}$$

Table 1.1: *Estimates of Cluster-Specifics Partial Effects ($\beta_1 + \gamma_g$) (population)*

Parameter	Estimate
X_1 & cluster: 1	0.488
X_1 & cluster: 2	0.756
X_1 & cluster: 3	1.016
X_1 & cluster: 4	1.269
X_1 & cluster: 5	1.484

$$\hat{\beta}_{X_1}^{\text{ATE}} = \sum_{g=1}^{5} L_g \hat{\beta}_{X_1,g} = 0.807 \tag{1.13}$$

The standard error is computed from the linear combination as well.

$$\sqrt{L^\top \left(\text{Var} \left[\hat{\beta}_{X_1,g} \right] \right) L} = 0.006 \tag{1.14}$$

The variance covariance estimators considered are the expected information matrix (EIM) and the cluster robust variance covariance estimator (Liang and Zeger 1986; Arellano 1987; Rogers 1993; Stock and Watson 2008; Cameron and Miller 2015) which is a generalization of the heteroscedasticity consistent Eicker-Huber-White (EHW) variance covariance estimator (Eicker 1967; Huber 1967; White 1980). The CRVE can be biased in many cases, but appropriate in certain cases with the presence of HTE and clustered sampling in the case of the pooling and within estimators (Abadie et al. 2017). It is also the recommended variance covariance estimator for the RWE estimator. The estimator uses equation 1.15

$$\mathbb{V}_{LZ} \left(\hat{\beta} \right) = \left(X^\top X \right)^{-1} \left(\sum_{j=1}^{J} X_j^\top \Omega_j X_j \right) \left(X^\top X \right)^{-1} \tag{1.15}$$

In this study, the CRVE was considered for the clustered sampling design conditions, but its performance was orders of magnitude worst than the EIM. For this reason, the reported empirical coverage rates are reported using the EIM for all cases.

1.4 Methodology: Monte Carlo Simulation

Consider the following DGP

$$y = \beta_0 + X_1 (\beta_1 + \gamma_g) + X_2 \beta_2 + u \tag{1.16}$$

where y is the outcome variable, X_1 and X_2 features, $(\beta_1 + \gamma_g)$ and β_2 the partial effects for each feature, β_0 is an intercept term, and u the error term. In this model, the DGP shows HTE for the X_1 dimension since the partial effect differs for each observation depending on which cluster g the observation belongs to. The partial effect for an observation in cluster $g = 1$ would then be the linear combination $\beta_1 + \gamma_1$ and so on.

Consider the following population, sampling designs, and treatment distributions:

- $(\beta_0, \beta_1, \beta_2) = (-0.2, 1.0, 0.5)$
- $(\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5) = (-0.5, -0.25, 0.00, 0.25, 0.50)$
- $X_2 \sim \mathcal{N}(-1.0, 0.5)$
- $u \sim \mathcal{N}(0.0, 1.0)$
- $X_1 \sim D$
- Random Sampling: $p = 0.1$
- Clustered Sampling: $(p_1, p_2, p_3, p_4, p_5) = \{0.10, 0.15, 0.20, 0.25, 0.30\}$ (in each trial each cluster dependent sampling probability sampled from the set without replacement)
- Under constant mean and variance $D = \mathcal{N}(0.0, 1.0)$
- Under variable mean $D = \mathcal{N}(\gamma_g, 1.0)$
- Under variable variance $D = \mathcal{N}(0.0, 0.25 + |\gamma_g|)$
- Under variable mean and variance $D = \mathcal{N}(\gamma_g, 0.25 + |\gamma_g|)$

Using the DGP specified above I generate a population with the following summary statistics:

- $m = 100,000$
- $(|G_1|, |G_2|, |G_3|, |G_4|, |G_5|) = (35013, 38502, 7093, 7455, 11937)$
- $\beta_{X_1}^{\text{ATE}} = 0.8070025$

Given the DGP and sampling designs, this study can explore the behavior of different estimators for the ATE using common metrics for the quality of the first and second moments under various conditions. Each Monte Carlo simulation consists of 1,000 trials. This exercise allows to survey various results in the literature in an accessible format.

1.5 Results

The results are reported through the empirical distribution of the parameters estimates in Figure 1.1, the estimated distribution moments in Table 1.2, and the empirical coverage rates in Table 1.4. Under the random sampling design all estimators show a bell curved distribution centered on the true parameter value except for the within estimator which shows bias when the variance of the treatment distribution is correlated with the clusters and the pooling estimator with the treatment distribution having a constant mean and correlated variance. The RWE is robust to the presence of correlated variance of treatment different from the pooling nor within estimators.

Under the clustered sampling design only the interaction-weighted estimator using population estimates showed a first-rate convergence rate. While the other estimators had a mean close to the parameter value, the standard deviation is about six times larger, has about three times as much skewness, and about one kurtosis less than under the random sampling counterpart.

Empirical coverage rates for the random sampling seem appropriate having rates close to the expected 0.95. One exception is the RWE which has significant lower coverage rates. The t-distribution used with the estimated variance covariance estimates had a residual degrees of freedom based on the single parameter à la manner in the implementations by the authors. However, it may be the case that the residual degrees of freedom should be adjusted to account for the intermediate estimates. Under the clustered sampling design both the CRVE and the EIM estimators (to lesser extend) are unable to capture the variability of the parameter estimates.

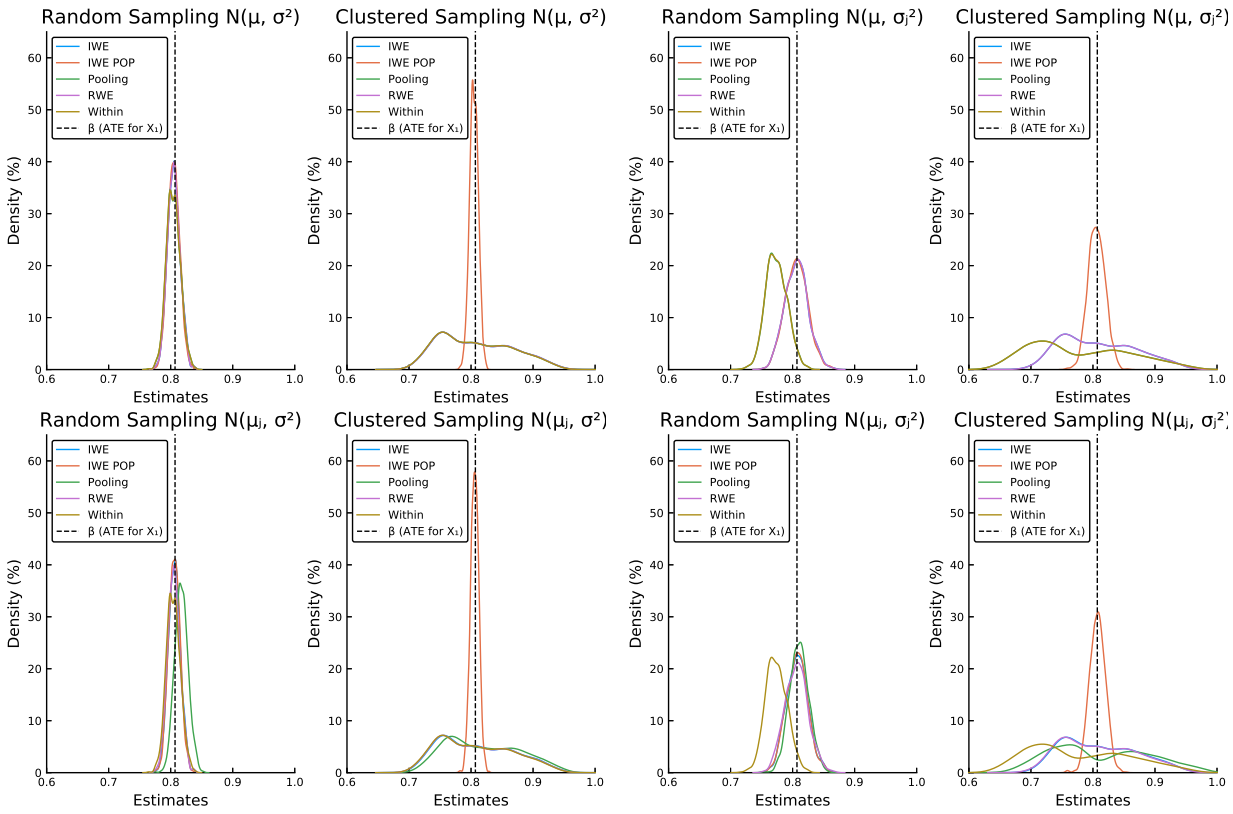

Figure 1.1: *Distribution of ATE Estimates*

Table 1.2: *Moments of Parameter Estimates (1,000 trials) With Variance of Treatment Independent*

Distribution	Estimator	Mean		Std Dev		Skewness		Kurtosis	
Sampling Design		R	C	R	C	R	C	R	C
$\mathcal{N}(\mu, \sigma^2)$	Pooling	0.80	0.81	0.01	0.06	0.12	0.34	0.03	-0.92
$\mathcal{N}(\mu, \sigma^2)$	Within	0.80	0.81	0.01	0.06	0.12	0.34	0.02	-0.92
$\mathcal{N}(\mu, \sigma^2)$	RWE	0.80	0.81	0.01	0.06	0.13	0.34	0.01	-0.93
$\mathcal{N}(\mu, \sigma^2)$	IWE	0.81	0.81	0.01	0.06	0.14	0.34	0.02	-0.93
$\mathcal{N}(\mu, \sigma^2)$	IWE POP	0.80	0.80	0.01	0.01	0.08	-0.05	-0.04	-0.04
$\mathcal{N}(\mu_j, \sigma^2)$	Pooling	0.82	0.82	0.01	0.06	0.05	0.28	0.09	-0.97
$\mathcal{N}(\mu_j, \sigma^2)$	Within	0.80	0.81	0.01	0.06	0.12	0.34	0.02	-0.92
$\mathcal{N}(\mu_j, \sigma^2)$	RWE	0.80	0.81	0.01	0.06	0.13	0.34	0.01	-0.93
$\mathcal{N}(\mu_j, \sigma^2)$	IWE	0.81	0.81	0.01	0.06	0.07	0.34	0.06	-0.93
$\mathcal{N}(\mu_j, \sigma^2)$	IWE POP	0.81	0.81	0.01	0.01	0.02	-0.05	-0.02	0.11

The average treatment effect is 0.807.
Sampling designs include: random (R) and clustered (C).

Table 1.3: *Moments of Parameter Estimates (1,000 trials) With Variance of Treatment Correlated*

Distribution	Estimator	Mean		Std Dev		Skewness		Kurtosis	
Sampling Design		R	C	R	C	R	C	R	C
$\mathcal{N}\left(\mu, \sigma_j^2\right)$	Pooling	0.77	0.78	0.02	0.08	0.14	0.35	-0.05	-0.95
$\mathcal{N}\left(\mu, \sigma_j^2\right)$	Within	0.77	0.78	0.02	0.08	0.14	0.35	-0.05	-0.95
$\mathcal{N}\left(\mu, \sigma_j^2\right)$	RWE	0.81	0.81	0.02	0.06	0.13	0.34	-0.03	-0.85
$\mathcal{N}\left(\mu, \sigma_j^2\right)$	IWE	0.81	0.81	0.02	0.06	0.13	0.34	0.00	-0.85
$\mathcal{N}\left(\mu, \sigma_j^2\right)$	IWE POP	0.81	0.81	0.02	0.01	0.10	0.00	-0.01	0.06
$\mathcal{N}\left(\mu_j, \sigma_j^2\right)$	Pooling	0.81	0.81	0.02	0.08	-0.02	0.24	0.05	-1.04
$\mathcal{N}\left(\mu_j, \sigma_j^2\right)$	Within	0.77	0.78	0.02	0.08	0.14	0.35	-0.05	-0.95
$\mathcal{N}\left(\mu_j, \sigma_j^2\right)$	RWE	0.81	0.81	0.02	0.06	0.13	0.34	-0.03	-0.85
$\mathcal{N}\left(\mu_j, \sigma_j^2\right)$	IWE	0.81	0.81	0.02	0.06	0.09	0.34	-0.04	-0.87
$\mathcal{N}\left(\mu_j, \sigma_j^2\right)$	IWE POP	0.81	0.81	0.02	0.01	0.07	-0.06	-0.08	0.34

The average treatment effect is 0.807.
Sampling designs include: random (R) and clustered (C).

Table 1.4: *Empirical Coverage Rates (95% nominal rate, 1,000 trials)*

		$\hat{\beta}^{\text{ATE}}_{X_1}$				
Distribution	Sampling	Pooling	Within	RWE	IWE	IWE POP
$\mathcal{N}(\mu, \sigma^2)$	Random	0.94	0.94	0.97	0.95	0.96
	Clustered	0.16	0.16	0.15	0.15	0.97
$\mathcal{N}(\mu_j, \sigma^2)$	Random	0.93	0.95	0.97	0.95	0.96
	Clustered	0.13	0.16	0.15	0.15	0.97
$\mathcal{N}\left(\mu, \sigma_j^2\right)$	Random	0.95	0.94	0.73	0.96	0.97
	Clustered	0.13	0.13	0.15	0.26	0.97
$\mathcal{N}\left(\mu_j, \sigma_j^2\right)$	Random	0.93	0.95	0.73	0.96	0.96
	Clustered	0.07	0.14	0.15	0.24	0.98

1.6 Case Study

In order to validate the results, this study includes a case study using data from the IPUMS Current Population Survey (CPS) for Social and Economic and Health Research (Flood et al. 2018a). Imagine you are a high school counselor at a men's high school in California. Something you may want to know if the return on education attainment for men in the civilian population living and working in California. Define the population of interest as 25 - 64 years old white or black men part of the civilian population living and working in California. Furthermore, define the treatment or variable of choice as education attainment: (1) drop high school, (2) high school diploma or equivalent, (3) bachelor's degree, or (4) a graduate degree. The outcome can be operationalized as the typical weekly earning before deductions (excludes self-employed). One covariate you would like to include is a career stage proxied by age groups: (1) 25 - 34, (2) 35 - 44, (3) 45 - 64 in years old. You might be concerned that the returns to education might have a racial disparity (i.e., heterogeneity in treatment effects), but even though you could employ a screening / discrimination choice, you prefer not to. Hence, you would like to use the average treatment effect.

After reflecting on the simulation results you decide to test the following estimators:

1. pooling

$$\ln(\text{inc}) = \beta_0 + \text{age}\ \beta_1 + \text{educ}\ \beta_2 + u \tag{1.17}$$

2. within

$$\ln(\text{inc}) = \beta_0 + \text{age}\ \beta_1 + \text{educ}\ \beta_2 + \text{race}\ \beta_3 + u \tag{1.18}$$

3. interaction

$$\ln(\text{inc}) = \beta_0 + \text{age}\ \beta_1 + \text{race} * \text{educ}\ \beta_2 + u \tag{1.19}$$

You use the IPUMS website and request a sample with the appropriate variables, sample selection, restrict the sample to that of interest, and make sure to use the appropriate sample weights such that your sample is representative. Given the quality of the survey design you operate under the assumption of random sampling design. Afterwards you fit the three models and ponder on whether and what estimates for the population composition should you use for the interaction estimator. You decide to consider three choices: (1) sample weights, (2) survey-weighted sample frequencies, and (3) best, potentially external, estimates of the population composition. For the "best" estimates of the population you generated the relevant tables using the CPS Table Creator tool.

You find the following results (see Figure 1.2) for the ATE. Your overall results suggest a high school diploma predicts an 11% increase in income compared to no high school diploma, a bachelor's degree predicts a 8% increase in income compared to a high school diploma or equivalent, and an advanced degree predicts a 3% increase in income compared to a bachelor's degree all interpretations under ceteris paribus and only about the defined population of interest. The direction and magnitudes are as expected as more education seems correlated with higher earnings and the percentage magnitudes are decreasing. The first observation is that the pooling estimator gives different parameter estimates than the other estimators suggesting the presence of HTE and failure of the pooling estimator to provide good estimates. From the contingency table you notice the distribution between

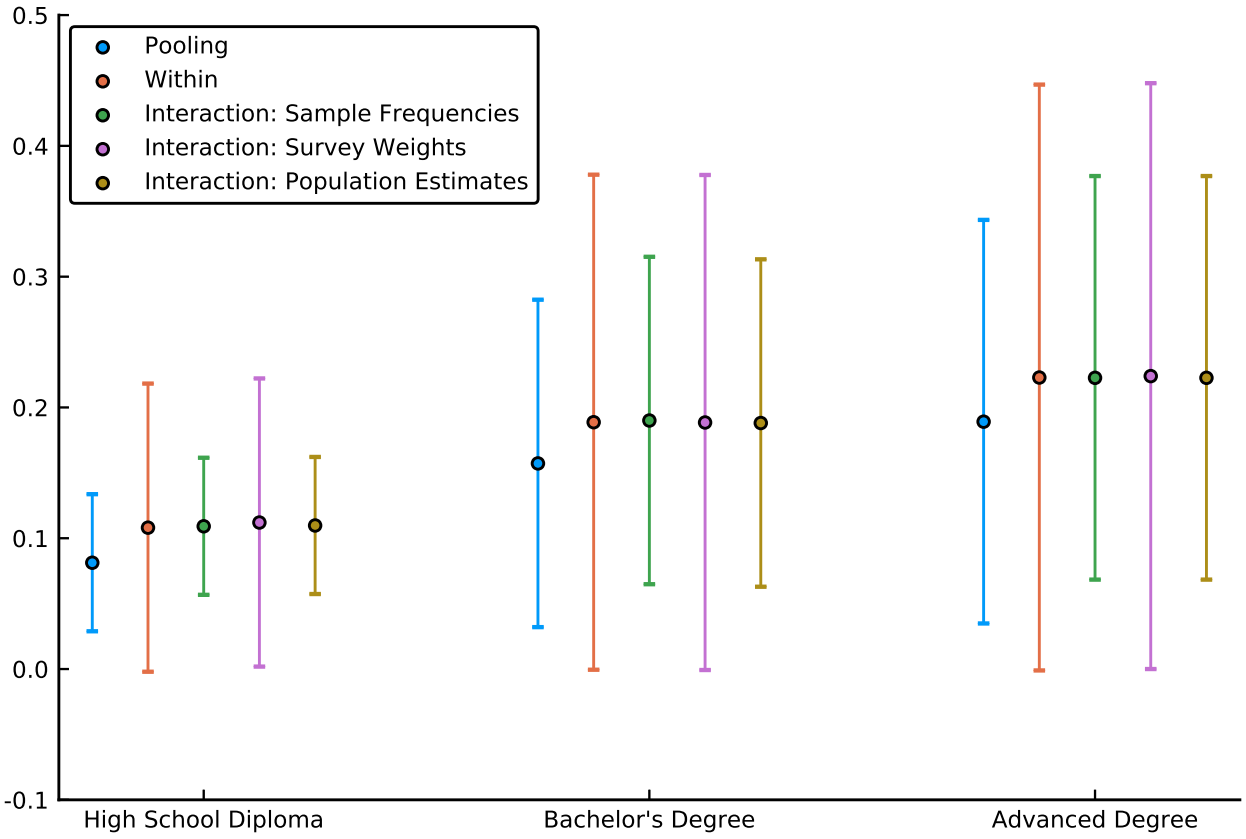

Figure 1.2: *Estimates of ATE for Returns to Education*

race by treatment is consistent across education levels (i.e., treatment conditions) and thus the within estimator should be an acceptable option. This observation is robust to using weighted/unweighted sample statistics or population composition which consistently yield an approximate 1:9 Blacks to Whites ratio.

You examine the parameters estimates and notice a difference not in point estimates, but in the confidence intervals. The interaction estimator with sample weights and the interaction estimator with survey weights are for all purposes equivalent. This phenomenon might be increasing with the adoption of replication weights by many widely used surveys. Likewise, the interaction estimator with sample frequencies is practically equivalent to the interaction estimator with the population composition estimates. Between the two classes of estimators, the interaction estimator with sample or population estimates of the composition offers a more efficient estimator and thus it is chosen as the preferred approach.

This case study covered various tools and consideration from an applied side that will hopefully help practitioners internalize the content of the study and make it easier to adopt in their research. As with the other sections, the scripts in the appendix can help practitioners incorporate these consideration in their day to day. For those purposes, the scripts for replicating the case study are provided as a Stata do file.

1.7 Conclusion

This study contributes to the literature of econometrics of program evaluation by exploring the performance of various estimators for ATE under HTE across applications fully characterized by simultaneously considering a set relevant conditions. The set of conditions considered to characterize an application were the sampling design, random or clustered, and the distribution of the treatment as a function of the first and second moments. The estimators explored in this study were: pooling, within, interaction-weighted (using sample frequencies or population estimates), and the regression-weighted estimator.

Based on the results from the Monte Carlo simulations, the sampling design does not change the expected value the parameter estimates. The variance of the distribution of parameter estimates increased three to six fold under clustered sampling compared to random sampling. Based on the distribution of the treatment effect in the population, which is skewed (1.02), under the clustered sampling, the skewness of the parameters estimates increased three fold compared to under random sampling. This result suggests that the parameters estimates are an image of the distribution of the treatment effect in the population under both sampling designs, but more so under clustered sampling. Clustered sampling also decreased the kurtosis of the parameter estimates by around one.

The within estimator was shown to be biased when the variance of the distribution of the treatment was correlated with the clusters. The pooling estimator seems to be biased as well when the mean of the distribution of the treatment is constant, but the variance is correlated with the clusters. As an alternative, the RWE provides a consistent and unbiased estimator, but lacks a proper estimator for the variability of the estimates as it shows significantly lower empirical coverage rates.

Under clustered sampling design the distribution of the parameters estimates is a function of the population composition which conventional variance covariance estimators fail to capture. Under random sampling, the distribution of the parameter estimates follows a t-distribution and the variance covariance estimators provide proper empirical coverage rates.

One estimator showed improved performance across all specifications: the interaction-weighted estimator using estimates of the population composition rather than sample frequencies. The IWE using sample frequencies was still a top performant option across the various conditions. These results suggest practitioners should consider interaction-weighted estimators as a top choice or at least as a robustness specification.

Estimates of the population composition beyond sample frequencies suggest a stark difference in robustness and efficiency provided for the interaction weighted estimator. In

addition, in many applications the partial effects in the context of HTE may be as insightful as the ATE. One last argument in favor of interaction-weighted estimators includes the high prevalence of correlated variance in the distribution of treatment effects with clusters. For example, in the case of seeking medical treatment, variability in promptness is many times correlated with seriousness and also with effectiveness of treatments.

Future work might consider the sensibility and robustness of the estimators based on various distributions of the ATE. Another consideration which is not in the scope of this study is the performance the estimators when there are true fixed effects (i.e., curse of dimensionality) or under-powered studies which provide bad estimates of the partial effects.

Practitioners are recommended to use the interaction-weighted estimators as the first-best with the best estimates for the population composition or sample weights if those are not available. The RWE is a suitable alternative which provides similar performance and does not rely on estimating the partial effects, but caution is advised on using common variance-covariance estimators for such models.

Chapter 2

Better Cluster Structures

Cluster robust models refer to models that attempt to estimate average treatment effects (ATE) in the potential presence of heterogeneity in treatment effects. Many strategies rely on obtaining an estimate of the cluster structure (e.g., within / fixed effects, interactions). It is customary to proxy the cluster structure based on theory and convenience. For example, potential treatment effects may be specified based on location (e.g., jurisdictions such as countries or states), period (e.g., year or seasonality), or based on the literature (e.g., race, ethnicity, gender). This study contributes to the field of econometrics of program evaluation by proposing a framework to understand the main characteristics of a cluster structure and method to help obtain better estimates for these.

The core problem is identifying a suitable cluster structure to use for estimating the average treatment effect for a parameter with heterogeneity in treatment effects. The solution should allocate each observation to a unique cluster where every cluster has the same partial treatment effect and clusters share the same partial effects. Each cluster structure should be defined and applied by each treatment effect. For a solution to be useful, one should have some way to map the estimated partial effects to a population composition. For example, using quantiles to allocate various clusters yields a clean way to estimate the population composition based on the cluster shares. The goal of this study is to help practitioners

understand, qualify, and estimate better cluster structures through a theoretical framework and examples that showcase available tools and thought processes.

The average treatment effect is given by

$$\beta_{ATE} = \frac{1}{m}\sum_{i=1}^{m}\frac{\partial y_i}{\partial x} = \sum_{j=1}^{J}\pi_j\beta_j \tag{2.1}$$

which can be understood as the weighted sum of the partial effects for each observation in the population of the interest or as the weighted sum of the partial effects times the its share in the population. The homogeneity in treatment effects leads to the simplified version where the partial effect is equal to the ATE. Evidently, one vital component in obtaining sound estimates for the ATE is the cluster structure which will influence both the estimates for the partial effects ($\hat{\beta}_j$) and the population composition ($\{\hat{\pi}_j\}$).

In certain applications cluster structures can be developed cleanly based on a sound theoretical model. For example, Bitler, Gelbach, and Hoynes (2006) studies the effect of Connecticut's Jobs First welfare reform experiment and estimates quantile treatment effects for various subgroups based on labor supply theory. As the treatment effect varies by income, it makes it a natural choice to explore a cluster structure based on the earnings dimension. The subsequent Bitler, Gelbach, and Hoynes (2017) explored more in details potential cluster structures and theoretical bases. For example, having the youngest child younger than five years old (e.g. operationalization of having small children) proxies for high fixed costs of work. Other dimensions included pre-intervention history (i.e., earnings and welfare), level of education of case head, etc.

Other applications lack a strong theoretical framework and may be good candidates for data-driven approaches. A few data-driven approaches include multiple hypothesis testing corrected for false discovery (List, Shaikh, and Xu 2016), subgroup analysis, *causal trees* which are modified regression trees (Athey and Guido W Imbens 2016), and other variants (Athey and Guido W Imbens 2017). The three main themes are (1) exhaustively explore

the search space accounting for multiple hypothesis testing to obtain valid inference, (2) exploit conditions in the experiment or process (e.g., requirements for menu), or (3) apply a supervised or unsupervised machine learning tool to strike a balance between robustness and parsimony.

This study has three purposes: (1) develop a framework for qualifying the quality of a cluster structure estimate, (2) examine the behavior of estimators as function of the quality of cluster structure proxies (i.e., does it matter?), and (3) examine potential ways to obtain better estimates for cluster structures (i.e., can we do better?).

Chapter 1 considered four estimators that are commonly used for estimating ATE with HTE: (1) pooling, (2) within / fixed effects, (3) interactions, and (4) the regression weighted estimator (RWE). The performance of these estimators was explored based on the sampling design (i.e., random vs clustered) and the distribution of the treatment (combination of first and second moments either cluster independent or dependent). This study explores the performance of the interaction weighted estimator which showed the best behavior for ATE with HTE.

2.1 Framework

Cluster robust models typically assume perfect information where each observation in the sample is properly identified as a member of its cluster on every relevant dimension. However, in applied work it is rarely if ever the case and one must use a good approximation based both in theory and educated guesses. For example, a survey design might collect information about individuals by households. In this case, for certain dimensions it is sensible to assume that members of the same household might share the same partial treatment effect for some dimensions and that information could lead to a suitable proxy.

Cluster proxies have certain properties that characterize their quality: (1) purity, (2) level, and (3) dimension. The purity of a cluster measures the degree to which observations

in a cluster are part of the same group and no foreign observations are included. The level refers to clusters sharing the same partial effects not being classified as the same cluster. The dimension refers to ensuring the clusters are defined by the partial effect of the treatment in question and not being defined by the partial effect of other treatments. Consider a study on self-image and one of the explanatory variables being exposure and use of social media (e.g., instrumented by hours spent daily on social media). The population of interest is heterogeneous in treatment effects by the exposure and use of social media by age groups: (1) 18 - 25, (2) 26 - 45, and (3) 46+. The researcher does not know the true data generation process and has access to their ages; thus models it by specifying the following clusters: (1) 18 - 25, (2) 26 - 35, (3) 36 - 45, (4) 46 - 60, (5) 61+. Using the 5-groups proxy, the researcher is able to achieve perfect purity as every cluster would only contain members of the same cluster. In terms of the level, the proxy has an incorrect level since clusters have been desegregated (i.e., groups 2 and 3 should be together as well as groups 4 and 5). Lastly, the model should specify the clusters as proper of the social media exposure and use variable and not extrapolate the cluster structure to other predictors.

2.2 Proposal

The proposal offers a way to address estimating cluster structures based on previous work, available tools, and new results that rise from the theoretical framework and analysis in this study.

Previous work suggests a broad and exhausting brute search (e.g., multiple hypotheses testing) or a data-driven (e.g., causal trees) approach. The results from the analysis of purity in cluster structures shows that obtaining consistent estimates of the partial effects is a first-order concern. Results from the analysis of precision levels shows that there is a low cost to erring on the side of consistency even if the solution is not a parsimonious one. Figures 2.2 and 2.3 show that even a high purity wrong level cluster structure estimate can

still yield a good estimate for the average treatment effect.

Given a set of suitable conditions, it is possible to learn the true cluster structure from a proxy that has high purity, but at an incorrect level. The technique requires that each sub-cluster has enough power to provide consistent estimates (e.g., a pair of observations or household level might not provide enough observations). For example, observations that are located in a hierarchical setting such as county-state, could have a cluster structure at some unknown level. The county level presents a potential high-purity proxy, but would lead to a model with a great number of parameter estimates. In addition, it may be that the population distribution is not accurately estimated at the county level. One potential way to construct a suitable proxy would be to estimate the model with the proxy at the county level and test whether the parameter estimates of certain heterogeneous parameters are the same (e.g., a joint significance test such as a Wald or score test). One can re-estimate the models until each heterogeneous in treatment parameter has been identified which leads to a parsimonious yet robust model.

In summary, the proposal provides a new approach to existing techniques in the way to search for a suitable cluster structure estimate. First, it highlights that parsimony should not be a primordial criterion for estimating the cluster structure. Second, it proposes using hypotheses tests not as a valid instrument for inference, but as a rule to determine the algorithms path (i.e., similar to bucket size and numbers in causal trees). Third, it recognizes that the cluster structure is a proxy which is a source of error for which it emphasizes in obtaining valid partial effects estimates. Lastly, it allows for the initial path to be theoretically driven or agnostic depending on the application.

2.3 Simulation Design

The following simulation shows the behavior of the interaction weighted estimator under random sampling as a function of the purity and precision of cluster structure estimates.

The analysis focuses on the first and second moments. However, the empirical coverage rates are orthogonal to cluster structures and thus results are presented only for the first moment (i.e., using the interaction weighted estimator the expected information matrix yields nominal coverage rates).

All simulations use the following design with a data generation process as follows

$$y = \beta_0 + X_1\gamma_g + X_2\beta_2 + u \tag{2.2}$$

where y is the outcome variables, X_1 and X_2 are explanatory variables, β_j and γ_g are the main parameters, and u is the idiosyncratic error term. The data generation process indicates that the population of interest is heterogeneous in treatment, namely in the X_1 dimension. Clusters $g \in G$ are determined by the relation between X_1 and y; meaning γ_g. The explanatory variables as well as the error term are standard normal distributed $X_1, X_2, u \sim \mathcal{N}(0, 1)$. The main parameters, $(\beta_0, \beta_1) = (-0.2, 0.5)$, and

$$\gamma_g = \begin{cases} -0.50 & \text{if } g = 1 \\ -0.25 & \text{if } g = 2 \\ 0.00 & \text{if } g = 3 \\ 0.25 & \text{if } g = 4 \\ 0.50 & \text{if } g = 5 \end{cases} \tag{2.3}$$

The population is finite with 100,000 observations and each observation is assigned a subgroup $g \in \{\mathbb{Z} \,|\, g \in [1, 5]\}$ with equal probability. The sampling design is a random sampling design with 5% probability of being observed. In other words, the probability of sampling an observation with of a particular cluster is the share of the cluster in the population times the sampling probability.

Each simulation performs an experiment based on repeating trial where a sample is drawn

and the model is specified as

$$y \sim X_1 \ \& \ G + X_2 \tag{2.4}$$

where G is a categorical variable and & represents the interaction operator. The model is fitted and the parameter estimates are used to compute the 95% confidence interval for the average treatment effect of X_1. The estimated average treatment effect is recorded as well for computing the empirical distribution of estimates.

During each section a different procedure is used to construct G which allows to examine the consequences varying conditions and emerging trade-offs for estimating and using a cluster structure. In order to test how purity affects the model, the varying purity condition perturbs G using the following rule: with probably equal to the chosen expected purity rule α, the cluster identifier is the true group and with $1-\alpha$ probably a different cluster identifier is chosen based on the relative distance in partial effects. The exact probability distribution for different cluster identifiers is given by the softmax function applied to the inverse of the squared root of absolute differences (see table 2.1 for resulting probabilities). This design allows for the mistakes in assigning observations to clusters to be a function of how different the clusters in question are in terms of partial effects. One could make the analogy of an observation belonging to age group 25-30 might be mislabeled more likely with the 31-35 than the 56-60 years old cluster.

Table 2.1: *Cluster Structure Based on Purity Levels*

True Cluster	**Assigned Label**				
	1	2	3	4	5
1	α	$(1-\alpha)$ 0.42	$(1-\alpha)$ 0.24	$(1-\alpha)$ 0.18	$(1-\alpha)$ 0.16
2	$(1-\alpha)$ 0.33	α	$(1-\alpha)$ 0.33	$(1-\alpha)$ 0.19	$(1-\alpha)$ 0.14
3	$(1-\alpha)$ 0.18	$(1-\alpha)$ 0.32	α	$(1-\alpha)$ 0.32	$(1-\alpha)$ 0.18
4	$(1-\alpha)$ 0.14	$(1-\alpha)$ 0.19	$(1-\alpha)$ 0.33	α	$(1-\alpha)$ 0.34
5	$(1-\alpha)$ 0.16	$(1-\alpha)$ 0.18	$(1-\alpha)$ 0.24	$(1-\alpha)$ 0.42	α

α is the expected purity level.

In the analysis of precision level conditions, G is obtained by allowing each observation to be a member of k clusters restricted such that all subgroups have perfect purity with equal probability. For example, given a true cluster structure by state, the state would be broken up into k various jurisdictions and the observations would be assigned to some of the jurisdiction within the state.

2.4 Analysis for Purity Conditions

Proxies for establishing cluster structures are not necessarily the true cluster structure in the data generation process. In most empirical work, one can use educated guesses that might have sensible accuracy, but it is important to understand how the results vary based on the quality of the proxy. This simulation studied the distribution of parameter estimates for a cluster robust model. Figure 2.1 shows the distribution of parameters estimates for the average treatment effect of X_1.

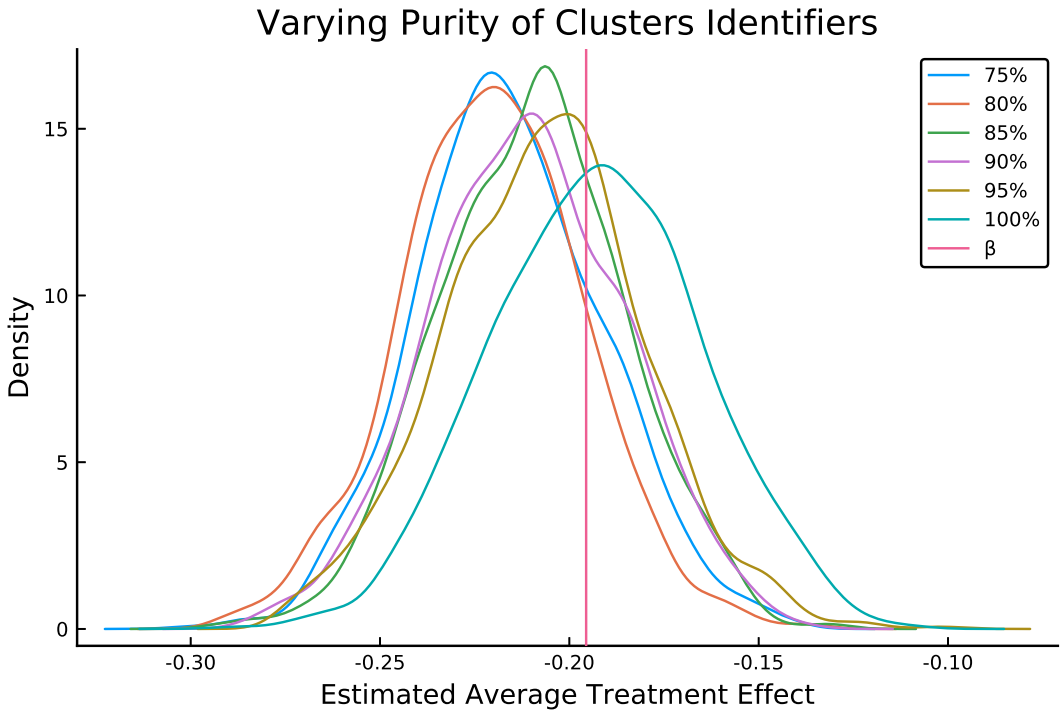

Figure 2.1: *Distribution of Parameter Estimates (1,000 replications)*

The performance of the interaction weighted estimator for the average treatment effect of the parameter with heterogeneity in treatment effects varies as a function of the purity of the cluster identifiers. Under the ground truth (i.e., 100% purity), the distribution is centered at the parameter value as the estimator is consistent and unbiased. However, as noise/error is introduced in the cluster structure, the estimator becomes biased and does not converge to the parameter value.

2.5 Analysis for Precision Level Conditions

Cluster robust models can be specified at a level that increases the probability of the proxy to being one of high purity. However, it may be the case that it leads to a large number of parameter estimates and difficult to obtain estimates of the population composition at that level of precision. An alternative is to estimate the model and use the results to develop a more parsimonious yet robust model based on the information gained (i.e., information about differentials in partial effects).

Figure 2.2 shows the distributions of partial effects for each condition and figure 2.3 shows the estimated average treatment effects. The k in the first figure indicates how many clusters share the same partial effects, but the proxy treats as different. As long as the partial effects are consistently estimated, the probability limits will be the same yielding similar partial effects estimates. Ideally, these would be consolidated to a model where $k = 1$, meaning a correct level specification where no cluster shares the same partial effects. The next figure shows that even if the model is not parsimonious, the same average treatment effects can be obtained regardless. Differences in inference could rise from lower power and changes to the degrees of freedom. These figures illustrate that cluster robust model with pure proxies, but at the wrong level still provide consistent estimates. Moreover, the model results could be used to construct a refined proxy which uses the estimates of heterogeneous in treatment parameters for constructing a proxy at the correct level.

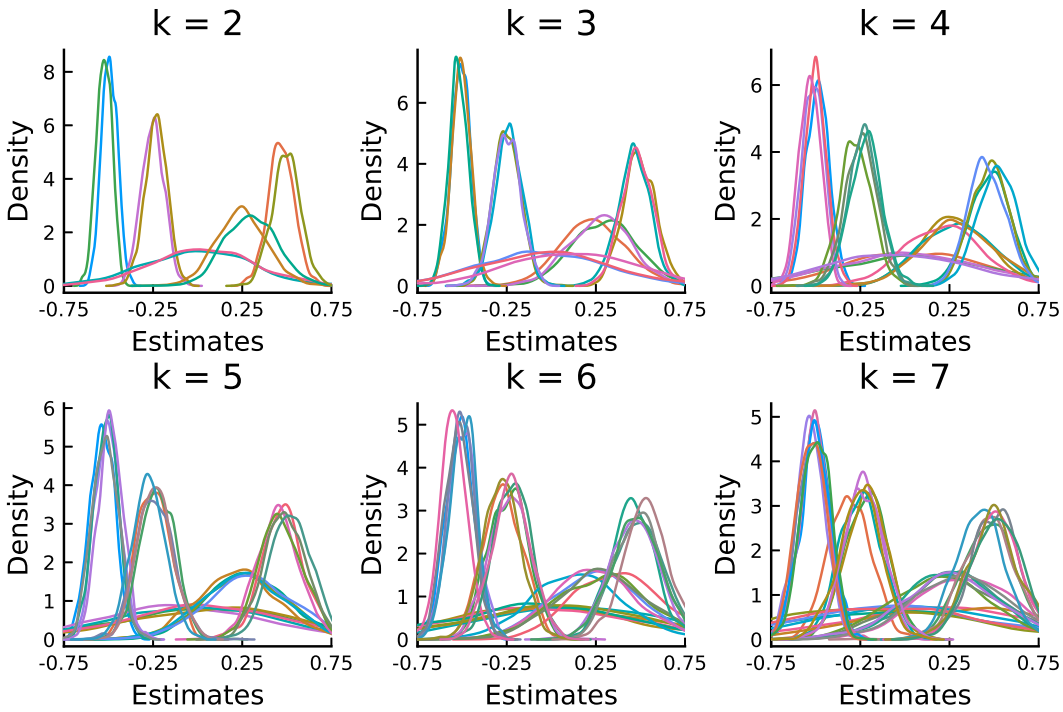

Figure 2.2: *Distribution of Partial Effects (1,000 replications)*

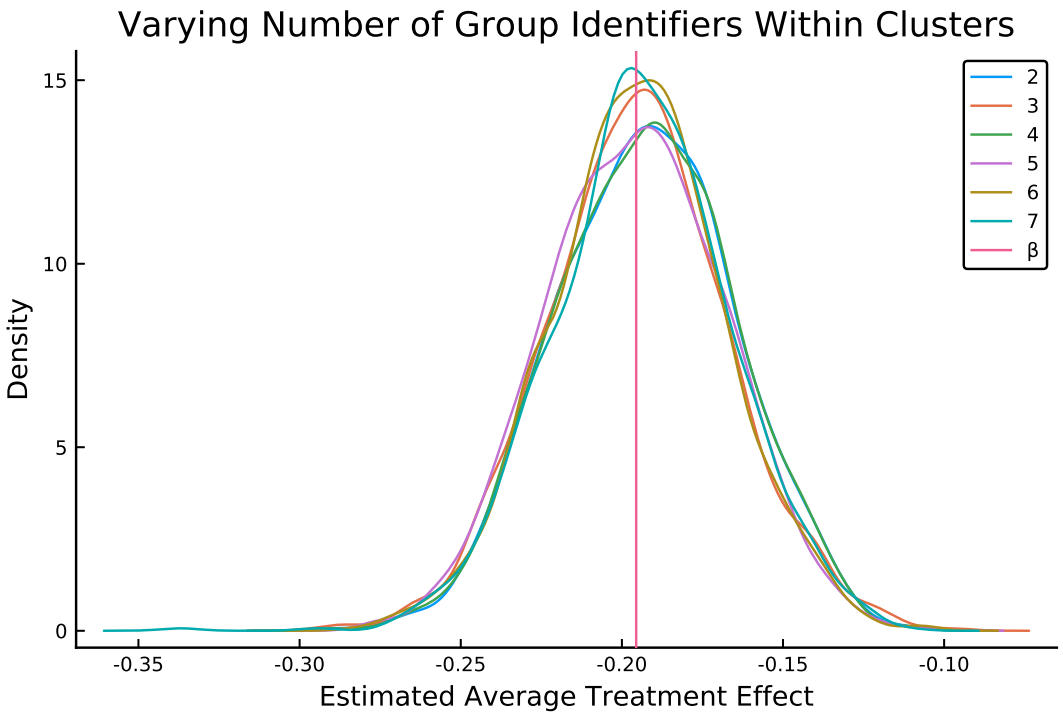

Figure 2.3: *Distribution of Parameter Estimates (1,000 replications)*

2.6 Higher Dimensional Issues

Regression analysis use features as opposed to variables in the sense that a feature is a mapping from an observable to an explanatory/predictive instrument. For example, income in thousands could be a variables, but it may be better for the model operationalize it through income brackets which shows the distinction between variables and features. Likewise, the cluster structure needs not to be a function of variables, but a feature. For example, rather than being age groups, a feature that allows to define a correct cluster structure might be the age groups and gender (i.e., the interaction). Another example could be a combination of spatial and temporal dimensions such as states and years.

Models may have multiple main parameters of interest. For example, one may want to estimate both a policy impact and the effect of a control variable. The analysis for obtaining average treatment effects with parameters which exhibit heterogeneity in treatment effects generalizes past the single parameter of interest. In other words, one would have to estimate the cluster structure for each parameter rather than imposing the same structure for every parameter as a requirement.

2.7 Caveats

In the case of homogeneity in treatment effects, any cluster structure with sufficient power to consistently estimate the partial effect will lead to a consistent estimate. In this regard, cluster robust models will are relatively safe which makes these appropriate as main estimation technique or as a robustness check.

High-purity cluster structure estimates provide a powerful tool for obtaining valid partial effect estimates. However, for estimating average treatment effects one might need to collapse the clusters up to the point that allows for estimating the population composition. A parsimonious/stop-rule criterion could be used is to collapse clusters up to the point that is feasible to estimate the population composition and apply the partial effect estimates.

Practitioners may face the problem where a potential heterogeneous in treatment parameter could have a cluster structure dependent on more than one possible dimension. For example, support for welfare programs could be heterogeneous in treatment by age, political ideology or some combination. The first step is to identify if the sampling cluster is clustered by any of the dimensions. If the sampling design is not clustered, a consistent estimator for the average partial effect in the population can be obtained if the samples are representative (both in cluster composition and cluster-specific distribution of attributes). In the case that the sampling design is clustered in a dimension that can pose an issue requiring a cluster robust model, the following guidelines can help. (1) If the dimension is unrelated to the cluster structure, the parameter estimates should converge to the same probability limit. (2) If the parameter estimates differ statistically, modify the cluster structure to a higher expected purity level for testing whether the cluster structure is at the wrong level.

2.8 Case Study

2.8.1 Objective

The motivation of the case study is to estimate the average treatment effect of education on earnings (i.e., returns to education). One potential source of heterogeneity in treatment effects for returns to education is race and ethnicity. Another potential source of heterogeneity in treatment effects could be through gender (Dougherty 2005). In other words, we would like to obtain estimates for returns to education that are robust to potential heterogeneity in treatment effects on basis of race/ethnicity/gender combinations. For these purposes, an intermediate objective is to obtain a suitable combination of cluster structures and population composition which can be used to estimate the average treatment effects (i.e., partial effects, population shares, and components for estimates of the second moment of the parameter estimates).

2.8.2 Target for Inference

For this case study, the target for inference is defined as the returns to education for California residents for work-age employed (excludes self-employed) in the civilian workforce. For these purposes, work-age is operationalized as 18 - 79 years old not enrolled in an educational program. A representative sample for this population is obtained through data from the Institute for Social Research and Data Innovation Current Population Survey Data for social, economic, and health research also known as IPUMS CPS (Flood et al. 2018b).

2.8.3 Data and Methodology

For developing the model specification, we first consider a previous study on differentials of returns to education. Barrow and Rouse (2005) used the National Longitudinal Survey of Youth, 1979 (NLSY79) 1993 round for estimating returns to education by groups across racial and ethnic characteristics. The NLSY79 survey is sponsored and directed by the U.S. Bureau of Labor Statistics and conducted by the Center for Human Resource Research at The Ohio State University. Interviews are conducted by the National Opinion Research Center at the University of Chicago. The estimates used an ordinary least squares estimator regressing the natural logarithm of hourly pay on years of completed education, a third-order polynomial in age, a gender indicator, a four geographic regions indicators, and a constant. The variance covariance estimator used is the expected information matrix. Robustness checks included using a weighted least squares estimator, using cluster robust variance covariance estimators, and fixed effects for ability based on the Armed Forces Qualification Test (AFQT) score and siblings effects. A second model includes restricting the sample to a within siblings sample. This second model is estimated both with ordinary least squares and instrumental variables estimators. The study concluded no significant differences to returns to education by race and ethnicity.

Consideration should be given to correctly define the treatment effect. For example, one potential operationalization may be a discrete scale on years of completed education.

Since a requirement for the interaction-weighted estimator is to obtain good estimates for the population composition, consider the Census Bureau Educational Attainment - Detailed variable. This variable has the following education attainments: (1) less than 9th grade, (2) 9th - 12th grade, no diploma (3) high school diploma or equivalent, (4) some college, but no degree (5) associate degree, (6) bachelor's degree, (7) master's degree, (8) professional degree, and (9) doctorate degree.

A theoretical argument for using education attainment over years of schooling is that signals in the job market are usually carried through degrees rather than years of schooling. The actual coding for the case study collapses *Less than 9th grade* and *9th - 12th grade, no diploma* conditions as the reference: *Less than high school diploma or equivalent. Some college, but no degree* is coded as *High school diploma or equivalent.* These transformations are consistent with having degrees as the actual signal / milestones. Robustness checks were performed that suggested said schema was appropriate (e.g., similar partial effects).

For estimating the cluster structure, first consider the treatment: educational attainment and the outcome variable: earnings. The outcome variable is operationalized as the natural log of weekly earnings. The mechanism can be argued to be that more education allows for higher productivity, serves as signal in the job market, and gives more options for job-seekers. Hence, what aspects could impact the mechanism? Gender, race, and ethnicity can interact with these mechanism through various paths. For example, statistical discrimination can influence signals through providing a more prominent prior. Glass ceiling discrimination could severely impact access to certain jobs even as education levels should suffice. This analysis provides a theoretical argument as to why race/ethnicity and gender could potentially present heterogeneity in treatment effects for returns to education on earnings.

The second step is to operationalize the features to estimate the cluster structure. Given the racial composition of the US and California, the racial/ethnic composition is defined in terms of White only, Black only, or Latino. These groups are the most prevalent within the population of interest and the US in general. This assumption excludes groups more prevalent

in California than in the US in general such as Asian Americans and other minorities such as Native Americans, other subgroups (e.g., heritage/nationality), and other specifications (e.g., multiple races), but it is only done as a simplification. Likewise, gender is restricted to male and female. For average treatment effects, one should take into consideration the population composition as small shares may be omitted without influencing the average.

The model to estimate is a weighted least squares as following

$$\begin{aligned} \ln(\text{weekly earnings}) = \beta_0 + \\ \beta_1(\text{Education Attainment \& Race and Ethnicity \& Gender}) + \\ \beta_2 \text{ Gender} + \\ \beta_3 \text{ Age} + \\ \beta_4 \text{ (Age)}^2 + \\ u \end{aligned} \tag{2.5}$$

u is the idiosyncratic error term and & is the interaction operator. The main effect of gender is included as a control since gender can influence earnings through other mechanisms other than educational attainment. A second degree polynomial on age is included as a control.

2.8.4 Causal Trees Approach

The treatment, education attainment, has several treatment conditions. These treatment conditions lead to six treatment coding contrasts (i.e., one for each non-reference category). A cluster structure would need to be defined for each treatment condition leading to eight estimated cluster structures. One approach is to perform a set of casual trees to identify most salient differences in partial effects by treatment condition.

The results for the causal trees align with what one would expect. More education is associated with higher earnings. The returns to education follow the expected second

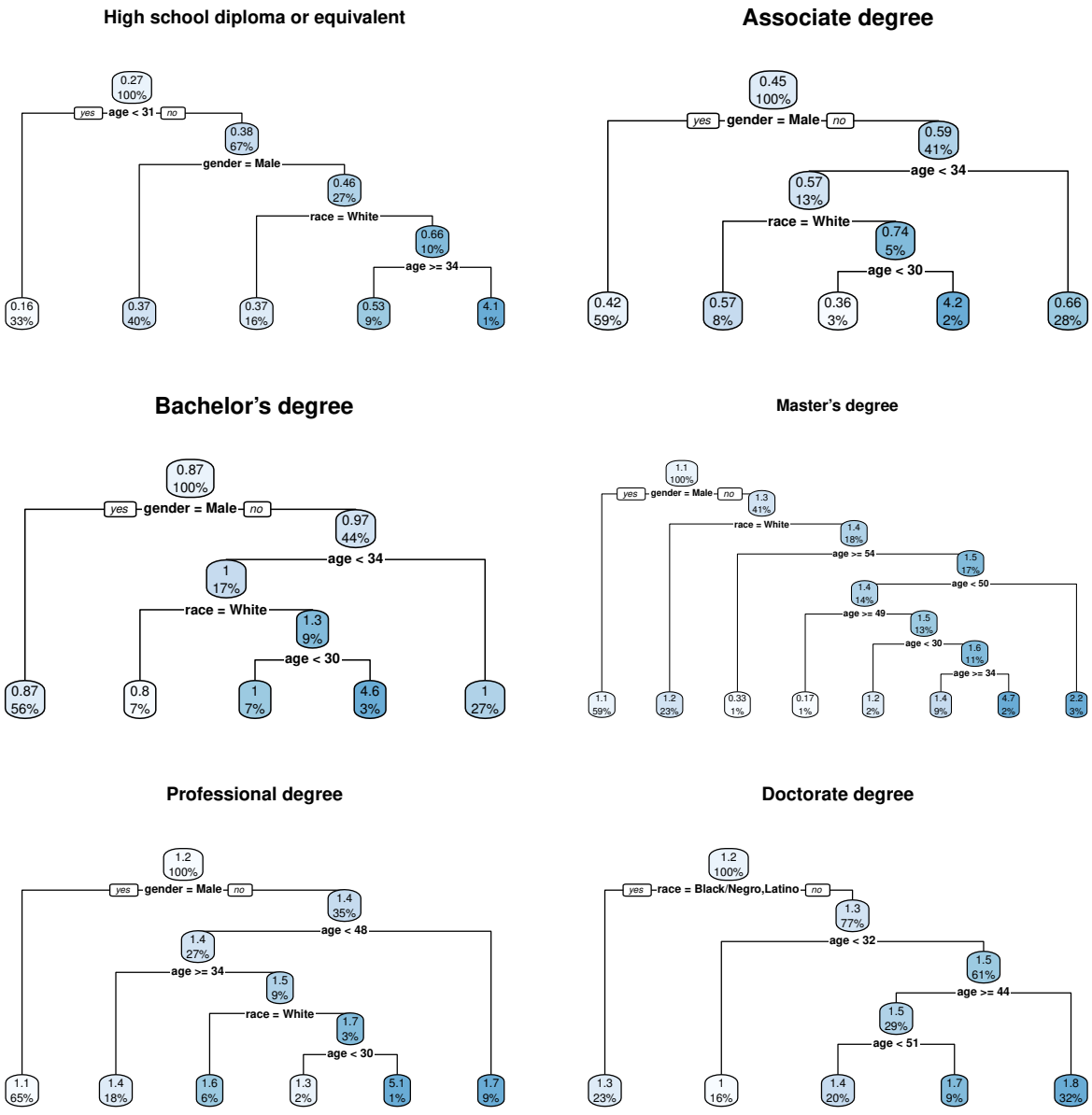

Figure 2.4: *Causal Trees*

degree polynomial relation with first positive and negative second derivatives. In order of salience, gender shows more explanatory power in differences of partial effects than than the race/ethnicity dimension. At the doctorate level race plays a more significant role than gender in explaining the potential heterogeneity in treatment effects which could be driven by race-dominated fields.

We can analyze the cluster structures for each treatment condition through exploring the causal trees. The first treatment condition (second since first is the control/reference), *9th - 12th grade with no diploma*, shows an effect of practical insignificance. The estimated partial effect is very close to zero and thus we could collapse the control and the first treatment effect. That would redefine the treatment variable reference class as less than high school diploma or equivalent.

The causal trees provide evidence consistent with the literature as it suggests a premium for men up to the point of an advanced degree after which the premium disappears. In other words, it pays for women to obtain a degree past bachelor's such as masters. The returns to education differentials at the highest levels of education reach parity. The evidence for a race/ethnicity disparity in returns to education is weaker than by gender. However, it seems that the potential disparity favors White non-Latino.

2.8.5 Estimates

This study will compare the average treatment effect estimates for returns to education using four different models. The models are: (1) pooling, (2) fixed effects, (3) gender cluster robust, and (4) race/ethnicity cluster robust. The models use the specifications given in equations 2.6 - 2.9. Figure 2.5 shows the average treatment effect, returns to education based on treatment condition, education attainment, and uses the population composition estimates from table 2.2. The population composition as expected is quite unbalanced in the racial/ethnicity dimension and more balanced in the gender dimension. Fixed effects are included in the models since race and gender might contribute to earnings through other

mechanisms not necessarily through differences in returns to education. Gender effects were consistently significant while race/ethnicity (i.e., Latino) was jointly significant.

Table 2.2: *Population Composition Estimates*

Education	Race/Ethnicity		Gender	
Attainment	White (Non-Latino)	Black or Latino	Male	Female
High School	81%	19%	57%	43%
Associate	84%	16%	49%	51%
Bachelor	92%	8%	56%	44%
Master	72%	28%	61%	39%
Professional	82%	18%	56%	44%
Doctorate	88%	12%	64%	36%

Tool: CPS Table Generator
Source: Current Population Survey and Annual Social and Economic Supplement, 2018
Description: Adult Civilian Year-Round Workers in California

$$\text{lnweekearn} = \beta_0 + \beta_1 \text{ educ} + \beta_2 \text{ age } + \beta_3(\text{age})^2 + u \tag{2.6}$$

$$\text{lnweekearn} = \beta_0 + \beta_1 \text{ educ} + \beta_2 \text{ age } + \beta_3(\text{age})^2 + \beta_4 \text{ race } + \beta_5 \text{ gender } + u \tag{2.7}$$

$$\text{lnweekearn} = \beta_0 + \beta_1(\text{educ \& gender}) + \beta_2 \text{ age } + \beta_3(\text{age})^2 + \beta_4 \text{ race } + \beta_5 \text{ gender } + u \tag{2.8}$$

$$\text{lnweekearn} = \beta_0 + \beta_1(\text{educ \& race}) + \beta_2 \text{ age } + \beta_3(\text{age})^2 + \beta_4 \text{ race } + \beta_5 \text{ gender } + u \tag{2.9}$$

The first two treatment conditions show very similar average treatment effect estimates across estimators. Similarly, the returns to education for terminal degrees or even advanced degrees are similar, but to a lesser extent. The pooling and within (fixed effects) estimators provide similar estimates suggesting the data obtained uses weights which provide a representative sample (i.e., earning weights are specifically designed for analysis using the earnings variable). When comparing the gender and the race/ethnicity cluster robust mod-

els, the estimates seem to differ the most in the professional degree treatment condition. The difference in returns to education by treatment condition is the largest in professional degrees under a gender or a race/ethnicity cluster structure. Another observation worth of notice is that the cluster-robust estimator based on race/ethnicity shows a larger confidence intervals than the other estimators.

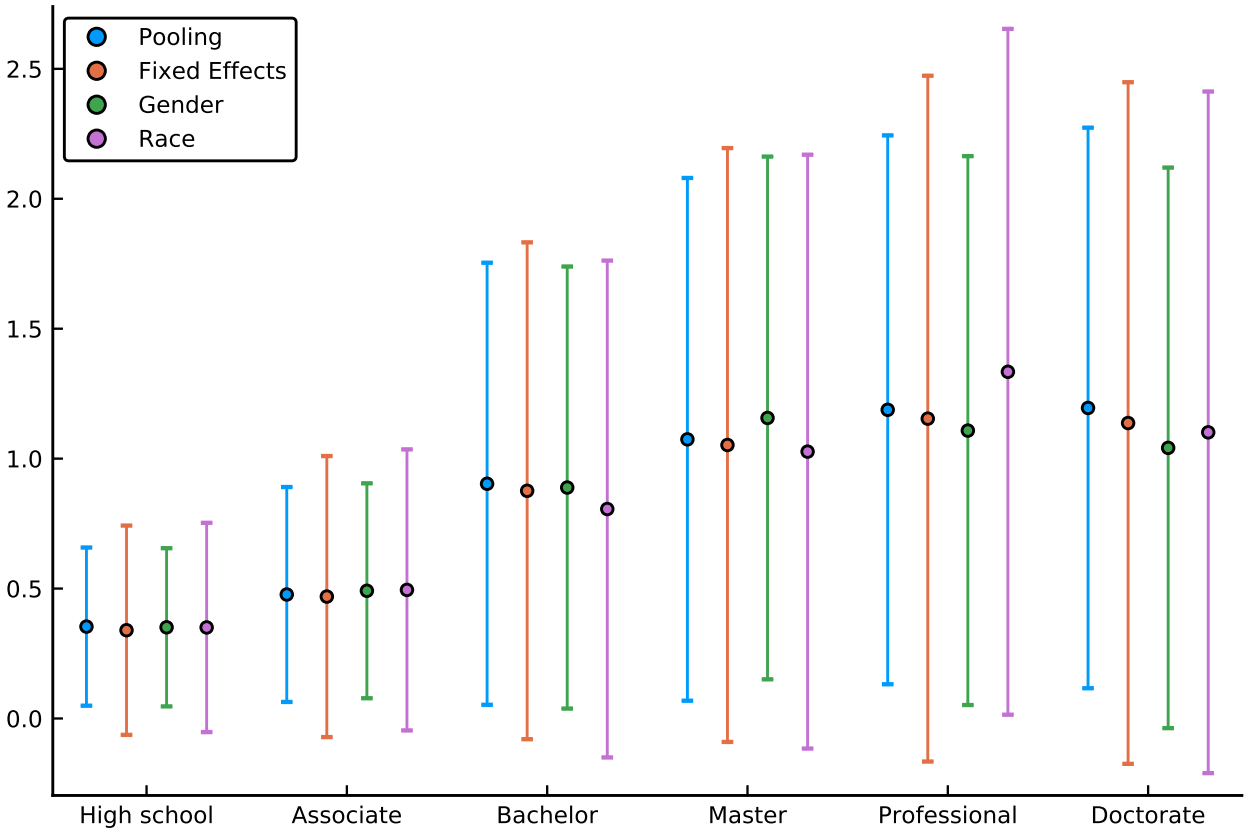

Figure 2.5: *Estimates for Average Treatment Effects (Returns to Education)*

2.9 Conclusion

Cluster robust models offer a way to consistently estimate average treatment effects with potential heterogeneity in treatment effects. Various approaches have been proposed to obtain estimates for cluster structures that range from theoretically driven, corrected inference tests, and data-driven approaches. This study contributes a comprehensive overview to

understand and evaluate cluster structure estimates and their effect in the performance of estimators. The analysis relaxes the unrealistic perfect information assumption and consider the various implications of having an imperfect cluster structure proxy. It shows that proxies can be understood within a framework of purity, level, and dimension. Purity relates to accurately each observation is correctly labeled, level considers if clusters should be aggregated, and the dimension offers tools to diagnose and improve proxies.

The second major contribution of the study is to provide a worked out case study that illustrates a thought process that practitioners can incorporate in their empirical work. One aspect to highlight is to develop a theoretical justification to a cluster structure directly related to identified mechanisms. Practitioners also gain access to data sources and tools (e.g., causal tress) they can use to incorporate in work that can benefit from refined cluster structures. For reporting results, the recommended approach is under uncertainty to report the multiple specification, but also apply a prior from the literature to access which cluster one believe is more likely.

Chapter 3

Econometrics.jl

Software to economists is what hammers are to blacksmiths. As scientists, the tools to work with data are as important as the know-how to perform the analysis. A requirement for good science is that the tools used in the process have a certain standard in terms of quality, transparency, and flexibility. The analysis must allow for the process to be verifiable and replicable.

This study will: (1) provide an overview of common routines and their statistical and technical requirements, (2) provide an overview of tools available with their peculiarities, (3) discuss design decisions in the development of these tools, and (4) showcase *Econometrics.jl* as a new tool for applied work.

Regression models may be used for several purposes. These may provide a basis for prediction models, causal inference, etc. Common targets include confidence intervals of the parameters estimates, joint-significance of a feature, out-of-sample predictive performance, and others. In other words, obtaining the estimates of a model is usually only part of the task. In order to judge a model, one may require to perform diagnostics and tests to justify potential conclusions for an analysis.

Regression analysis is at the core of applied econometrics. It can be challenging to grasp the extent what makes up the broad term of regression analysis. The following describes a

very brief survey of what this technique might entail. Regression analysis can be used for both observational and experimental settings and it allows great flexibility for a multitude of applications. The main idea is to find estimates for model parameters to optimize some objective such as the likelihood in maximum likelihood estimation (MLE). Other potential objectives include restricted maximum likelihood (REML) or a Bayesian approach such as maximum a posteriori probability (MAP). One framework is the generalized linear model (GLM) which use a linear predictor that is mapped through a link function to a distribution modeling the response. Continuous responses might use a Normal distribution, count responses a log link with a Negative Binomial distribution, and probability models might use a categorical distribution with links that map to valid probabilities such as the Logit link. In cases such as probability models where the responses are multidimensional, the generalization is known as vector generalized linear models (VGLM). Other generalizations include relaxing the relation between the linear predictor and the outcome to be the sum of smoothing functions through a generalized additive model (GAM) framework or incorporating random effects through a mixed models approach. Some estimators address challenges such as endogeneity, censored responses, and zero-inflated responses through various solutions such as instrumental variables or censored regression model. Others, exploit aspects of the data to overcome challenges or increase efficiency such as random effects in longitudinal data. In relation to the second moment of the estimator, robust variance covariance estimators or boostrapping may be required for inference.

Out of the many potential tools practitioners may require, what are some of the most common? Not every estimator is as widely accessible or commonly used. Some educated guesses may be well justified such as ordinary least squares being more widely used than spatially-weighted regressions. In order to avoid speculation, I defer to a reasonable assumption that the most common estimators are those usually taught in academic programs and available in widely used software similarly to Renfro (2009). Most programs teach tools to address the most common response types: continuous, count, rates, nominal, and ordinal

outcomes. Hence, some routines might be linear models, Poisson/negative binomial, multinomial logistic regression, and ordinal logistic regression with proportional odds assumption. Topics in time series and panel data are usually offered in most programs. Perhaps the most common topic is short panels (many units of observations through relatively small number of repeated observations). Common estimators include pooling, first-difference, fixed effects / within estimator, and one-way random effects. The between estimator is usually masked as an intermediate model for estimating the error component in the random effects model. Lastly, the two big challenges taught in most programs are endogeneity and heteroscedasticity. These challenges are usually countered through instrumental variables (e.g., 2SLS) or robust variance-covariance estimators (e.g., heteroscedasticity consistent estimators).

Renfro (2009) surveyed the functionality of 24 alternatives for common econometrics routines. Throughout the history of econometrics software, alternatives have risen and fallen in following. Some high contenders by market share include Stata (StataCorp 2017), R (R Core Team 2018), MATLAB, Python (Python Software Foundation 2018), IBM SPSS Statistics, SAS software, and EViews. These include both commercial and open-source alternatives. Functionality may be provided by the base/standard libraries in the statistical software environment, as a product such as a toolkit or user contributed such as a module/package that is distributed. Some examples of user-contributed functionality include the *reghdfe* Stata module and a series of R packages such as *MASS* (Venables and Ripley 2002), *lmtest* (Zeileis and Hothorn 2002), *sandwich* (Zeileis 2004), *plm* (Croissant and Millo 2008), and *mlogit* (Croissant 2018).

The Julia language (Bezanson et al. 2017) is an upcoming language especially well-suited for scientific computing such as econometrics, data science, machine learning, and other related tasks. The following sections describe commons estimators, the Julia ecosystem supporting tools, and Econometrics.jl which provides further functionality.

3.1 Common Estimators

3.1.1 Weighted Least Squares

Weighted least squares solves

$$\beta = \left(X^\top W X\right)^{-1} X^\top W y \tag{3.1}$$

with information matrix

$$\Psi = \left(X^\top W X\right)^{-1} \tag{3.2}$$

where X is the full rank version of a model matrix, W a diagonal matrix with positive values (e.g., frequency), and y the response. The common solution method is to factorize X as either its QR decomposition or Cholesky decomposition. Singular value decomposition may also be used, but it is rare as the computational complexity is significantly higher. In the case of the QR decomposition the solution method comes down to, transforming the model matrix and the response by row-wise multiplying them by the square root of the weights. Afterwards, the factorization is used to solve the system of equations using the appropriate method. In the case of a QR decomposition, R is an upper triangular matrix which enables back substitution to obtain the solution efficiently without matrix inversion. However, a Cholesky decomposition would still be required if the information matrix is desired. The solution method with QR decomposition is delineated in equation 3.3.[1]

The case for the Cholesky decomposition follows closely and without loss of generality other variants could be used such as Bunch-Kaufman decomposition or the upper triangular form ($U^\top U$). The QR decomposition is more numerically stable, but more expensive than Cholesky.[2]. Equation 3.4 delineates the solution method with Cholesky decomposition. Since the information matrix is an important component, Bunch-Kaufman decomposition, a Cholesky variant, is the preferred method used in *Econometrics.jl*.

1. The refers multiplication of b by the inverse of A on the left
2. $\mathcal{O}\left(n^3\right) > \mathcal{O}\left(2mn^2 - \frac{2}{3}2n^3\right)$ where the matrix has m rows and n columns

$$\begin{aligned} \tilde{X} &= X. * \sqrt{w} \\ \tilde{y} &= y. * \sqrt{w} \\ QR &= \tilde{X} \\ \beta &= R \setminus \left(Q^\top \tilde{y}\right) \end{aligned} \tag{3.3}$$

$$\begin{aligned} LL &= \left(X^\top W X\right) \\ \beta &= L \setminus \left(X^\top W y\right) \\ \Psi &= \left(L^{-1}\right)^\top L^{-1} \end{aligned} \tag{3.4}$$

The remaining estimators will assume a Cholesky decomposition as part of the estimation technique. The QR decomposition will be used in models estimated through iterative reweighted least squares (IRLS) as the factorization may be computed once and recycled.

3.1.2 Within Estimator

The within estimator is an application of the Frisch-Waugh-Lovell theorem (Frisch and Waugh 1933; Lovell 2008). The estimator allows to compute the parameters estimates and information matrix for a subset of predictors without having to include the full set of categorical features. For example, one may include individual fixed effects in a large data set that may increase the dimension of the model matrix to several thousand making the problem unfeasible or inefficient. Moreover, some parameters may not be consistently estimated in certain contexts. For example, individual fixed effects are not consistently estimated when there is a fixed length for the panels (i.e., more observations implies more parameters).

Consider the following model,

$$y = X\beta + D\theta + e \tag{3.5}$$

where y is the response, β the parameters of interest, X the features of the parameters of interest, D a high dimensional representation of categorical features as control, θ the parameters on said covariates, and e the error term. In order to obtain the parameter estimates β and the associated information matrix, we can estimate an alternative specification.

$$\tilde{y} = \tilde{X}\beta + e \tag{3.6}$$

where $\tilde{X}$ and $\tilde{y}$ are obtained by using projections, such as the annihilator matrix (i.e., $I - X\left(X^\top W X\right)^{-1} X^\top$). There are several methods to obtain a suitable alternative regression and these are not unique. Correia (2017) presents several approaches to solving these problem including specialized methods in certain applications. Some implementations include *reghdfe*, Stata module, and Matthieu Gomez *FixedEffectModels.jl* package. The two most common approaches are solving for the residuals through a sparse least-squares problems such as with LSMR (Fong and Saunders 2011) or using some variant for the method of alternating projections. The residuals approach tends to be more efficient, but degrades certain aspects of the model (e.g., no longer able to obtain the mean response). The method of alternating projections is able to preserve under certain conditions artifacts of the original regression such as obtaining the same estimate for the intercept even though it is not particularly meaningful.

3.1.3 Between Estimator

The between estimator estimates

$$\tilde{y} = \tilde{X}\beta + e \tag{3.7}$$

where the transformed model components are collapsed through some dimension through the mean function. For example, one approach to obtaining the error component for a random effects model is to use model statistics of the between estimator collapsing by panel. The weighted version of the model uses the observation weights to compute the weighted mean values and may use the weight fractions by the collapsing dimension as weights for the weighted least squares regression on the transformed model.

3.1.4 Random Effects Model

The random effects model relies in estimating the unobserved error components. Random effects requires a particular schema for the data which has a panel component and a temporal component. There are multiple approaches, but the most common one is the Swamy-Arora approach (Swamy and Arora 1972). This estimator uses the mean squared residuals estimates (i.e., deviance divided by residual degrees of freedom) of the between and within models using the panel dimension as the collapsing / dimension to absorb. The error components are estimated as

$$\begin{aligned} \theta_g &= 1 - \sqrt{\frac{\sigma_e^2}{T_g * \sigma_u^2 + \sigma_e^2}} \\ \sigma_e^2 &= W \\ \sigma_u^2 &= \max\left\{0, B - \sigma_e^2 * \bar{T}\right\} \end{aligned} \tag{3.8}$$

where W is the mean squared residuals of the within model, B the mean squared residuals of the between model, and T_g is the length of the panel g, and $\bar{T}$ is the harmonic mean of the panel lengths.

The model terms are then transformed by partial demeaning

$$\begin{aligned} \tilde{y}_{it} &= y_{it} - \theta_g * \bar{y}_{.t} \\ \tilde{X}_{it} &= X_{it} - \theta_g * \bar{X}_{.t} \end{aligned} \tag{3.9}$$

and these are used in the standard regression setting.

3.1.5 First-Difference

The first-difference estimator is a special case that use time / panel context for feature designs. The most common transformations include contrasts such as treatment coding (dummy coding), sum coding (effects coding) or Helmert coding which apply to categorical variables. Other common feature engineering techniques include log-transform and polynomial terms. However, certain transformations require a context such as a time dimension. Some examples include shift operations (lag, lead) and differentiating (e.g., first-difference). These operations may optionally require a group context such that the operations are performed group wise. Time-context operations have important concepts such as frequency and gaps. The frequency describes the difference between periods/observations and gaps describe observations that are skipped and should be understood as missing.

3.1.6 Instrumental Variables

Every estimator thus far can be generalized to include endogenous covariates through instrumental variables. The most common method is through two stages least squares (2SLS). The idea is to first apply all the relevant transformations to the model terms and apply the 2SLS standard procedure. In the case of the random effects model, the within and between models are estimated using 2SLS to obtain the error component estimates. After applying the random effects transformation to each model term the 2SLS process is employed in the final regression model à la Balestra and Varadharajan-Krishnakumar (1987).

The standard 2SLS estimator uses,

$$\begin{aligned} \hat{z} &= [XZ] \left[\left([XZ]^\top W [XZ] \right)^{-1} [XZ]^\top W z \right] \\ \hat{\beta} &= \left([X\hat{z}]^\top W [X\hat{z}] \right)^{-1} [X\hat{z}]^\top W y \\ \Psi &= \left([X\hat{z}]^\top W [X\hat{z}] \right)^{-1} \\ \hat{y} &= [Xz]\hat{\beta} \end{aligned} \tag{3.10}$$

for each model where z is endogenous variables and Z the additional instruments.

3.1.7 Nominal Response Model

Multinomial logistic regression is a probability model for estimating probabilities across multiple categories. It is a vector generalized linear model with softmax link function and the categorical distribution. It is estimated through iterative re-weighted least squares (IRLS) methods such as the QR Newton variant (O'Leary 1990). The data schema for discrete choice models include the response (observed behavior), unit of observation covariates, and outcomes-specific covariates. The initial implementation allows for the base case of no-outcome specific features.

3.1.8 Ordinal Response Model

Ordinal logistic regression is a probability model for estimating probabilities across multiple ordered categories. Similarly to its nominal counterpart, it has a pool of alternatives, and observed outcome, unit of observation covariates, and outcome-specific covariates. A common assumption is the proportional odds assumption which may be relaxed in other models.

The log-likelihood function has the same form as the general form for computing the cost associated with a categorical distribution and predicted probability for realization. More

specific,

$$\ell\ell = \sum_{i=1}^{m}\sum_{k}^{K} \mathbb{1}\left(y_i = k\right) \ln\left[F\left(\alpha_{k+1} - \eta\right) - F\left(\alpha_k - \eta\right)\right] \tag{3.11}$$

where F is the cumulative distribution function of the logistic distribution with zero location and unit scale, η is the linear projection, and α_k is the threshold for lower threshold (McKelvey and Zavoina 1975). The log-likelihood function and the gradient are passed to the *Optim.jl* framework (K Mogensen and N Riseth 2018) using *ForwardDiff.jl* (Revels, Lubin, and Papamarkou 2016) forward mode automatic differentiation (AD) for the Newtonian solver.

3.1.9 Count/Rate Model

Count/rate models are generalized linear models and follow a similar description as nominal models. The most common distribution choices are Poisson and Negative Binomial with the log link function. Negative Binomial is a generalization of the Poisson model, which adds an extra parameter for modeling the second moment (i.e., relaxes the mean equal variance assumption in the Poisson model). For the Negative Binomial to be a distribution in the exponential family it needs a restriction parameter which may be optimized through maximum likelihood estimation. An offset may be included to handle rates, a generalization of counts, that account for differences in exposures. Other generalizations include additive or multiplicative errors relations.

3.1.10 Duration Models

Duration models deal with responses of the type time until an event. One such model is the Cox proportional hazards model which relies on the proportional hazards assumption. Various models of these kind may be re-specified in a generalized linear model framework relating to the previous descriptions.

3.2 Technical Challenges

One technical challenge that is prevalent through every model is the issue of rank deficient terms. Rank deficient systems of linear equations are not identifiable. One approach is to error out and let the user explore and find a subset of features such that the no multi-collinearity assumption holds. The second approach is to automatically promote the system to a full rank version by excluding linearly dependent features. How much collinearity is too much is not an exact science. Some potential criteria include using the absolute values of the diagonal in the triangular matrix of the factorization (e.g., L in $LL^\top$, R in QR, D in $LDL^\top$, Σ in $U\Sigma V^\top$). These values are then compared against a chosen tolerance and the column of the term is deemed linearly independent if the values are greater than the tolerance. Note that Cholesky, QR, and Bunch-Kaufman decomposition allow to identify which columns are independent while singular values only allow to determine the rank. It may be arbitrary to choose among linearly dependent features. An additional level of complexity in probability models is the issue of linearly separability. Konis (2007) provides an overview of potential approaches to identifying the issue.

3.3 Julia Ecosystem

The usual pipeline for regression analysis involves (1) accessing data (I/O), (2) obtaining a tabular data representation, (3) data wrangling, and (4) employing regression analysis tools. The Julia ecosystem follows this canonical pipeline. The following sections provides an overview of the pipeline available in Julia.

3.3.1 Data to Modeling

StatsBase.jl builds on top of Statistics.jl (standard library) to provide additional statistical functionality. One which includes the abstraction for Statistical Models (and Regression models which inherit from the former). It provides a simple and powerful API for the whole

Julia ecosystem to use. It allows packages to implement the API and easily support a common functionality users can expect and interact with in a familiar manner. For example, *coef* will extract the parameter estimates from any object that implements the API. The full API include model statistics such as: coefficient of determination (or adjusted), information criteria statistics such as AIC/BIC (and corrected), statistics about the model fitness such as deviance, log-likelihood, and usual queries such as point estimates, variance covariance estimates, standard errors, confidence intervals, degrees of freedom (or residual degrees of freedom), etc. Lastly, several accessors are available for fitted values, response, model matrix, information matrix, leverage values, error components, etc. Lastly, it also provides an abstraction for weights including frequency weights and analytical weights.

Tables.jl provide an interface for tabular data. This API allows users to choose from various solutions the tabular data implementation of their choosing without having to worry that their choice will limit potential functionality. Many tabular implementations such as *DataFrames.jl* provide robust functionality to many routines such as handling categorical features, dates/time, missing values, reshaping data, split/apply operations, and others. Users need not to worry about any I/O issues as a rich array of options exist for importing and exporting across different file formats such as delimiter-separated values, JSON, Feather, HDF5, MATLAB, Stata, SPSS, SAS, and R.

StatsModels.jl is a package that provides the means to go from data to model terms. It provides the formulae language (e.g., similar syntax to R's formulae syntax). A model is then build using a formula, data, and additional model specific arguments. The process can be summarized as (1) collecting the information in the formula, (2) parsing its meaning by applying a schema based on the data, user-specified contrasts or other arguments, and (3) generating the model terms such as a response, model matrices, etc. Lastly, a package fits said model and implements the API.

3.3.2 Regression Analysis

The regression analysis ecosystem in Julia has *GLM.jl* as its flagship. *GLM.jl* provides the typical functionality for fitting generalized linear models through Fisher scoring. This includes linear models, Poisson/Negative Binomial, Logit/Probit, and other non-canonical link models. *CovarianceMatrices.jl* provides various variance covariance estimators for *GLM.jl* models à la R's *sandwich* package. *LinearMixedModels.jl* (Bates et al. 2019) extends *GLM.jl* for mixed-effects models. *FixedEffectModels.jl* provides fast estimation of linear models with instrumental variables and high dimensional categorical variables à la *reghdfe*. *Survival.jl* provides a series of estimators for duration models. Two major gaps in the ecosystem include estimating nominal and ordinal response models (i.e., discrete choice) with more than two alternatives and support for longitudinal estimators.

3.4 Econometrics.jl

Econometrics.jl is a package for performing several common econometrics routines in the Julia language. It aims to provide the following functionality for two major gaps in the ecosystem, longitudinal estimators and discrete choice models. Developing the package has resulted in many contributions in the current ecosystem. However, the development of this package serves multiple purposes beyond the immediate effect. As the statistics ecosystem evolves and matures, *Econometrics.jl* aims to serve as inspiration and an alternative to design decisions, standards, and option for user.

3.4.1 Fitting Models

This section will showcase some examples of using the package for various estimators. For each estimator a brief description of the data, model, syntax, and output will be provided. Results will be provided for Econometrics.jl and some alternatives such as R or Stata.

For linear models, the examples use the crime data set from Cornwell and Trumbull

(1994). The data set is a balanced longitudinal data set with 90 counties in North Carolina from 1981 to 1987. The outcome variable is the crime rate and the explanatory variables include the probability of conviction, average sentence, and probability of prison sentence. Estimating the pooling estimator or between estimators can be accomplished as in figure 3.1. Table 3.1 shows the estimated 95% confidence intervals using Econometrics.jl, Stata, and R's plm package.

```
julia> fit(EconometricModel,
           @formula(CRMRTE ~ PrbConv + AvgSen + PrbPris),
           data)
Continuous Response Model
Number of observations: 630
Null Loglikelihood: 1633.30
Loglikelihood: 1643.30
R-squared: 0.0312
LR Test: 19.99 ~ χ²(3) ⟹ Pr > χ² = 0.0002
Formula: CRMRTE ~ PrbConv + AvgSen + PrbPris
                 PE          SE     t-value Pr > |t|         2.5%       97.5%
(Intercept)    0.0186413  0.00431066  4.32447    <1e-4    0.0101762   0.0271064
PrbConv      -0.00116473 0.000422062 -2.75961   0.0060  -0.00199356  -0.0003359
AvgSen       0.000236185 0.000268216  0.88058   0.3789 -0.000290526 0.000762897
PrbPris        0.0273394  0.00817642  3.34369   0.0009    0.0112829   0.0433959


julia> fit(EconometricModel,
           @formula(CRMRTE ~ PrbConv + AvgSen + PrbPris + between(County)),
           data)
Between Estimator
County with 630 groups
Balanced groups with size 7
Number of observations: 90
Loglikelihood: 244.40
R-squared: 0.1029
Wald: 3.29 ~ F(3, 86) ⟹ Pr > F = 0.0246
Formula: CRMRTE ~ PrbConv + AvgSen + PrbPris + :(between(County))
                 PE         SE     t-value Pr > |t|        2.5%      97.5%
(Intercept)  -0.00464339  0.0186571 -0.24888   0.8040  -0.0417325  0.0324457
PrbConv      -0.00354113  0.0018904 -1.87322   0.0644 -0.00729911 0.00021685
AvgSen       0.000848049 0.00116477 0.728086   0.4685 -0.00146743 0.00316353
PrbPris        0.0730299  0.0332057  2.19932   0.0305   0.0070191   0.139041
```

Figure 3.1: *Estimation of the pooling and between panel estimator*

Table 3.1: *Pooling and Between Estimators*

Model	Parameter	Econometrics.jl		Stata		R (plm)	
Pooling	Intercept	0.0102	0.0271	0.0102	0.0271	0.0102	0.0271
	PrbConv	-0.0020	-0.0003	-0.0020	-0.0003	-0.0020	-0.0003
	AvgSen	-0.0003	0.0008	-0.0003	0.0008	-0.0003	0.0008
	PrbPris	0.0113	0.0434	0.0113	0.0434	0.0113	0.0434
Between	Intercept	-0.0417	0.0324	-0.0417	0.0324	-0.0412	0.0319
	PrbConv	-0.0073	0.0002	-0.0073	0.0002	-0.0072	0.0002
	AvgSen	-0.0015	0.0032	-0.0015	0.0032	-0.0014	0.0031
	PrbPris	0.0070	0.1390	0.0070	0.1390	0.0079	0.1381

The fixed effects model or within estimator can be estimated as in figure 3.2 which estimates the panel effects and the two-ways fixed effects model (i.e., fixed effects for time dimension as well). Table 3.2 shows the estimated 95% confidence intervals using Econometrics.jl, Stata, and R's plm package.

Table 3.2: *Absorbing Panel or Panel and Temporal Indicators*

Model	Parameter	Econometrics.jl		Stata (reghdfe)		R (plm)	
Within PID	Intercept	0.0274	0.0355	0.0274	0.0355		
	PrbConv	-0.0004	0.0004	-0.0004	0.0004	-0.0004	0.0004
	AvgSen	-0.0002	0.0003	-0.0002	0.0003	-0.0002	0.0003
	PrbPris	-0.0093	0.0066	-0.0093	0.0066	-0.0093	0.0066
Within PTID	Intercept	0.0279	0.0360	0.0279	0.0360		
	PrbConv	-0.0003	0.0005	-0.0003	0.0005	-0.0003	0.0005
	AvgSen	-0.0004	0.0002	-0.0004	0.0002	-0.0004	0.0002
	PrbPris	-0.0070	0.0089	-0.0070	0.0089	-0.0069	0.0089

```
julia> fit(EconometricModel,
           @formula(CRMRTE ~ PrbConv + AvgSen + PrbPris + absorb(County)),
           data)
Continuous Response Model
Number of observations: 630
Loglikelihood: 2273.95
R-squared: 0.0009
Wald: 0.16 ~ F(3, 537) ⟹ Pr > F = 0.9258
Formula: CRMRTE ~ PrbConv + AvgSen + PrbPris + :(absorb(County))
                 PE          SE      t-value Pr > |t|         2.5%       97.5%
(Intercept)   0.0314527  0.00204638   15.3699   <1e-43    0.0274328   0.0354726
PrbConv      6.65981e-6   0.0001985 0.0335507   0.9732 -0.000383272 0.000396592
AvgSen       7.83181e-5 0.000127904  0.612318   0.5406 -0.000172936 0.000329572
PrbPris      -0.0013419  0.00405182 -0.331185   0.7406  -0.00930126  0.00661746


julia> fit(EconometricModel,
           @formula(CRMRTE ~ PrbConv + AvgSen + PrbPris + absorb(County + Year)),
           data)
Continuous Response Model
Number of observations: 630
Loglikelihood: 2290.25
R-squared: 0.0013
Wald: 0.23 ~ F(3, 531) ⟹ Pr > F = 0.8767
Formula: CRMRTE ~ PrbConv + AvgSen + PrbPris + :(absorb(County + Year))
                  PE          SE      t-value Pr > |t|         2.5%       97.5%
(Intercept)    0.0319651  0.00206378   15.4886   <1e-44     0.027911   0.0360193
PrbConv       7.90362e-5 0.000196089  0.403062   0.6871  -0.00030617 0.000464242
AvgSen        -9.4788e-5 0.000134924 -0.702531   0.4827 -0.000359838 0.000170262
PrbPris      0.000979533   0.0040432  0.242267   0.8087  -0.00696309  0.00892216
```

Figure 3.2: *Estimation of the within estimator*

The random effects model can be estimated with similar syntax by including the *PID* (panel identifier) and *TID* (temporal identifier) tags such as in figure 3.3 which shows estimating a random effects model as well as its instrumental variables counterpart. Table 3.3 shows the estimated 95% confidence intervals using Econometrics.jl, Stata's reghdfe module, and R's plm package.

```
julia> fit(EconometricModel,
           @formula(CRMRTE ~ PrbConv + AvgSen + PrbPris + PID(County) + TID(Year)),
           data)
One-way Random Effect Model
Longitudinal dataset: County, Year
Balanced dataset with 90 panels of length 7
individual error component: 0.0162
idiosyncratic error component: 0.0071
ρ: 0.8399
Number of observations: 630
Loglikelihood: 2226.01
R-squared: 0.0007
Wald: 0.15 ~ F(3, 626) ⟹ Pr > F = 0.9291
Formula: CRMRTE ~ PrbConv + AvgSen + PrbPris + :(PID(County)) + :(TID(Year))
                    PE          SE      t-value Pr > |t|         2.5%       97.5%
(Intercept)     0.0309293  0.00266706   11.5968   <1e-27    0.0256918   0.0361668
PrbConv       -3.96634e-5 0.000198471 -0.199845   0.8417 -0.000429413 0.000350086
AvgSen          8.2428e-5 0.000127818  0.644887   0.5192 -0.000168575 0.000333431
PrbPris      -0.000123327  0.00404272 -0.030506   0.9757  -0.00806226  0.00781561


julia> fit(EconometricModel,
           @formula(CRMRTE ~ PrbConv + (AvgSen ~ PrbPris) + PID(County) + TID(Year)),
           data)
One-way Random Effect Model
Longitudinal dataset: County, Year
Balanced dataset with 90 panels of length 7
individual error component: 0.0413
idiosyncratic error component: 0.0074
ρ: 0.9691
Number of observations: 630
Loglikelihood: 2248.34
R-squared: NaN
Wald: 0.03 ~ F(2, 626) ⟹ Pr > F = 0.9671
Formula: CRMRTE ~ PrbConv + AvgSen ~ PrbPris + :(PID(County)) + :(TID(Year))
                     PE          SE      t-value Pr > |t|         2.5%       97.5%
(Intercept)      0.037747   0.0241684   1.56183   0.1188  -0.00971405    0.085208
PrbConv       1.39581e-5 0.000200244 0.0697054   0.9445 -0.000379273 0.000407189
AvgSen       -0.000688924  0.00266514 -0.258494   0.7961  -0.00592262  0.00454478
```

Figure 3.3: *Estimation of the random effects model*

Table 3.3: *Random Effects and Instrumental Variables*

Model	Parameter	Econometrics.jl		Stata		R (plm)	
Random	Intercept	0.0257	0.0362	0.0257	0.0362		
	PrbConv	-0.0004	0.0004	-0.0004	0.0003	-0.0004	0.0004
	AvgSen	-0.0002	0.0003	-0.0002	0.0003	-0.0002	0.0003
	PrbPris	-0.0081	0.0078	-0.0080	0.0078	-0.0093	0.0066
IV Random	Intercept	-0.0097	0.0852	-0.0096	0.0851	-0.0096	0.0851
	PrbConv	-0.0004	0.0004	-0.0004	0.0004	-0.0004	0.0004
	AvgSen	-0.0059	0.0045	-0.0059	0.0045	-0.0059	0.0045

The sysdsn1 Stata example health insurance data set is used to illustrate the multinomial logistic regression when the response is nominal as seen in figure 3.4. A comparison with the estimates for the 95% confidence intervals between Econometrics.jl and Stata is shown in table 3.4.

```
julia> fit(EconometricModel,
           @formula(insure ~ age + male + nonwhite + site),
           data)
Probability Model for Nominal Response
Categories: Indemnity, Prepaid, Uninsure
Number of observations: 615
Null Loglikelihood: -555.85
Loglikelihood: -534.36
R-squared: 0.0387
LR Test: 42.99 ~ χ²(10) ⟹ Pr > χ² = 0.0000
Formula: insure ~ age + male + nonwhite + site
                                  PE         SE      t-value Pr > |t|       2.5%       97.5%
insure: Prepaid ~ (Intercept)   0.269725   0.328443  0.821223   0.4118  -0.375302    0.914752
insure: Prepaid ~ age         -0.0117451 0.00619461  -1.89602   0.0584 -0.0239107 0.000420447
insure: Prepaid ~ male          0.561703   0.202747   2.77046   0.0058   0.163529    0.959877
insure: Prepaid ~ nonwhite       0.97479   0.236322   4.12483    <1e-4   0.510678      1.4389
insure: Prepaid ~ site: 2       0.113027   0.210191  0.537735   0.5910  -0.299765    0.525819
insure: Prepaid ~ site: 3      -0.587996   0.227936  -2.57966   0.0101   -1.03564   -0.140355
insure: Uninsure ~ (Intercept)  -1.28693   0.592315  -2.17271   0.0302   -2.45018   -0.123689
insure: Uninsure ~ age        -0.00779625 0.0114417 -0.681388   0.4959 -0.0302666   0.0146741
insure: Uninsure ~ male         0.451858   0.367481   1.22961   0.2193  -0.269836     1.17355
insure: Uninsure ~ nonwhite      0.21707   0.425628  0.510001   0.6102  -0.618817     1.05296
insure: Uninsure ~ site: 2      -1.21157    0.47051  -2.57502   0.0103    -2.1356   -0.287539
insure: Uninsure ~ site: 3     -0.207819   0.366289 -0.567364   0.5707  -0.927171    0.511533
```

Figure 3.4: *Estimation of the multinomial logistic regression*

Table 3.4: *Multinomial Logistic Regression*

Response	Parameter	Econometrics.jl		Stata	
Indemnity	(Intercept)	-0.3753	0.9148	-0.3740	0.9134
	Age	-0.0239	0.0004	-0.0239	0.0004
	Gender: Male	0.1635	0.9599	0.1643	0.9591
	Nonwhite	0.5107	1.4389	0.5116	1.4380
	Site: 2	-0.2998	0.5258	-0.2989	0.5250
	Site: 3	-1.0356	-0.1404	-1.0347	-0.1412
Prepaid	(Intercept)	-2.4502	-0.1237	-2.4479	-0.1260
	Age	-0.0303	0.0147	-0.0302	0.0146
	Gender: Male	-0.2698	1.1736	-0.2684	1.1721
	Nonwhite	-0.6188	1.0530	-0.6172	1.0513
	Site: 2	-2.1356	-0.2875	-2.1338	-0.2894
	Site: 3	-0.9272	0.5115	-0.9257	0.5101

The fullauto Stata example automobile models data set is used to illustrate the proportional ordinal logistic regression when the response is ordinal as seen in figure 3.5. A comparison with the estimates for the 95% confidence intervals between Econometrics.jl, Stata, and R's MASS is shown in table 3.5.

Table 3.5: *Parallel Ordinal Logistic Regression*

Parameter	Econometrics.jl		Stata		R's MASS	
Foreign	1.3168	4.4768	1.3472	4.4464	1.4111	4.5293
Length	0.0374	0.1282	0.0383	0.1274	0.0395	0.1292
MPG	0.0900	0.3716	0.0927	0.3689	0.0986	0.3781
(Intercept): Poor | Fair	6.8343	29.0206	7.0473	28.8076		
(Intercept): Fair | Average	8.6814	31.0487	8.8962	30.8340		
(Intercept): Average | Good	10.6949	33.5117	10.9140	33.2926		
(Intercept): Good | Excellent	12.9204	36.4639	13.1465	36.2378		

```
julia> fit(EconometricModel,
           @formula(rep77 ~ foreign + length + mpg),
           data)
Probability Model for Ordinal Response
Categories: Poor < Fair < Average < Good < Excellent
Number of observations: 66
Null Loglikelihood: -89.90
Loglikelihood: -78.25
R-squared: 0.1295
LR Test: 23.29 ~ χ²(3) ⟹ Pr > χ² = 0.0000
Formula: rep77 ~ foreign + length + mpg
                                PE        SE    t-value Pr > |t|      2.5%   97.5%
foreign: Foreign              2.89681  0.790641 3.66387   0.0005   1.31684 4.47678
length                      0.0828275   0.02272 3.64558   0.0005 0.0374253 0.12823
mpg                          0.230768 0.0704548  3.2754   0.0017 0.0899749 0.37156
(Intercept): Poor | Fair      17.9275   5.55119 3.22948   0.0020   6.83431 29.0206
(Intercept): Fair | Average   19.8651   5.59648 3.54956   0.0007   8.68139 31.0487
(Intercept): Average | Good   22.1033   5.70894  3.8717   0.0003   10.6949 33.5117
(Intercept): Good | Excellent 24.6921   5.89075 4.19168    <1e-4   12.9204 36.4639
```

Figure 3.5: *Estimation of the proportional odds logistic regression*

3.4.2 Design Decisions

Statistical software developers play a very powerful role in shaping culture and norms. For example, whether to default to maximum likelihood estimation (MLE) or restricted maximum likelihood estimation (REML), can shape not only choices by practitioners, but by stakeholders, regulatory agencies, and expected components for reports. These changes may be good or bad depending on the case. For example, advances in econometrics are rarely widely adopted without buy-in from software developers. The following discussions will survey survey some of the decisions relevant to *Econometrics.jl*.

Should software be dummy-proof? Many times software developers have to choose between exposing users to make mistakes on their own volition or put safeguard against potential misuses by restricting behavior that may be correct under rare scenarios. For example, a basic tool might allow users to mix and match link and distributions in a GLM settings even if the combinations are nonsensical. A safer approach would be to restrict combinations to those "safe" combinations such as distributions with canonical links. The trade-off occurs when users may encounter a specification that while uncommon it may be the correct one for that particular model. Currently, *Econometrics.jl* takes a conservative approach that pro-

vides "dummy-proof" experience as well as ease of use. For example, rather than requiring users to specify the model, the estimator is inferred based on the type of the response and information provided. Other examples of this approach include making model statistics such as promoting the coefficient of determination to a pseudo-version for non-linear models (a generalization of the linear case), but providing a *not a number* (NaN) value for instrumental variable models or those that do not include an intercept. Similarly, the software will promote terms to full rank as required and inform of the behavior.

Software analysis should always include diagnostics and tools to make it easier for dissemination. Many tests and diagnostics are applicable to a wide set of implementations. The best manner to make these available and for these to "play nicely" with one another is to have an effective API. Sadly, the Julia ecosystem has yet to experience wide adoption of this pattern. For example, packages might need to access components for computing a test or providing some estimates such as variance covariance estimates. The test might include components such as the residual degrees of freedom, the information matrix, residuals, and score. Many implementations might access internals of one particular implementation and compute these such as obtaining the model matrix, response and coefficients to compute the residuals. This behavior would produce incorrect estimates in cases of instrumental variable as the fitted / linear projection should not use the model matrix, but replace the projection of endogenous features with their actual values. A better approach would be to request the response and fitted values to compute the residuals. This approach is less prone to errors, but it can be more robust by calling residuals directly. Specifying precise components can alleviate risk of errors or relying on assumptions such as whether the components are weighted or should be weighted for the procedure.

Various decisions are software specific with asymptotic justification, but significant finite-sample consequences. Software may differ on whether to report statistics using finite-sample statistics (t-distribution, F-distribution) or asymptotic equivalent counterparts (Normal, Chi squared). These tend to have negligible effect in most applications, but other decisions such

as degrees of freedom may have larger consequences. For example, software may differs on how it computes the degrees of freedom for instrumental variables or absorbed variables depending on the context (e.g., main regression or auxiliary regression for estimating error components). Refinements and robustness checks can also contribute to a better analysis such as verifying gaps for time variant operations such as first-difference or purging singletons and other degree of freedom adjustments à la Correia (2015).

3.4.3 Best Practices

Econometrics.jl adopts the best practices standards for open-source statistical software. These include adhering to semantic versioning (semver) for descriptive versioning, continuous integration for development, software validation through a comprehensive code coverage and test suite, and lastly online hosted documentation for the public API.

3.5 Conclusion

Econometrics.jl is a new addition to the Julia ecosystem that brings highly demanded functionality concerning longitudinal estimators and discrete choice models. This study serves as a complement to the software documentation providing context to the development, design considerations, and roadmap of the project. A philosophical motivation for the project is to make econometrics accessible to practitioners not only through functionality, but transparency in the code readability, replicability, and correctness. For example, transparent well-written code is easier to maintain, inspect / audit, and can be useful for learning and teaching.

Community contributions and feedback are highly encouraged in order to best continue developing the project. The release will be available at the Github repository and licensed under a permissive license.

Bibliography

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. 2017. *When Should You Adjust Standard Errors for Clustering?* Working Paper, Working Paper Series 24003. National Bureau of Economic Research. doi:10.3386/w24003. (Cited on pages 3, 8).

Arellano, Manuel. 1987. "Computing Robust Standard Errors for Within-groups Estimators." *Oxford Bulletin of Economics and Statistics* 49 (4): 431–434. doi:10.1111/j.1468-0084.1987.mp49004006.x. (Cited on page 8).

Athey, Susan, and Guido W Imbens. 2016. "Recursive partitioning for heterogeneous causal effects." *Proceedings of the National Academy of Sciences* 113, no. 27 (July 5): 7353–7360. ISSN: 0027-8424, 1091-6490. doi:10.1073/pnas.1510489113. (Cited on page 20).

Athey, Susan, and Guido W. Imbens. 2016. "The Econometrics of Randomized Experiments." *ArXiv e-prints.* arXiv: 1607.00698 [stat.ME]. https://arxiv.org/abs/1607.00698. (Cited on pages 3, 5).

Athey, Susan, and Guido W Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31, no. 2 (May): 3–32. doi:10.1257/jep.31.2.3. (Cited on page 20).

Balestra, Pietro, and Jayalakshmi Varadharajan-Krishnakumar. 1987. "Full Information Estimations of a System of Simultaneous Equations with Error Component Structure." *Econometric Theory* 3 (2): 223–246. doi:10.1017/S0266466600010318. (Cited on page 46).

Barrow, Lisa, and Cecilia Elena Rouse. 2005. "Do Returns to Schooling Differ by Race and Ethnicity?" *American Economic Review* 95, no. 2 (April): 83–87. ISSN: 0002-8282. doi:10.1257/000282805774670130. (Cited on page 31).

Bates, Douglas, José Bayoán Santiago Calderón, Dave Kleinschmidt, Tony Kelman, Simon Babayan, Patrick Kofod Mogensen, Morten Piibeleht, et al. 2019. *MixedModels.jl.* doi:10.5281/zenodo.2592615. https://github.com/dmbates/MixedModels.jl. (Cited on page 51).

Bezanson, Jeff, Alan Edelman, Stefan Karpinski, and Viral B. Shah. 2017. "Julia: A Fresh Approach to Numerical Computing." *SIAM Review* 59, no. 1 (January): 65–98. ISSN: 0036-1445, 1095-7200. doi:10.1137/141000671. (Cited on page 41).

Bitler, Marianne P, Jonah B Gelbach, and Hilary W Hoynes. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96, no. 4 (August): 988–1012. doi:10.1257/aer.96.4.988. (Cited on page 20).

———. 2017. "Can Variation in Subgroups' Average Treatment Effects Explain Treatment Effect Heterogeneity? Evidence from a Social Experiment." *The Review of Economics and Statistics* 99, no. 4 (October): 683–697. ISSN: 0034-6535, 1530-9142. doi:10.1162/REST_a_00662. (Cited on page 20).

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *The Journal of Human Resources* 8 (4): 436. ISSN: 0022166X. doi:10.2307/144855. (Cited on page 2).

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50 (2): 317–372. doi:10.3368/jhr.50.2.317. (Cited on page 8).

Card, David, and Alan B. Krueger. 1992. "School Quality and Black-White Relative Earnings: A Direct Assessment*." *The Quarterly Journal of Economics* 107 (1): 151–200. doi:10.2307/2118326. (Cited on page 3).

Cornwell, Christopher, and William N. Trumbull. 1994. "Estimating the Economic Model of Crime with Panel Data." *The Review of Economics and Statistics* 76 (2): 360–366. ISSN: 15309142. doi:10.2307/2109893. (Cited on page 51).

Correia, Sergio. 2015. *Singletons, cluster-robust standard errors and fixed effects: A bad mix.* Technical Note. Duke University. (Cited on page 60).

———. 2017. *Linear Models with High-Dimensional Fixed Effects: An Efficient and Feasible Estimator.* Technical report. Working Paper. (Cited on page 44).

Croissant, Yves. 2018. *mlogit: Multinomial Logit Models.* R package version 0.3-0. https://CRAN.R-project.org/package=mlogit. (Cited on page 41).

Croissant, Yves, and Giovanni Millo. 2008. "Panel Data Econometrics in R : The **plm** Package." *Journal of Statistical Software* 27 (2). ISSN: 1548-7660. doi:10.18637/jss.v027.i02. (Cited on page 41).

Dettori, Joseph, Daniel Norvell, Andrea Skelly, and Jens Chapman. 2011. "Heterogeneity of treatment effects: from "How to treat" to "Whom to treat"." *Evidence-Based Spine-Care Journal* 2, no. 2 (May): 7–10. ISSN: 1663-7976, 1869-4136. doi:10.1055/s-0030-1267099. (Cited on page 2).

Dougherty, Christopher. 2005. "Why Are the Returns to Schooling Higher for Women than for Men?" *Journal of Human Resources* XL (4): 969–988. ISSN: 1548-8004. doi:10.3368/jhr.XL.4.969. (Cited on page 30).

Eicker, Friedhelm. 1967. "Limit theorems for regressions with unequal and dependent errors," 59–82. Berkeley, California: University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512981. (Cited on page 8).

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, and J. Robert Warren. 2018a. *Integrated Public Use Microdata Series, Current Population Survey: Version 6.0.* Type: dataset. doi:10.18128/D030.V6.0. https://cps.ipums.org. (Cited on page 13).

———. 2018b. *Integrated Public Use Microdata Series, Current Population Survey: Version 6.0.* Type: dataset. doi:10.18128/D030.V6.0. (Cited on page 31).

Fong, David Chin-Lung, and Michael Saunders. 2011. "LSMR: An Iterative Algorithm for Sparse Least-Squares Problems." *SIAM Journal on Scientific Computing* 33, no. 5 (January): 2950–2971. ISSN: 1095-7197. doi:10.1137/10079687X. (Cited on page 44).

Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1, no. 4 (October): 387. ISSN: 00129682. doi:10.2307/1907330. (Cited on page 43).

Gibbons, Charles E., Juan Carlos Súarez Serrato, and Michael B. Urbancic. 2018. "Broken or Fixed Effects?" *Journal of Econometric Methods.* doi:10.3386/w20342. (Cited on pages 3, 6).

Huber, Peter J. 1967. "Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics," 221–233. Berkeley, California: University of California Press. https://projecteuclid.org/euclid.bsmsp/1200512988. (Cited on page 8).

K Mogensen, Patrick, and Asbjørn N Riseth. 2018. "Optim: A mathematical optimization package for Julia." *Journal of Open Source Software* 3, no. 24 (April 4): 615. ISSN: 2475-9066. doi:10.21105/joss.00615. (Cited on page 48).

Konis, Kjell. 2007. "Linear Programming Algorithms for Detecting Separated Data in Binary Logistic Regression Models." PhD diss., Worcester College, University of Oxford. (Cited on page 49).

Liang, Kung-Yee, and Scott L. Zeger. 1986. "Longitudinal data analysis using generalized linear models." *Biometrika* 73 (1): 13–22. doi:10.1093/biomet/73.1.13. (Cited on page 8).

List, John, Azeem Shaikh, and Yang Xu. 2016. *Multiple Hypothesis Testing in Experimental Economics.* Artefactual Field Experiments 00402. The Field Experiments Website. htt ps://ideas.repec.org/p/feb/artefa/00402.html. (Cited on page 20).

Lovell, Michael C. 2008. "A Simple Proof of the FWL Theorem." *The Journal of Economic Education* 39, no. 1 (January): 88–91. ISSN: 2152-4068. doi:10.3200/JECE.39.1.88-91. (Cited on page 43).

McKelvey, Richard D., and William Zavoina. 1975. "A statistical model for the analysis of ordinal level dependent variables." *The Journal of Mathematical Sociology* 4 (1): 103–120. doi:10.1080/0022250X.1975.9989847. (Cited on page 48).

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14, no. 3 (October): 693. ISSN: 00206598. doi:10.2307/2525981. (Cited on page 2).

O'Leary, Dianne P. 1990. "Robust Regression Computation Using Iteratively Reweighted Least Squares." *SIAM Journal on Matrix Analysis and Applications* 11, no. 3 (July): 466–480. ISSN: 0895-4798, 1095-7162, accessed September 10, 2018. doi:10.1137/0611032. (Cited on page 47).

Python Software Foundation. 2018. *Python Software.* (Cited on page 41).

R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/. (Cited on page 41).

Renfro, Charles G. 2009. *The Practice of Econometric Theory.* Vol. 44. Advanced Studies in Theoretical and Applied Econometrics. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN: 978-3-540-75571-5. doi:10.1007/978-3-540-75571-5. (Cited on pages 40 sq.).

Revels, Jarrett, Miles Lubin, and Theodore Papamarkou. 2016. "Forward-Mode Automatic Differentiation in Julia." *CoRR* abs/1607.07892. arXiv: 1607.07892. http://arxiv.org/abs/1607.07892. (Cited on page 48).

Rogers, William. 1993. "Regression standard errors in clustered samples." *Stata Technical Bulletin* 13 (17): 19–23. https://www.stata.com/products/stb/journals/stb13.pdf. (Cited on page 8).

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301–316. doi:10.3368/jhr.50.2.301. (Cited on page 6).

Stapleton, SM, TO Oseni, YJ Bababekov, Y Hung, and DC Chang. 2018. "Race/ethnicity and age distribution of breast cancer diagnosis in the united states." *JAMA Surgery.* doi:10.1001/jamasurg.2018.0035. (Cited on page 3).

StataCorp. 2017. *Stata Statistical Software: Release 15.* College Station, TX: StataCorp LLC. (Cited on page 41).

Stock, James H., and Mark W. Watson. 2008. "Heteroskedasticity-Robust Standard Errors for Fixed Effects Panel Data Regression." *Econometrica* 76, no. 1 (January): 155–174. doi:10.1111/j.0012-9682.2008.00821.x. (Cited on page 8).

Swamy, P. A. V. B., and S. S. Arora. 1972. "The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models." *Econometrica* 40, no. 2 (March): 261. ISSN: 00129682. doi:10.2307/1909405. (Cited on page 45).

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S.* Fourth. ISBN 0-387-95457-0. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4. (Cited on page 41).

White, Halbert L. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48 (4): 817. doi:10.2307/1912934. (Cited on page 8).

Zeileis, Achim. 2004. "Econometric Computing with HC and HAC Covariance Matrix Estimators." *Journal of Statistical Software* 11 (10): 1–17. doi:10.18637/jss.v011.i10. (Cited on page 41).

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/. (Cited on page 41).