1-1-2014

Inference for Approximating Regression Models

Emil Pitkin
University of Pennsylvania, pitkin@wharton.upenn.edu

Inference for Approximating Regression Models

Abstract
The assumptions underlying the Ordinary Least Squares (OLS) model are regularly and sometimes severely violated. In consequence, inferential procedures presumed valid for OLS are invalidated in practice. We describe a framework that is robust to model violations, and describe the modifications to the classical inferential procedures necessary to preserve inferential validity. As the covariates are assumed to be stochastically generated ("Random-X"), the sought after criterion for coverage becomes marginal rather than conditional. We focus on slopes, mean responses, and individual future observations. For slopes and mean responses, the targets of inference are redefined by means of least squares regression at the population level. The partial slopes that that regression defines, rather than the slopes of an assumed linear model, become the population quantities of interest, and they can be estimated unbiasedly. Under this framework, we estimate the Average Treatment Effect (ATE) in Randomized Controlled Studies (RCTs), and derive an estimator more efficient than one commonly used. We express the ATE as a slope coefficient in a population regression and immediately prove unbiasedness that way. For the mean response, the conditional value of the best least squares approximation to the response surface in the population - rather than the conditional value of y, is aimed to be captured. A calibration through pairs bootstrap can markedly improve such coverage. Moving to observations, we show that when attempting to cover future individual responses, a simple in-sample calibration technique that widens the empirical interval to contain $(1-\alpha)*100\%$ of the sample residuals is asymptotically valid, even in the face of gross model violations. OLS is startlingly robust to model departures when a future y needs to be covered, but nonlinearity, combined with a skewed **X**-distribution, can severely undermine coverage of the mean response. Our ATE estimator dominates the common estimator, and the stronger the R squared of the regression of a patient's response on covariates, treatment indicator, and interactions, the better our estimator's relative performance. By considering a regression model as a semi-parametric approximation to a stochastic mechanism, and not as its description, we rest assured that a coverage guarantee is a coverage guarantee.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Lawrence D. Brown

Keywords
ATE, Calibration, Model mis-specification, Regression, Semi-parametric

INFERENCE FOR APPROXIMATING REGRESSION MODELS

Emil Pitkin

A DISSERTATION

in

Statistics

For the Graduate Group in
Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

Lawrence D. Brown
Miers Busch Professor
of Statistics

Graduate Group Chairperson

Eric Bradlow
K.P. Chao Professor, Marketing,
Statistics and Education

Dissertation Committee

Lawrence D. Brown, Miers Busch Professor of Statistics
Richard A. Berk, Professor of Statistics and Criminology
Andreas Buja, Liem Sioe Liong/First Pacific Company Professor of Statistics
Edward I. George, Universal Furniture Professor of Statistics
Linda Zhao, Professor

INFERENCE FOR APPROXIMATING REGRESSION MODELS

Emil Pitkin

Acknowledgments

I thank first Mama and Papa – Drs. Pitkin – together with Dedushka Rafa, Babushka Rita, and Babushka Sonya. You were my first teachers. I learn from you always. Love and reverence for your suffuses my every step. I dedicate this dissertation to you. I thank my incomparable, good, and wise advisor Larry Brown, who is a towering flame and who cultivated all my sparks, however slight. I thank Richard Berk, Andreas Buja, Ed George, and Linda Zhao who have also generously and significantly shaped my thinking as a statistician and this work. I thank my early mentor Ken Stanley, whose unfailing trust and reservoir of wisdom has propelled me on. I thank my classmates Alex Goldstein, Adam Kapelner, Justin Rising, Jordan Rodu, Jose Zubizaretta, and also Justin Bleich. We few, we happy few! I thank the members of the Salon and of the Captain's Club. Words and equations and flint are all the same – rub them the right way and the sparks will fly.

I thank Mr. Waldman and Moreh Shem for their love of learning and of teaching, Ethan Schaff for being the best math teacher I ever had, Mr. Randall and Mr. Jarvis for their perfect advice, Mr. Astrue for his frequent support, Dr. Fichtner for his example, Dr. Damjanovic and Dr. Morris for their important role.

ABSTRACT

INFERENCE FOR APPROXIMATING REGRESSION MODELS

Emil Pitkin

Lawrence D. Brown

The assumptions underlying the Ordinary Least Squares (OLS) model are regularly and sometimes severely violated. In consequence, inferential procedures presumed valid for OLS are invalidated in practice. We describe a framework that is robust to model violations, and describe the modifications to the classical inferential procedures necessary to preserve inferential validity. As the covariates are assumed to be stochastically generated (Random-X), the sought after criterion for coverage becomes marginal rather than conditional. We focus on slopes, mean responses, and individual future observations. For slopes and mean responses, the targets of inference are redefined by means of least squares regression at the population level. The partial slopes that that regression defines, rather than the slopes of an assumed linear model, become the population quantities of interest, and they can be estimated unbiasedly. Under this framework, we estimate the Average Treatment Effect (ATE) in Randomized Controlled Studies (RCTs), and derive an estimator more efficient than one commonly used. We express the ATE as a slope coefficient in a population

regression and immediately prove unbiasedness that way. For the mean response, the conditional value of the best least squares approximation to the response surface in the population rather than the conditional value of y, is aimed to be captured. A calibration through pairs bootstrap can markedly improve such coverage. Moving to observations, we show that when attempting to cover future individual responses, a simple in-sample calibration technique that widens the empirical interval to contain $(1-\alpha)*100\%$ of the sample residuals is asymptotically valid, even in the face of gross model violations. OLS is startlingly robust to model departures when a future y needs to be covered, but nonlinearity, combined with a skewed $\boldsymbol{X}$-distribution, can severely undermine coverage of the mean response. Our ATE estimator dominates the common estimator, and the stronger the R^2 of the regression of a patient's response on covariates, treatment indicator, and interactions, the better our estimator's relative performance. By considering a regression model as a semi-parametric approximation to a stochastic mechanism, and not as its description, we rest assured that a coverage guarantee is a coverage guarantee.

Contents

List of Tables

List of Figures

1 Introduction

The organizing principle of this work, which will be repeated in each chapter, is two-fold: 1) that classical regression theory does not accommodate non-linearity or heteroskedasticity in the presence of random predictors, and 2) that a re-examination of the target of inference can and does give rise to valid, marginal inference. Halbert White wrote a series of three papers (White, 1980b), (White, 1980a), (White, 1982) in which he addressed and solved the question of inference for misspecified models. The sandwich estimator he introduced is asymptotically equivalent to the non-parametric "pairs bootstrap," which we will employ often in this work. Chapter 2, an adaptation of (Buja et al., 2013), examines this form of valid inference, which includes a comparison of the relative performance of classical, or "conventional" standard errors, and those implied by the sandwich or the bootstrap. The key insight is that regression slope estimates derived through OLS are asymptotically unbiased for regression coefficients derived through population least squares. It is the randomness of the joint distribution of the predictors and response that motivates the population least squares procedure.

Chapter 3 changes orientation but preserves the statistical framework. We turn to Randomized Controlled Trials (RCTs) and the attendant estimation of the Average Treatment Effect (ATE). We briefly trace the evolution of its estimators, from the

progenitor in Neyman's thesis (Splawa-Neyman et al., 1990) to contemporary ones that consider, as we do, semi-parametric settings (Zhang et al., 2008), (Rosenblum and van der Laan, 2010), and then we explicitly define an estimator that is asymptotically efficient relative to the difference in means estimator. Our estimator can be expressed as a slope coefficient in a regression model of the sort defined in chapter 2, and it is therefore asymptotically unbiased. This work sets a principled foundation to the study of efficient ATE estimators, and admits many natural extensions to more complex study designs.

The problem of predicting future observations in a regression setting without invoking normal-theory parametric intervals is not new. (Stine, 1985), for example, examines the coverage of bootstrap prediction intervals. In his scheme the operating assumption is that the model is correctly specified, and hence that the distribution of the errors is known. (Schmoyer, 1992), who conscientiously avoids resampling methods, creates an estimator derived from a convolution of the empirical distribution of the regression residuals. More resonant with our work, which assumes only a joint distribution $\boldsymbol{P}$ between the $\vec{\boldsymbol{X}}$ and y, and more recently, (Politis, 2013) states as a common sense principle that in the absence of a model (the "model - free" case), prediction intervals should be based on quantiles of the observed predictive distribution. We adapt this principle to generate prediction intervals based on quantiles of the empirical distribution of the residuals. In chapter 4 we show how a simple in-sample calibration technique, which places minimal assumptions on the data generating process, gives desired, asymptotically valid coverage. Chapter 4 continues with an exploration of valid coverage for the mean response. Again, our target of inference is based on the population least squares approximation to the conditional mean. We in simulations show examples where $\vec{\boldsymbol{X}}\boldsymbol{\beta}$ is covered with probability lower than 20% for nominally 95% confidence intervals, when intervals based on classical

theory are applied to misspecified models with random predictors. Again relying on the bootstrap, we illustrate a technique that improves asymptotic marginal coverage.

2

Random Predictors and Model Violations*

Excerpted and adapted from Buja, A., Berk, Richard A., Brown, Lawrence D., George, Edward I., Pitkin, E., Traskin, M. Zhao, L., Zhang, K.: A Conspiracy of Random X and Model Violation Against Classical Inference in Linear Regression.

2.1 Abstract

This chapter reviews the insights of Halbert White's asymptotically correct inference in the presence of "model misspecification." This form of inference, which is pervasive in econometrics, relies on the "sandwich estimator" of standard error. White permits models to be "misspecified" and predictors to be random. Careful reading of his theory shows that it is a synergistic effect — a "conspiracy" — of nonlinearity and randomness of the predictors that has the deepest consequences for statistical inference. A valid alternative to the sandwich estimator is given by the "pairs bootstrap." We continue with an asymptotic comparison of the sandwich estimator and the standard error estimator from classical linear models theory. The comparison shows that when standard errors from linear models theory deviate from their sandwich analogs, they are usually too liberal, but occasionally they can be too conservative as well. The

*Joint work with Dana Chandler

chapter concludes by answering why we might be interested in inference for models that are not correct.

2.2 Introduction

The classical Gaussian linear model reads as follows:

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_N, \sigma^2 \boldsymbol{I}_{N\times N}) \qquad (\boldsymbol{y}, \boldsymbol{\epsilon} \in \mathbb{R}^N,\ \boldsymbol{X} \in \mathbb{R}^{N\times(p+1)},\ \boldsymbol{\beta} \in \mathbb{R}^{p+1}). \tag{2.1}$$

Important for the present focus are two aspects of how the model is commonly interpreted: (1) the model is assumed correct, that is, the conditional response means are a linear function of the predictors and the errors are independent, homoskedastic and Gaussian; (2) the predictors are treated as known constants even when they arise as random observations just like the response. Starting with Halbert White's (White, 1980a), ((White, 1980b), (White, 1982)) seminal articles, econometricians have used multiple linear regression without making the many assumptions of classical linear models theory. While statisticians use **assumption-laden exact finite sample inference**, econometricians use **assumption-lean asymptotic inference** based on the so-called "sandwich estimator" of standard error. The approach in this chapter is to interpret linear regression in a semi-parametric fashion: the generally nonlinear response surface is decomposed into a linear and a "residualized" nonlinear component. The modeling assumptions can then be reduced to i.i.d. sampling from largely arbitrary joint $(\vec{\boldsymbol{X}}, Y)$ distributions that satisfy a few moment conditions. It is in this assumption-lean framework that the sandwich estimator produces asymptotically correct standard errors.

We also connect the assumption-lean econometric framework to the "pairs boot-

strap." As the name indicates, the pairs bootstrap consists of resampling pairs $(\vec{x}_i, y_i)$, which contrasts with the "residual bootstrap" which resamples residuals r_i. Asymptotic theory exists to justify both types of bootstrap under different assumptions (Freedman, 1981), (Mammen, 1993). It is intuitively clear that the pairs bootstrap can be asymptotically justified in the assumption-lean framework mentioned above. In what follows we will use the general term "**assumption-lean estimators of standard error**" to refer to either the sandwich estimators or the pairs bootstrap estimators of standard error.

The chapter concludes by comparing the standard error estimates from assumption-lean theory and from classical linear models theory. The ratio of asymptotic variances — "***RAV***" for short — describes the discrepancies between the two types of standard error estimates in the asymptotic limit. If $\boldsymbol{RAV} \neq 1$, then there exist deviations from the linear model in the form of nonlinearities and/or heteroskedasticities. If $\boldsymbol{RAV} = 1$, then either the model is correct, or there has been a false negative.

2.3 Discrepancies between Standard Errors Illustrated

The table below shows regression results for a dataset in a sample of 505 census tracts in Los Angeles that has been used to examine homelessness in relation to covariates for demographics and building usage (Berk et al., 2008). We show the raw results of linear regression to illustrate the degree to which discrepancies can arise among three types of standard errors: SE_{lin} from linear models theory, SE_{boot} from the pairs bootstrap ($N_{boot} = 100,000$) and SE_{sand} from the sandwich estimator (according to (MacKinnon and White, 1985)). Ratios of standard errors are shown in bold font when they indicate a discrepancy exceeding 10%.

Table 1: Regression coefficients along with their standard errors estimated by different means.

	$\hat{\beta}_j$	SE_{lin}	SE_{boot}	SE_{sand}	$\frac{SE_{boot}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{boot}}$	t_{lin}	t_{boot}	t_{sand}
Intercept	0.760	22.767	16.505	16.209	**0.726**	**0.712**	0.981	0.033	0.046	0.047
MedianInc ($K)	-0.183	0.187	0.114	0.108	**0.610**	**0.576**	0.944	-0.977	-1.601	-1.696
PercVacant	4.629	0.901	1.385	1.363	**1.531**	**1.513**	0.988	5.140	3.341	3.396
PercMinority	0.123	0.176	0.165	0.164	0.937	0.932	0.995	0.701	0.748	0.752
PercResidential	-0.050	0.171	0.112	0.111	**0.653**	**0.646**	0.988	-0.292	-0.446	-0.453
PercCommercial	0.737	0.273	0.390	0.397	**1.438**	**1.454**	1.011	2.700	1.892	1.857
PercIndustrial	0.905	0.321	0.577	0.592	**1.801**	**1.843**	1.023	2.818	1.570	1.529

The ratios SE_{sand}/SE_{boot} show that the standard errors from the pairs bootstrap and the sandwich estimator are in rather good agreement. Not so for the standard errors based on linear models theory: we have $SE_{boot}, SE_{sand} > SE_{lin}$ for the predictors `PercVacant`, `PercCommercial` and `PercIndustrial`, and $SE_{boot}, SE_{sand} < SE_{lin}$ for `Intercept`, `MedianInc ($1000)`, `PercResidential`. Only for `PercMinority` is SE_{lin} off by less than 10% from SE_{boot} and SE_{sand}. The discrepancies affect outcomes of some of the t-tests: under linear models theory the predictors `PercCommercial` and `PercIndustrial` have commanding t-values of 2.700 and 2.818, respectively, which are reduced to unconvincing values below 1.9 and 1.6, respectively, if the pairs bootstrap or the sandwich estimator are used. On the other hand, for `MedianInc ($K)` the t-value -0.977 from linear models theory becomes borderline significant with the bootstrap or sandwich estimator if the plausible one-sided alternative with negative sign is used.

The second illustration of discrepancies between types of standard errors, shown in the table below, is with the Boston Housing data (Harrison Jr and Rubinfeld, 1978). We focus only on the comparison of standard errors. Here, too, SE_{boot} and SE_{sand} are mostly in agreement as they fall within less than 2% of each other, an exception being `CRIM` with a deviation of about 10%. By contrast, SE_{boot} and SE_{sand} are larger than their linear models cousin SE_{lin} by a factor of about 2 for `RM` and `LSTAT`, and about 1.5 for the intercept and the dummy variable `CHAS`. On the opposite side, SE_{boot} and

SE_{sand} are only a fraction of about 0.73 of SE_{lin} for `TAX`. Also worth stating is that for several predictors there is no substantial discrepancy among all three standard errors, namely `ZN`, `NOX`, `B`, and even for `CRIM`, SE_{lin} falls between the somewhat discrepant values of SE_{boot} and SE_{sand}.

Table 2: Regression coefficients along with their standard errors estimated by different means.

	$\hat{\beta}_j$	SE_{lin}	SE_{boot}	SE_{sand}	$\frac{SE_{boot}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{lin}}$	$\frac{SE_{sand}}{SE_{boot}}$	t_{lin}	t_{boot}	t_{sand}
(Intercept)	36.459	5.103	8.038	8.145	**1.575**	**1.596**	1.013	7.144	4.536	4.477
CRIM	-0.108	0.033	0.035	0.031	1.055	0.945	0.896	-3.287	-3.115	-3.478
ZN	0.046	0.014	0.014	0.014	1.005	1.011	1.006	3.382	3.364	3.345
INDUS	0.021	0.061	0.051	0.051	**0.832**	**0.823**	0.990	0.334	0.402	0.406
CHAS	2.687	0.862	1.307	1.310	**1.517**	**1.521**	1.003	3.118	2.056	2.051
NOX	-17.767	3.820	3.834	3.827	1.004	1.002	0.998	-4.651	-4.634	-4.643
RM	3.810	0.418	0.848	0.861	**2.030**	**2.060**	1.015	9.116	4.490	4.426
AGE	0.001	0.013	0.016	0.017	**1.238**	**1.263**	1.020	0.052	0.042	0.042
DIS	-1.476	0.199	0.214	0.217	1.075	1.086	1.010	-7.398	-6.882	-6.812
RAD	0.306	0.066	0.063	0.062	0.949	0.940	0.990	4.613	4.858	4.908
TAX	-0.012	0.004	0.003	0.003	**0.736**	**0.723**	0.981	-3.280	-4.454	-4.540
PTRATIO	-0.953	0.131	0.118	0.118	0.899	0.904	1.005	-7.283	-8.104	-8.060
B	0.009	0.003	0.003	0.003	1.026	1.009	0.984	3.467	3.379	3.435
LSTAT	-0.525	0.051	0.100	0.101	**1.980**	**1.999**	1.010	-10.347	-5.227	-5.176

Important messages are the following: (1) SE_{boot} and SE_{sand} are in substantial agreement; (2) SE_{lin} on the one hand and $\{SE_{boot}, SE_{sand}\}$ on the other hand can show substantial discrepancies; (3) these discrepancies are specific to predictors. In what follows we describe how the discrepancies arise from nonlinearities in the conditional mean and/or heteroskedasticities in the conditional variance of the response given the predictors. Furthermore, it will turn out that SE_{boot} and SE_{sand} are asymptotically correct while SE_{lin} is not.

2.4 Populations and Targets of Estimation

Before we compare standard errors it is necessary to define targets of estimation in a semi-parametric framework. Targets of estimation will no longer be parameters in a generative model but statistical functionals that are well-defined for a large

nonparametric class of data distributions. A seminal work that inaugurated this approach is P.J. Huber's 1967 article whose title is worth citing in full: "The behavior of maximum likelihood estimation under nonstandard conditions." The "nonstandard conditions" are essentially arbitrary distributions for which certain moments exist.

A population view of regression with random predictors has as its ingredients random variables X_1, ..., X_p and Y, where Y is singled out as the response. At this point the only assumption is that these variables have a joint distribution

$$\boldsymbol{P} \;=\; \boldsymbol{P}(\mathrm{d}y, \mathrm{d}x_1, ..., \mathrm{d}x_p)$$

whose second moments exist and whose predictors have a full rank covariance matrix. We write

$$\vec{\boldsymbol{X}} \;=\; (1, X_1, ..., X_p)^T.$$

for the *column* random vector consisting of the predictor variables with a constant 1 prepended to accommodate an intercept term. Values of the random vector $\vec{\boldsymbol{X}}$ will be denoted by lower case $\vec{x} = (1, x_1, ..., x_p)^T$. We write any function $f(X_1, ..., X_p)$ of the predictors equivalently as $f(\vec{\boldsymbol{X}})$ because the prepended constant 1 is irrelevant. Correspondingly we also use the notations

$$\boldsymbol{P} = \boldsymbol{P}(\mathrm{d}y, \mathrm{d}\vec{x}), \quad \boldsymbol{P}(\mathrm{d}\vec{x}), \quad \boldsymbol{P}(\mathrm{d}y \,|\, \vec{x}) \quad \text{or} \quad \boldsymbol{P} = \boldsymbol{P}_{Y,\vec{\boldsymbol{X}}}, \quad \boldsymbol{P}_{\vec{\boldsymbol{X}}}, \quad \boldsymbol{P}_{Y|\vec{\boldsymbol{X}}} \tag{2.2}$$

for the joint distribution of $(Y, \vec{\boldsymbol{X}})$, the marginal distribution of $\vec{\boldsymbol{X}}$, and the conditional distribution of Y given $\vec{\boldsymbol{X}}$, respectively. Nonsingularity of the predictor covariance matrix is equivalent to nonsingularity of the cross-moment matrix $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$.

Among functions of the predictors, a special one is the best $L_2(\boldsymbol{P})$ approximation

to the response Y, which is the conditional expectation of Y given $\vec{\boldsymbol{X}}$:

$$\mu(\vec{\boldsymbol{X}}) \;:=\; \text{argmin}_{f(\vec{\boldsymbol{X}})\in L_2(\boldsymbol{P})}\boldsymbol{E}[(Y-f(\vec{\boldsymbol{X}}))^2] \;=\; \boldsymbol{E}[\,Y\,|\,\vec{\boldsymbol{X}}\,]\,. \tag{2.3}$$

This is sometimes called the "conditional mean function" or the "response surface". Importantly we do not assume that $\mu(\vec{\boldsymbol{X}})$ is a linear function of $\vec{\boldsymbol{X}}$.

Among *linear* functions $l(\vec{\boldsymbol{X}}) = \boldsymbol{\beta}^T\vec{\boldsymbol{X}}$ of the predictors, one stands out as the best *linear* $L_2(\boldsymbol{P})$ or population LS linear approximation to Y:

$$\boldsymbol{\beta}(\boldsymbol{P}) \;:=\; \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\boldsymbol{E}[(Y-\boldsymbol{\beta}^T\vec{\boldsymbol{X}})^2] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}\boldsymbol{E}[\,\vec{\boldsymbol{X}}Y\,]\,. \tag{2.4}$$

The right hand expression follows from the normal equations $\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{\beta}-\boldsymbol{E}[\vec{\boldsymbol{X}}Y] = \boldsymbol{0}$ that are the stationarity conditions for minimizing the population LS criterion $\boldsymbol{E}[(Y-\boldsymbol{\beta}^T\vec{\boldsymbol{X}})^2] = -2\boldsymbol{\beta}^T\boldsymbol{E}[\vec{\boldsymbol{X}}Y] + \boldsymbol{\beta}^T\boldsymbol{E}[\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\boldsymbol{\beta} + \text{const}$.

By abuse of terminology, we use the expressions "population coefficients" for $\boldsymbol{\beta}(\boldsymbol{P})$ and "population approximation" for $\boldsymbol{\beta}(\boldsymbol{P})^T\vec{\boldsymbol{X}}$We will often write $\boldsymbol{\beta}$, omitting the argument $\boldsymbol{P}$ when it is clear from the context that $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$.

The population coefficients $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ form a *statistical functional* that is defined for a large class of data distributions $\boldsymbol{P}$. The question of how $\boldsymbol{\beta}(\boldsymbol{P})$ relates to coefficients in the classical linear model (2.1) will be answered in Section 2.6.

The population coefficients $\boldsymbol{\beta}(\boldsymbol{P})$ provide also the best linear $L_2(\boldsymbol{P})$ approximation to $\mu(\vec{\boldsymbol{X}})$:

$$\boldsymbol{\beta}(\boldsymbol{P}) \;=\; \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}\boldsymbol{E}[(\mu(\vec{\boldsymbol{X}})-\boldsymbol{\beta}^T\vec{\boldsymbol{X}})^2] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}\boldsymbol{E}[\,\vec{\boldsymbol{X}}\mu(\vec{\boldsymbol{X}})\,]\,. \tag{2.5}$$

This fact shows that $\boldsymbol{\beta}(\boldsymbol{P})$ depends on $\boldsymbol{P}$ only in a limited way, as will be spelled out below.

The response Y has the following natural decompositions:

$$\begin{aligned} Y &= \boldsymbol{\beta}^T\vec{\boldsymbol{X}} + \underbrace{(\mu(\vec{\boldsymbol{X}}) - \boldsymbol{\beta}^T\vec{\boldsymbol{X}})}_{\eta(\vec{\boldsymbol{X}})} + \underbrace{(Y - \mu(\vec{\boldsymbol{X}})}_{\epsilon} \\ &= \boldsymbol{\beta}^T\vec{\boldsymbol{X}} + \underbrace{\eta(\vec{\boldsymbol{X}}) + \epsilon}_{\delta} \\ &= \boldsymbol{\beta}^T\vec{\boldsymbol{X}} + \delta \end{aligned} \tag{2.6}$$

These equalities define the random variable $\eta = \eta(\vec{\boldsymbol{X}})$, called "nonlinearity", and ϵ, called "error" or "noise", as well $\delta = \epsilon + \eta$, for which there is no standard term so that "linearity deviation" may suffice. Unlike $\eta = \eta(\vec{\boldsymbol{X}})$, the error ϵ and the linearity deviation δ are not functions of $\vec{\boldsymbol{X}}$ alone; if there is a need to refer to the conditional distribution of either given $\vec{\boldsymbol{X}}$, we may write them as $\epsilon|\vec{\boldsymbol{X}}$ and $\delta|\vec{\boldsymbol{X}}$, respectively.

The error ϵ is not assumed homoskedastic, and indeed its conditional distributions $\boldsymbol{P}(d\epsilon|\vec{\boldsymbol{X}})$ can be quite arbitrary except for being centered and having second moments almost surely:

$$\boldsymbol{E}[\,\epsilon\,|\vec{\boldsymbol{X}}] \overset{\boldsymbol{P}}{=} 0, \qquad \sigma^2(\vec{\boldsymbol{X}}) := \boldsymbol{V}[\,\epsilon\,|\vec{\boldsymbol{X}}] = \boldsymbol{E}[\,\epsilon^2\,|\,\vec{\boldsymbol{X}}] \overset{\boldsymbol{P}}{<} \infty. \tag{2.7}$$

We will also need a quantity that describes the total conditional variation of the response around the LS linear function:

$$m^2(\vec{\boldsymbol{X}}) := \boldsymbol{E}[\,\delta^2\,|\,\vec{\boldsymbol{X}}] = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}}). \tag{2.8}$$

We refer to it as the "conditional mean squared error" of the population LS function.

Equations (2.6) above can be given the following *semi-parametric interpretation*:

$$\underbrace{\mu(\vec{\boldsymbol{X}})}_{\textit{semi-parametric part}} = \underbrace{\boldsymbol{\beta}^T\vec{\boldsymbol{X}}}_{\textit{parametric part}} + \underbrace{\eta(\vec{\boldsymbol{X}})}_{\textit{nonparametric part}} \tag{2.9}$$

The purpose of linear regression is to extract the parametric part of the response surface and provide statistical inference for the parameters even in the presence of a nonparametric part.

To make the decomposition (2.9) identifiable one needs an orthogonality constraint:

$$\boldsymbol{E}[\,(\boldsymbol{\beta}^T\vec{\boldsymbol{X}})\,\eta(\vec{\boldsymbol{X}})\,]\;=\;0.$$

For $\eta(\vec{\boldsymbol{X}})$ as defined above, this equality follows from the more general fact that the nonlinearity $\eta(\vec{\boldsymbol{X}})$ is uncorrelated with all predictors. Because we will need similar facts for ϵ and δ as well, we state them all at once:

$$\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\eta\,]\;=\;\mathbf{0},\qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\epsilon\,]\;=\;\mathbf{0},\qquad \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\delta\,]\;=\;\mathbf{0}. \tag{2.10}$$

Proofs: The nonlinearity η is uncorrelated with the predictors because it is the population residual of the regression of $\mu(\vec{\boldsymbol{X}})$ on $\vec{\boldsymbol{X}}$ according to (2.5). The error ϵ is uncorrelated with $\vec{\boldsymbol{X}}$ because $\boldsymbol{E}[\vec{\boldsymbol{X}}\epsilon] = \boldsymbol{E}[\vec{\boldsymbol{X}}\boldsymbol{E}[\epsilon|\vec{\boldsymbol{X}}]] = \mathbf{0}$. Finally, δ is uncorrelated with $\vec{\boldsymbol{X}}$ because $\delta = \eta + \epsilon$.

While the nonlinearity $\eta = \eta(\vec{\boldsymbol{X}})$ is uncorrelated with the predictors, it is not independent from them as it still is a function of them. By comparison, the error ϵ as defined above is *not* independent of the predictors either, but it enjoys a stronger orthogonality property than η: $\boldsymbol{E}[\,g(\vec{\boldsymbol{X}})\,\epsilon\,] = 0$ for all $g(\vec{\boldsymbol{X}}) \in L_2(\boldsymbol{P})$.

It is important to note that $\boldsymbol{\beta}(\boldsymbol{P})$ does *not* depend on the predictor distribution if and only if $\mu(\vec{\boldsymbol{X}})$ is linear. More precisely, for a fixed measurable function $\mu_0(\vec{x})$ consider the class of data distributions $\boldsymbol{P}$ for which $\mu_0(.)$ is a version of their conditional mean function: $\boldsymbol{E}[Y|\vec{\boldsymbol{X}}] = \mu(\vec{\boldsymbol{X}}) \stackrel{\boldsymbol{P}}{=} \mu_o(\vec{\boldsymbol{X}})$. In this class we have:

$$\begin{aligned}
&\mu_0(.)\ \textit{is nonlinear} &&\Longrightarrow &&\exists \boldsymbol{P}_1, \boldsymbol{P}_2: \quad \boldsymbol{\beta}(\boldsymbol{P}_1) \neq \boldsymbol{\beta}(\boldsymbol{P}_2),\\
&\mu_0(.)\ \textit{is linear} &&\Longrightarrow &&\forall \boldsymbol{P}_1, \boldsymbol{P}_2: \quad \boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}(\boldsymbol{P}_2).
\end{aligned}$$

(For proof details, see Appendix 2.12.1.) Two population LS lines for two different predictor distributions may differ when the conditional response is nonlinear, while they will be identical when it is linear in the covariates.

In the presence of nonlinearity the LS functional $\boldsymbol{\beta}(\boldsymbol{P})$ depends on the predictor distribution, hence the predictors are not ancillary for $\boldsymbol{\beta}(\boldsymbol{P})$.

2.5 Observational Datasets and Estimation

The term "observational data" means in this context "cross-sectional data" consisting of i.i.d. cases $(Y_i, X_{i,1}, ..., X_{i,p})$ drawn from a joint multivariate distribution $\boldsymbol{P}(\mathrm{d}y, \mathrm{d}x_1, ..., \mathrm{d}x_p)$ $(i = 1, 2, ..., N)$. We collect the predictors of case i in a column $(p+1)$-vector $\vec{\boldsymbol{X}}_i = (1, X_{i,1}, ..., X_{i,p})^T$, prepended with 1 for an intercept. We stack the N samples to form random column N-vectors and a random predictor $N \times (p+1)$-matrix:

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ .. \\ .. \\ Y_N \end{bmatrix}, \quad \boldsymbol{X}_j = \begin{bmatrix} X_{1,j} \\ .. \\ .. \\ X_{N,j} \end{bmatrix}, \quad \boldsymbol{X} = [\boldsymbol{1}, \boldsymbol{X}_1, ..., \boldsymbol{X}_p] = \begin{bmatrix} \vec{\boldsymbol{X}}_1^T \\ ... \\ ... \\ \vec{\boldsymbol{X}}_N^T \end{bmatrix}.$$

Similarly we stack the values $\mu(\vec{\boldsymbol{X}}_i)$, $\eta(\vec{\boldsymbol{X}}_i)$, $\epsilon_i = Y_i - \mu(\vec{\boldsymbol{X}}_i)$, δ_i, and $\sigma(\vec{\boldsymbol{X}}_i)$ to form random column N-vectors:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu(\vec{\boldsymbol{X}}1) \\ .. \\ .. \\ \mu(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta(\vec{\boldsymbol{X}}_1) \\ .. \\ .. \\ \eta(\vec{\boldsymbol{X}}_N) \end{bmatrix}, \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ .. \\ .. \\ \epsilon_N \end{bmatrix}, \quad \boldsymbol{\delta} = \begin{bmatrix} \delta_1 \\ .. \\ .. \\ \delta_N \end{bmatrix}, \quad \boldsymbol{\sigma} = \begin{bmatrix} \sigma(\vec{\boldsymbol{X}}_1) \\ .. \\ .. \\ \sigma(\vec{\boldsymbol{X}}_N) \end{bmatrix}. \quad (2.11)$$

The definitions of $\eta(\vec{\boldsymbol{X}})$, ϵ and δ in (2.6) translate to vectorized forms:

$$\boldsymbol{\eta} \;=\; \boldsymbol{\mu} - \boldsymbol{X}\boldsymbol{\beta}, \qquad \boldsymbol{\epsilon} \;=\; \boldsymbol{Y} - \boldsymbol{\mu}, \qquad \boldsymbol{\delta} \;=\; \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}. \tag{2.12}$$

It is important to keep in mind the distinction between population and sample properties. In particular, the N-vectors $\boldsymbol{\delta}$, $\boldsymbol{\epsilon}$ and $\boldsymbol{\eta}$ are *not* orthogonal to the predictor columns $\boldsymbol{X}_j$ in the sample. Writing $\langle \cdot, \cdot \rangle$ for the usual Euclidean inner product on $\mathbb{R}^N$, we have in general $\langle \boldsymbol{\delta}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\epsilon}, \boldsymbol{X}_j \rangle \neq 0$, $\langle \boldsymbol{\eta}, \boldsymbol{X}_j \rangle \neq 0$, even though the associated random variables are orthogonal to X_j in the population: $\boldsymbol{E}[\,\delta X_j] = \boldsymbol{E}[\,\epsilon X_j]$ $= \boldsymbol{E}[\,\eta(\vec{\boldsymbol{X}}) X_j] = \; 0$.

The **sample linear LS estimate** of $\boldsymbol{\beta}$ is the random column $(p+1)$-vector

$$\hat{\boldsymbol{\beta}} \;=\; (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)^T \;=\; \mathrm{argmin}_{\tilde{\boldsymbol{\beta}}} \, \|\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}\|^2 \;=\; (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\,. \tag{2.13}$$

Randomness stems from both the random response $\boldsymbol{Y}$ and the random predictors in $\boldsymbol{X}$. Associated with $\hat{\boldsymbol{\beta}}$ are the following:

the hat or projection matrix: $\quad \boldsymbol{H} \;=\; \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T,$

the vector of LS fits: $\quad \hat{\boldsymbol{Y}} \;=\; \boldsymbol{X}\hat{\boldsymbol{\beta}} \;=\; \boldsymbol{H}\boldsymbol{Y},$

the vector of residuals: $\quad \boldsymbol{r} \;=\; \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$

The vector $\boldsymbol{r}$ of residuals is of course distinct from the vector $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}$ as the latter arises from $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$.

2.6 Decomposition of the LS Estimate According to Two Sources of Variation

When the predictors are random and linear regression is interpreted semi-parametrically as the extraction of the linear part of a nonlinear response surface, the sampling variation of the LS estimate $\hat{\boldsymbol{\beta}}$ can be additively decomposed into two components: one component due to error $\boldsymbol{\epsilon}$ and another component due to nonlinearity interacting with randomness of the predictors. This decomposition is a direct reflection of the decomposition $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$, according to (2.6) and (2.12). We give elementary asymptotic normality statements for each part of the decomposition. The relevance of the decomposition is that it explains what the pairs bootstrap estimates, while the associated asymptotic normalities are necessary to justify the pairs bootstrap.

In the classical linear models theory, which is conditional on $\boldsymbol{X}$, the target of estimation is $\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$. When $\boldsymbol{X}$ is treated as random and nonlinearity is permitted, the target of estimation is the population LS solution $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{P})$ defined in (2.4). In this case, $\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]$ is a random vector that sits between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \;=\; (\hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \;+\; (\boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta}) \tag{2.14}$$

This decomposition corresponds to the decomposition $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$ as the following lemma shows.

Definition and Lemma: *The following quantities will be called "Estimation Offsets" or "EO" for short, and they will be prefixed as follows:*

$$\begin{aligned}
&\textit{Total EO}: && \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} &&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\delta},\\
&\textit{Error EO}: && \hat{\boldsymbol{\beta}} - \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] &&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon},\\
&\textit{Nonlinearity EO}: && \boldsymbol{E}[\hat{\boldsymbol{\beta}}|\boldsymbol{X}] - \boldsymbol{\beta} &&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}.
\end{aligned} \tag{2.15}$$

This follows immediately from the decompositions (2.12), $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{\mu}$, $\boldsymbol{\eta} = \boldsymbol{\mu} - \boldsymbol{X\beta}$, $\boldsymbol{\delta} = \boldsymbol{Y} - \boldsymbol{X\beta}$, and these facts:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}, \quad \boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}] = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\mu}, \quad \boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{X\beta}).$$

The first equality is the definition of $\hat{\boldsymbol{\beta}}$, the second uses $\boldsymbol{E}[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{\mu}$, and the third is a tautology.

The variance/covariance matrix of $\hat{\boldsymbol{\beta}}$ has a canonical decomposition with regard to conditioning on $\boldsymbol{X}$:

$$\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \boldsymbol{E}[\,\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]] \;+\; \boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]]. \tag{2.16}$$

This decomposition reflects the estimation decomposition (2.14) and $\boldsymbol{\delta} = \boldsymbol{\epsilon} + \boldsymbol{\eta}$ in view of (2.15):

$$\begin{aligned}
\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,] &= \boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\delta}\,], && (2.17)\\
\boldsymbol{E}[\,\boldsymbol{V}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]] &= \boldsymbol{E}[\,\boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\epsilon}\,|\boldsymbol{X}]\,], && (2.18)\\
\boldsymbol{V}[\,\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}\,|\boldsymbol{X}]] &= \boldsymbol{V}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\,]. && (2.19)
\end{aligned}$$

(Note that in general $\boldsymbol{E}[\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta}\,] \neq \boldsymbol{0}$ even though $\boldsymbol{E}[\,\boldsymbol{X}^T\boldsymbol{\eta}\,] = \boldsymbol{0}$ and hence $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\eta} \longrightarrow \boldsymbol{0}$ a.s.)

2.7 Assumption-Lean Central Limit Theorems

The three EOs arise from the decomposition $\delta = \epsilon + \eta$ (2.6). The respective CLTs draw on the analogous conditional second moment decomposition $m^2(\vec{\boldsymbol{X}}) = \sigma^2(\vec{\boldsymbol{X}}) + \eta^2(\vec{\boldsymbol{X}})$ (2.8). The asymptotic variance/covariance matrices have the well-known sandwich

form:

Proposition: *The three EOs follow central limit theorems under usual multivariate CLT assumptions:*

$$N^{1/2}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\,\boldsymbol{E}[\,\delta^{2}\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\right) \tag{2.20}$$

$$N^{1/2}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}]) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\,\boldsymbol{E}[\,\epsilon^{2}\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\right) \tag{2.21}$$

$$N^{1/2}\,(\boldsymbol{E}[\,\hat{\boldsymbol{\beta}}|\boldsymbol{X}]-\boldsymbol{\beta}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\,\boldsymbol{E}[\,\eta^{2}\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\right) \tag{2.22}$$

Proof Outline: The three cases follow the same way; we consider the first. Using $\boldsymbol{E}[\,\delta\vec{\boldsymbol{X}}\,]=\mathbf{0}$ from (2.10) we have:

$$\begin{aligned}
N^{1/2}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &= \left(\tfrac{1}{N}\,\boldsymbol{X}^{T}\boldsymbol{X}\right)^{-1}\left(\tfrac{1}{N^{1/2}}\,\boldsymbol{X}^{T}\boldsymbol{\delta}\right)\\
&= \left(\tfrac{1}{N}\,\textstyle\sum \vec{\boldsymbol{X}}_{i}\vec{\boldsymbol{X}}_{i}^{T}\right)^{-1}\left(\tfrac{1}{N^{1/2}}\,\textstyle\sum \vec{\boldsymbol{X}}_{i}\,\delta_{i}\right)\\
&\xrightarrow{\mathcal{D}} \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\mathcal{N}\left(\mathbf{0},\boldsymbol{E}[\,\delta^{2}\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]\right)\\
&= \mathcal{N}\left(\mathbf{0},\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\,\boldsymbol{E}[\,\delta^{2}\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\right)\,,
\end{aligned} \tag{2.23}$$

The proposition can be specialized in a few ways to cases of partial or complete well-specification:

- **First order well-specification:** When there is no nonlinearity, $\eta(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$, then

  $$N^{1/2}\,(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\,\boldsymbol{E}[\,\epsilon^{2}\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^{T}]^{-1}\right)$$

  The sandwich form of the asymptotic variance/covariance matrix is solely due to heteroskedasticity.

- **First and second order well-specification:** When additionally homoskedasticity holds, $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} \sigma^2$, then

$$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(\mathbf{0},\ \sigma^2\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}^T]^{-1}\right)$$

  The familiar simplified form is asymptotically valid under first and second order well-specification but without the assumption of Gaussian errors.

- **Deterministic nonlinear response:** $\sigma^2(\vec{\boldsymbol{X}}) \overset{\boldsymbol{P}}{=} 0$, then

$$N^{1/2}\,(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \quad \xrightarrow{\mathcal{D}} \quad \mathcal{N}\left(\mathbf{0},\ \boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}^T]^{-1}\,\boldsymbol{E}[\,\eta^2\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}^T]\,\boldsymbol{E}[\,\vec{\boldsymbol{X}}\,\vec{\boldsymbol{X}}^T]^{-1}\right)$$

  The sandwich form of the asymptotic variance/covariance matrix is solely due to nonlinearity and random predictors.

2.8 The Sandwich Estimator and the M-of-N Pairs Bootstrap

Empirically one observes that standard error estimates obtained from the pairs bootstrap and from the sandwich estimator are generally close to each other. This is intuitively unsurprising as they both estimate the same asymptotic variances. A closer connection between them will be established below.

2.8.1 The Plug-In Sandwich Estimator of Asymptotic Variance

The simplest form of the sandwich estimator of asymptotic variance is the plug-in version of the asymptotic variance as it appears in the CLT of (2.20), replacing the hard-to-estimate quantity $m^2(\vec{\boldsymbol{X}})$ with the easy-to-estimate quantity $\delta^2 = (Y - \boldsymbol{\beta}\vec{\boldsymbol{X}})^2$

according to (2.20). For plug-in one estimates the population expectations $\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$ and $\boldsymbol{E}[\,(Y-\vec{\boldsymbol{X}}^T\boldsymbol{\beta})\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]$ with sample means and the population parameter $\boldsymbol{\beta}$ with the LS estimate $\hat{\boldsymbol{\beta}}$. For this we use the notation $\hat{\boldsymbol{E}}[...]$ to express sample means:

$$\begin{aligned}\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] &= \tfrac{1}{N}\textstyle\sum_{i=1\ldots N}\vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i^T &&= \tfrac{1}{N}\,(\boldsymbol{X}^T\boldsymbol{X})\\ \hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}\hat{\boldsymbol{\beta}})^2\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T] &= \tfrac{1}{N}\textstyle\sum_{i=1\ldots N}(Y_i-\vec{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}_i\vec{\boldsymbol{X}}_i^T &&= \tfrac{1}{N}\,(\boldsymbol{X}^T D_{\boldsymbol{r}}^2\,\boldsymbol{X}),\end{aligned}$$

where $D_{\boldsymbol{r}}^2$ is the diagonal matrix with squared residuals $r_i^2 = (Y_i - \vec{\boldsymbol{X}}_i\hat{\boldsymbol{\beta}})^2$ in the diagonal. With this notation the simplest and original form of the sandwich estimator of asymptotic variance can be written as follows (White, 1980b):

$$\hat{\boldsymbol{AV}}_{sand} \;:=\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\hat{\boldsymbol{E}}[\,(Y-\vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\,\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1} \tag{2.24}$$

The sandwich standard error estimate for the j'th regression coefficient is therefore defined as

$$\hat{SE}_{sand}(\hat{\beta}_j) \;:=\; \frac{1}{N^{1/2}}(\hat{\boldsymbol{AV}}_{sand})_{jj}^{1/2}. \tag{2.25}$$

2.8.2 The M-of-N Pairs Bootstrap Estimator of Asymptotic Variance

To connect the sandwich estimator (2.24) to its bootstrap counterpart we need the M-of-N bootstrap whereby the *resample size* M is allowed to differ from the sample size N. It is at this point important not to confuse

- M-of-N resampling *with* replacement, and
- M-out-of-N subsampling *without* replacement.

In resampling the resample size M can be any $M < \infty$, whereas for subsampling it is necessary that the subsample size M satisfy $M < N$. We are here concerned with

bootstrap resampling, and we will focus on the extreme case $M \gg N$, namely, the limit $M \to \infty$.

Because resampling is i.i.d. sampling from some distribution, there holds a CLT as the resample size grows, $M \to \infty$. It is immaterial that in this case the sampled distribution is the empirical distribution $\boldsymbol{P}_N$ of a given dataset $\{(\vec{\boldsymbol{X}}_i, Y_i)\}_{i=1\ldots N}$, which is frozen of size N as $M \to \infty$.

Proposition: *For any fixed dataset of size N, there holds a CLT for the M-of-N bootstrap as $M \to \infty$. Denoting by $\boldsymbol{\beta}^*_M$ the LS estimate obtained from a bootstrap resample of size M, we have*

$$M^{1/2}\,(\boldsymbol{\beta}^*_M - \hat{\boldsymbol{\beta}}) \;\xrightarrow{\mathcal{D}}\; \mathcal{N}\left(\mathbf{0},\; \hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\,\hat{\boldsymbol{E}}[\,(Y - \vec{\boldsymbol{X}}^T\hat{\boldsymbol{\beta}})^2\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]\;\hat{\boldsymbol{E}}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T]^{-1}\right) \;\; (M \to \infty). \tag{2.26}$$

This is a straight application of the CLT of the previous section to the empirical distribution rather than the actual distribution of the data, where the middle part (the "meat") of the asymptotic formula is based on the empirical counterpart $r_i^2 = (Y_i - \vec{\boldsymbol{X}}_i^T\hat{\boldsymbol{\beta}})^2$ of $\delta^2 = (Y - \vec{\boldsymbol{X}}^T\boldsymbol{\beta})^2$. A comparison of (2.24) and (2.26) results in the following:

Observation: *The sandwich estimator* (2.24) *is the asymptotic variance estimated by the limit of the M-of-N pairs bootstrap as $M \to \infty$ for a fixed sample of size N.*

2.9 Adjusted Predictors

The adjustment formulas of this section serve to express the slopes of multiple regressions as slopes in simple regressions using adjusted single predictors. The goal is to analyze the discrepancies between the proper and improper standard errors of regression estimates in subsequent sections.

2.9.1 Adjustment formulas for the population

To express the population LS regression coefficient $\beta_j = \beta_j(\boldsymbol{P})$ as a simple regression coefficient, let the adjusted predictor $X_{j\bullet}$ be defined as the "residual" of the population regression of X_j, used as the response, on all other predictors. In detail, collect all other predictors in the random p-vector $\vec{\boldsymbol{X}}_{-j} = (1, X_1, ..., X_{j-1}, X_{j+1}, ..., X_p)^T$, and let $\boldsymbol{\beta}_{j\bullet}$ be the coefficient vector from the regression of X_j onto $\vec{\boldsymbol{X}}_{-j}$:

$$\boldsymbol{\beta}_{j\bullet} \;=\; \text{argmin}_{\tilde{\boldsymbol{\beta}} \in \mathbb{R}^p} \boldsymbol{E}[\,(X_j - \tilde{\boldsymbol{\beta}}^T \vec{\boldsymbol{X}}_{-j})^2] \;=\; \boldsymbol{E}[\,\vec{\boldsymbol{X}}_{-j} \vec{\boldsymbol{X}}_{-j}^T]^{-1} \boldsymbol{E}[\,\vec{\boldsymbol{X}}_{-j} X_j]\,.$$

The adjusted predictor $X_{j\bullet}$ is the residual from this regression:

$$X_{j\bullet} \;=\; X_j - \boldsymbol{\beta}_{j\bullet}^T \vec{\boldsymbol{X}}_{-j}\,. \tag{2.27}$$

The representation of β_j as a simple regression coefficient is as follows:

$$\beta_j \;=\; \frac{\boldsymbol{E}[\,Y X_{j\bullet}]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]} \;=\; \frac{\boldsymbol{E}[\,\mu(\vec{\boldsymbol{X}}) X_{j\bullet}]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]}\,. \tag{2.28}$$

2.9.2 Adjustment formulas for samples

To express estimates of regression coefficients as simple regressions, collect all predictor columns other than $\boldsymbol{X}_j$ in a $N \times p$ random predictor matrix $\boldsymbol{X}_{-j} = (\boldsymbol{1}, ..., \boldsymbol{X}_{j-1}, \boldsymbol{X}_{j+1}, ...)$ and define

$$\hat{\boldsymbol{\beta}}_{j\bullet} \;=\; \text{argmin}_{\tilde{\boldsymbol{\beta}}} \,\|\boldsymbol{X}_j - \boldsymbol{X}_{-j}\tilde{\boldsymbol{\beta}}\|^2 \;=\; (\boldsymbol{X}_{-j}^T \boldsymbol{X}_{-j})^{-1} \boldsymbol{X}_{-j}^T \boldsymbol{X}_j\,.$$

Using the notation "$\hat{\bullet}$" to denote sample-based adjustment to distinguish it from population-based adjustment "$\bullet$", we write the sample-adjusted predictor as

$$\boldsymbol{X}_{j\hat{\bullet}} \;=\; \boldsymbol{X}_j - \boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{j\hat{\bullet}} \;=\; (\boldsymbol{I} - \boldsymbol{H}_{-j})\boldsymbol{X}_j\,. \tag{2.29}$$

where $\boldsymbol{H}_{-j} = \boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^T\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^T$ is the associated projection or hat matrix. The j'th slope estimate of the multiple linear regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1, ..., \boldsymbol{X}_p$ can then be expressed in the well-known manner as the slope estimate of the simple linear regression without intercept of $\boldsymbol{Y}$ on $\boldsymbol{X}_{j\hat{\bullet}}$:

$$\hat{\beta}_j \;=\; \frac{\langle \boldsymbol{Y}, \boldsymbol{X}_{j\hat{\bullet}}\rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}\,. \tag{2.30}$$

With the above notation we can make the following distinctions: $X_{i,j\bullet}$ refers to the i'th i.i.d. replication of the population-adjusted random variable $X_{j\bullet}$, whereas $X_{i,j\hat{\bullet}}$ refers to the i'th component of the sample-adjusted random column $\boldsymbol{X}_{j\hat{\bullet}}$. Note that the former, $X_{i,j\bullet}$, are i.i.d. for $i = 1, ..., N$, whereas the latter, $X_{i,j\hat{\bullet}}$, are not because sample adjustment introduces dependencies throughout the components of the random N-vector $\boldsymbol{X}_{j\hat{\bullet}}$. As $N \to \infty$ for fixed p, however, this dependency disappears asymptotically, and we have for the empirical distribution of the values $\{X_{i,j\hat{\bullet}}\}_{i=1...N}$ the obvious convergence in distribution:

$$\{X_{i,j\hat{\bullet}}\}_{i=1...N} \quad \xrightarrow{\mathcal{D}} \quad X_{j\bullet} \overset{\mathcal{D}}{=} X_{i,j\bullet} \qquad (N \to \infty).$$

2.9.3 Adjustment Formulas for Decompositions and Their CLTs

The vectorized formulas for estimation offsets (2.14) have the following component analogs:

$$
\begin{aligned}
&\textit{Total EO}: && \hat{\beta}_j - \beta_j &&= \frac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\delta} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}, \\
&\textit{Error EO}: && \hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] &&= \frac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\epsilon} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}, \\
&\textit{Nonlinearity EO}: && \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j &&= \frac{\langle \boldsymbol{X}_{j\hat{\bullet}}, \boldsymbol{\eta} \rangle}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}.
\end{aligned}
\tag{2.31}
$$

Asymptotic normality can also be expressed for each $\hat{\beta}_j$ separately using population adjustment:

Corollary:

$$
\begin{aligned}
N^{1/2}(\hat{\beta}_j - \beta_j) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) = \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,\delta^2 X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) \\
N^{1/2}(\hat{\beta}_j - \boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}]) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right) \\
N^{1/2}(\boldsymbol{E}[\,\hat{\beta}_j|\boldsymbol{X}] - \beta_j) &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})X_{j\bullet}{}^2]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]^2}\right)
\end{aligned}
\tag{2.32}
$$

2.10 Proper and Improper Asymptotic Variances Expressed with Adjusted Predictors

The following prepares the ground for an asymptotic comparison of linear models standard errors with correct assumption-lean standard errors. We know the former to be potentially incorrect, hence a natural question is this: by how much can linear models standard errors deviate from valid assumption-lean standard errors? We look for an answer in the asymptotic limit, which frees us from issues related to how the standard errors are estimated.

Here is generic notation that can be used to describe the proper asymptotic variance of $\hat{\beta}_j$ as well as its decomposition into components due to error and due to nonlinearity:

Definition:

$$\boldsymbol{AV}^{(j)}_{lean}(f^2(\vec{\boldsymbol{X}})) \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]^2} \tag{2.33}$$

The proper asymptotic variance of $\hat{\beta}_j$ and its decomposition is therefore according to (2.32)

$$\begin{aligned} \boldsymbol{AV}^{(j)}_{lean}(m^2(\vec{\boldsymbol{X}})) \;&=\; \boldsymbol{AV}^{(j)}_{lean}(\sigma^2(\vec{\boldsymbol{X}})) \;+\; \boldsymbol{AV}^{(j)}_{lean}(\eta^2(\vec{\boldsymbol{X}})) \\ \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]^2} \;&=\; \frac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]^2} \;+\; \frac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}}){X_{j\bullet}}^2]}{\boldsymbol{E}[\,{X_{j\bullet}}^2]^2} \end{aligned} \tag{2.34}$$

The next step is to derive an asymptotic form for the conventional standard error estimate in the assumption-lean framework. This asymptotic form will have the appearance of an asymptotic variance but it is valid only in the assumption-loaded framework of first and second order well-specification. This "improper" standard error depends on an estimate $\hat{\sigma}^2$ of the error variance, usually $\hat{\sigma}^2 = \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\|^2/(N-p-1)$. In an assumption-lean context, with both heteroskedastic error variance and nonlinearity, $\hat{\sigma}^2$ has the following limit:

$$\hat{\sigma}^2 \;\stackrel{N\to\infty}{\longrightarrow}\; \boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,] \;=\; \boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})\,] + \boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})\,]$$

Standard error estimates are therefore given by

$$\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \;=\; \hat{\sigma}^2\,(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \qquad \hat{SE}^2_{lin}[\,\hat{\beta}_j] \;=\; \frac{\hat{\sigma}^2}{\|\boldsymbol{X}_{j\hat{\bullet}}\|^2}. \tag{2.35}$$

Their scaled limits are (a.s.) under usual assumptions as follows:

$$N\,\hat{\boldsymbol{V}}_{lin}[\,\hat{\boldsymbol{\beta}}\,] \overset{N\to\infty}{\longrightarrow} \boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,]\;\boldsymbol{E}[\,\vec{\boldsymbol{X}}\vec{\boldsymbol{X}}^T\,]^{-1}, \qquad N\,\hat{SE}^2_{lin}[\,\hat{\beta}_j] \overset{N\to\infty}{\longrightarrow} \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})\,]}{\boldsymbol{E}[\,X_{j\bullet}^2\,]}. \tag{2.36}$$

These are the asymptotic expressions that describe the limiting behavior of linear models standard errors in an assumption-lean context. Even though they are not proper asymptotic variances except in an assumption-loaded context, they are intended and used as such. We introduce the following generic notation for improper asymptotic variance where $f^2(\vec{\boldsymbol{X}})$ is again a placeholder for any one among $m^2(\vec{\boldsymbol{X}})$, $\sigma^2(\vec{\boldsymbol{X}})$ and $\eta^2(\vec{\boldsymbol{X}})$:

Definition:
$$\boldsymbol{AV}^{(j)}_{lin}(f^2(\vec{\boldsymbol{X}})) \;:=\; \frac{\boldsymbol{E}[\,f^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \tag{2.37}$$

Here is the improper asymptotic variance of $\hat{\beta}_j$ and its decomposition into components due to error and nonlinearity:

$$\begin{aligned} \boldsymbol{AV}^{(j)}_{lin}(m^2(\vec{\boldsymbol{X}})) \;&=\; \boldsymbol{AV}^{(j)}_{lin}(\sigma^2(\vec{\boldsymbol{X}})) \;+\; \boldsymbol{AV}^{(j)}_{lin}(\eta^2(\vec{\boldsymbol{X}})) \\ \frac{\boldsymbol{E}[\,m^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \;&=\; \frac{\boldsymbol{E}[\,\sigma^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \;+\; \frac{\boldsymbol{E}[\,\eta^2(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[\,X_{j\bullet}{}^2]} \end{aligned} \tag{2.38}$$

We examine next the discrepancies between proper and improper asymptotic variances.

2.10.1 Comparison of Proper and Improper Asymptotic Variances

It will be shown that the conventional asymptotic variances can be too small or too large to unlimited degrees compared to the proper marginal asymptotic variances. A comparison of asymptotic variances can be done separately for $\sigma^2(\vec{\boldsymbol{X}})$, $\eta^2(\vec{\boldsymbol{X}})$ and

$m^2(\vec{\boldsymbol{X}})$. To this end we form the ratios $\boldsymbol{RAV}_j(...)$ as follows:

Definition and Lemma: *Ratios of Proper and Improper Asymptotic Variances*

$$\begin{aligned}
\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) &:= \frac{\boldsymbol{AV}^{(j)}_{lean}(m^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{lin}(m^2(\vec{\boldsymbol{X}}))} = \frac{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})X_j{\bullet}^2]}{\boldsymbol{E}[m^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_j{\bullet}^2]} \\
\boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}})) &:= \frac{\boldsymbol{AV}^{(j)}_{lean}(\sigma^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{lin}(\sigma^2(\vec{\boldsymbol{X}}))} = \frac{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})X_j{\bullet}^2]}{\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_j{\bullet}^2]} \qquad (2.39) \\
\boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}})) &:= \frac{\boldsymbol{AV}^{(j)}_{lean}(\eta^2(\vec{\boldsymbol{X}}))}{\boldsymbol{AV}^{(j)}_{lin}(\eta^2(\vec{\boldsymbol{X}}))} = \frac{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})X_j{\bullet}^2]}{\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}})]\,\boldsymbol{E}[X_j{\bullet}^2]}
\end{aligned}$$

The second equality on each line follows from (2.38) and (2.34). The ratios in (2.39) express by how much the improper conventional asymptotic variances need to multiplied to match the proper asymptotic variances. Among the three ratios the relevant one for the overall comparison of improper conventional and proper inference is $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$. For example, if $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) = 4$, say, then, for large sample sizes, the correct marginal standard error of $\hat{\beta}_j$ is about twice as large as the incorrect conventional standard error. In general $\boldsymbol{RAV}_j$ expresses the following:

- If $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) = 1$, the conventional standard error for $\hat{\beta}_j$ is asymptotically correct;
- if $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) > 1$, the conventional standard error for β_j is asymptotically too small/optimistic;
- if $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}})) < 1$, the conventional standard error for β_j is asymptotically too large/pessimistic.

The ratios $\boldsymbol{RAV}_j(\sigma^2(\vec{\boldsymbol{X}}))$ and $\boldsymbol{RAV}_j(\eta^2(\vec{\boldsymbol{X}}))$ express the degrees to which heteroskedasticity and/or nonlinearity contribute asymptotically to the defects of conventional standard errors.

2.10.2 Meaning and Range of the *RAV*

Observations:

(a) If $X_{j\bullet}$ has unbounded support on at least one side, that is, if $\boldsymbol{P}[{X_{j\bullet}}^2 > t] > 0 \; \forall t > 0$, then

$$\sup_f \boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) = \infty \,. \tag{2.40}$$

(b) If the closure of the support of the distribution of $X_{j\bullet}$ contains zero but there is no pointmass at zero, that is, if $\boldsymbol{P}[{X_{j\bullet}}^2 < t] > 0 \; \forall t > 0$ but $\boldsymbol{P}[{X_{j\bullet}}^2 = 0] = 0$, then

$$\inf_f \boldsymbol{RAV}_j(f^2(\vec{\boldsymbol{X}})) = 0 \,. \tag{2.41}$$

Even though the ***RAV*** is not a correlation, it is nevertheless a measure of association between $f^2(\vec{\boldsymbol{X}})$ and ${X_{j\bullet}}^2$:

- Heteroskedasticities $\sigma^2(\vec{\boldsymbol{X}})$ with large average variance $\boldsymbol{E}[\sigma^2(\vec{\boldsymbol{X}}) \,|\, {X_{j\bullet}}^2]$ in the tail of ${X_{j\bullet}}^2$ imply an upward contribution to the overall $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$; heteroskedasticities with large average variance concentrated near ${X_{j\bullet}}^2 = 0$ imply a downward contribution to the overall $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$.

- Nonlinearities $\eta^2(\vec{\boldsymbol{X}})$ with large average values $\boldsymbol{E}[\eta^2(\vec{\boldsymbol{X}}) \,|\, {X_{j\bullet}}^2]$ in the tail of ${X_{j\bullet}}^2$ imply an upward contribution to the overall $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$; nonlinearities with large average values concentrated near ${X_{j\bullet}}^2 = 0$ imply a downward contribution to the overall $\boldsymbol{RAV}_j(m^2(\vec{\boldsymbol{X}}))$.

2.11 Discussion

We compared statistical inference from classical linear models theory with inference from assumption-lean semiparametric theory. The former is a finite-sample theory

that relies on strong assumptions and treats the predictors as fixed even when they are random, whereas the latter uses asymptotic theory that relies on few assumptions and treats the predictors as random. At a practical level, inferences differ in the type of standard error estimates they use: linear models theory is based on the "ususal" standard error which is a scaled version of the error standard deviation, whereas econometric theory is based on the so-called "sandwich standard error" which derives from an assumption-lean asymptotic variance. We observe the following:

- As the semiparametric framework makes no demands on the correctness of the linearity and homoskedasticity assumptions of linear models theory, a new interpretation of the targets of estimation is needed: linear fits estimate the best linear approximation to a usually nonlinear response surface.

- The discrepancies between standard errors from assumption-rich linear models theory and assumption-lean econometric theory can be of arbitrary magnitude in the asymptotic limit, but real data examples indicate discrepancies by a factors of 2 to be common. This is obviously relevant because such factors can change a t-statistic from significant to insignificant and vice versa.

- The pairs bootstrap is seen to be an alternative the sandwich estimate of standard error. In fact, the latter is the asymptotic limit in the M-of-N bootstrap as $M \to \infty$.

2.12 Proofs

2.12.1 Proofs from Section 2.4

The linear case is trivial: if $\mu_0(\vec{\boldsymbol{X}})$ is linear, that is, $\mu_0(\vec{x}) = \boldsymbol{\beta}^T\vec{x}$ for some $\boldsymbol{\beta}$, then $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}$ irrespective of $\boldsymbol{P}(\mathrm{d}\vec{x})$ according to (2.5). The nonlinear case is proved as follows: For any set of points $\vec{x}_1, ...\vec{x}_{p+1} \in \mathbb{R}^{p+1}$ in general position and with 1 in the first coordinate, there exists a unique linear function $\boldsymbol{\beta}^T\vec{x}$ through the values of $\mu_0(\vec{x}_i)$. Define $\boldsymbol{P}(\mathrm{d}\vec{x})$ by putting mass $1/(p+1)$ on each point; define the conditional distribution $\boldsymbol{P}(\mathrm{d}y \,|\, \vec{x}_i)$ as a point mass at $y = \mu_o(\vec{x}_i)$; this defines $\boldsymbol{P}$ such that $\boldsymbol{\beta}(\boldsymbol{P}) = \boldsymbol{\beta}$. Now, if $\mu_0()$ is nonlinear, there exist two such sets of points with differing linear functions $\boldsymbol{\beta}_1^T\vec{x}$ and $\boldsymbol{\beta}_2^T\vec{x}$ to match the values of $\mu_0()$ on these two sets; by following the preceding construction we obtain $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ such that $\boldsymbol{\beta}(\boldsymbol{P}_1) = \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 = \boldsymbol{\beta}(\boldsymbol{P}_2)$.

2.12.2 Conditional Expectation of RSS

The conditional expectation of the RSS allowing for nonlinearity and heteroskedasticity:

$$
\begin{aligned}
\boldsymbol{E}[\|\boldsymbol{r}\|^2|\boldsymbol{X}] &= E[\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{Y}|\boldsymbol{X}] && (2.42)\\
&= \boldsymbol{E}[(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\eta}+\boldsymbol{\epsilon})'(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\eta}+\boldsymbol{\epsilon})|\boldsymbol{X}] && (2.43)\\
&= \boldsymbol{E}[(\boldsymbol{\eta}+\boldsymbol{\epsilon})^T(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta}+\boldsymbol{\epsilon})|\boldsymbol{X}] && (2.44)\\
&= \mathrm{tr}(\boldsymbol{E}[(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta}+\boldsymbol{\epsilon})(\boldsymbol{\eta}+\boldsymbol{\epsilon})^T|\boldsymbol{X}]) && (2.45)\\
&= \mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta}\boldsymbol{\eta}^T+\boldsymbol{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T|\boldsymbol{X}]) && (2.46)\\
&= \mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{\eta}\boldsymbol{\eta}^T+\boldsymbol{D}_{\boldsymbol{\sigma}^2}) && (2.47)\\
&= |(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\eta}|^2+\mathrm{tr}((\boldsymbol{I}-\boldsymbol{H})\boldsymbol{D}_{\boldsymbol{\sigma}^2}) && (2.48)
\end{aligned}
$$

2.12.3 Limit of Squared Adjusted Predictors

The asymptotic limit of $\|\boldsymbol{X}_{j\hat{\bullet}}\|^2$:

$$\begin{aligned}
\frac{1}{N}\|\boldsymbol{X}_{j\hat{\bullet}}\|^2 &= \frac{1}{N}\boldsymbol{X}_j^T(\boldsymbol{I}-\boldsymbol{H}_{-j})\boldsymbol{X}_j \\
&= \frac{1}{N}\left(\boldsymbol{X}_j^T\boldsymbol{X}_j - \boldsymbol{X}_j^T\boldsymbol{H}_{-j}\boldsymbol{X}_j\right) \\
&= \frac{1}{N}X_{i,j}^2 \;-\; \left(\frac{1}{N}\sum X_{i,j}\vec{\boldsymbol{X}}_{i,-j}^T\right)\left(\sum_i \vec{\boldsymbol{X}}_{i,-j}\vec{\boldsymbol{X}}_{i,-j}^T\right)^{-1}\left(\sum_i \vec{\boldsymbol{X}}_{i,-j}X_{i,j}\right) \\
&\xrightarrow{\boldsymbol{P}} \boldsymbol{E}[X_j^2] \;-\; \boldsymbol{E}[X_j\vec{\boldsymbol{X}}_{-j}]\boldsymbol{E}[\vec{\boldsymbol{X}}_{-j}\vec{\boldsymbol{X}}_{-j}^T]^{-1}\boldsymbol{E}[\vec{\boldsymbol{X}}_{-j}X_j] \\
&= \boldsymbol{E}[X_{j\bullet}{}^2]
\end{aligned}$$

2.12.4 Asymptotic Normality in Terms of Adjustment

We gave the asymptotic limit of the conditional bias in vectorized form after (2.20)-(2.22). Here we derive the equivalent element-wise limit using adjustment The variance of the conditional bias is the marginal inflator of SE.

$$N^{1/2}(\boldsymbol{E}[\hat{\beta}_j|\boldsymbol{X}]-\beta_j) = N^{1/2}\frac{\langle X_{j\bullet},\boldsymbol{\eta}\rangle}{\|X_{j\bullet}\|^2} = \frac{\frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{\eta} - \frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{H}_{-j}\boldsymbol{\eta}}{\frac{1}{N}\|X_{j\bullet}\|^2}$$

$$\begin{aligned}
\frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{H}_{-j}\boldsymbol{\eta} &= \frac{1}{N^{1/2}}\boldsymbol{X}_j^T\boldsymbol{X}_{-j}(\boldsymbol{X}_{-j}^T\boldsymbol{X}_{-j})^{-1}\boldsymbol{X}_{-j}^T\boldsymbol{\eta} \\
&= \left(\frac{1}{N}\sum_i X_{i,j}\vec{\boldsymbol{X}}_{i,-j}^T\right)\left(\frac{1}{N}\sum_i \vec{\boldsymbol{X}}_{i,-j}\vec{\boldsymbol{X}}_{i,-j}^T\right)^{-1} \\
&\quad \left(\frac{1}{N^{1/2}}\sum_i \vec{\boldsymbol{X}}_{i,-j}\eta(\vec{\boldsymbol{X}}_i)\right) \\
&\overset{\mathcal{D}}{\approx} \boldsymbol{E}[X_j\vec{\boldsymbol{X}}_{-j}]\boldsymbol{E}[\vec{\boldsymbol{X}}_{-j}\vec{\boldsymbol{X}}_{-j}^T]\left(\frac{1}{N^{1/2}}\sum_i \vec{\boldsymbol{X}}_{i,-j}\eta(\vec{\boldsymbol{X}}_i)\right) \\
&= \boldsymbol{\beta}_{j\cdot}^T\left(\frac{1}{N^{1/2}}\sum_i \vec{\boldsymbol{X}}_{i,-j}\eta(\vec{\boldsymbol{X}}_i)\right)
\end{aligned}$$

$$= \frac{1}{N^{1/2}} \sum_i (\boldsymbol{\beta}_{j\cdot}^T \vec{\boldsymbol{X}}_{i,-j}) \eta(\vec{\boldsymbol{X}}_i)$$

$$\begin{aligned}
\frac{1}{N^{1/2}} \left(\boldsymbol{X}_j^T \boldsymbol{\eta} - \boldsymbol{X}_j^T \boldsymbol{H}_{-j} \boldsymbol{\eta} \right) &\overset{\mathcal{D}}{\approx} \frac{1}{N^{1/2}} \sum_i \left(X_{i,j} - \boldsymbol{\beta}_{j\cdot}^T \vec{\boldsymbol{X}}_{i,-j} \right) \eta(\vec{\boldsymbol{X}}_i) \\
&\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \boldsymbol{V}[(X_j - \boldsymbol{\beta}_{j\cdot}^T \vec{\boldsymbol{X}}_{-j}) \eta(\vec{\boldsymbol{X}})]\right) \\
&= \mathcal{N}\left(0, \boldsymbol{V}[X_{j\bullet} \eta(\vec{\boldsymbol{X}})]\right)
\end{aligned}$$

$$N^{1/2}(\boldsymbol{E}[\hat{\beta}_j | \boldsymbol{X}] - \beta_j) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{\boldsymbol{V}[X_{j\bullet} \eta(\vec{\boldsymbol{X}})]}{\boldsymbol{E}[{X_{j\bullet}}^2]^2}\right)$$

3

Improved Precision in Estimating Average Treatment Effects

3.1 Abstract

The Average Treatment Effect (ATE) is a global measure of the effectiveness of an experimental treatment intervention. In the context of randomized trials, classical methods of its estimation either ignore relevant covariates or do not fully exploit them. Regression based adjustment has primarily considered covariates as fixed, or the model as correctly specified. We relax these assumptions and present a method for improving the precision of the ATE estimate: the treatment and control responses are estimated via a regression, and information is pooled between the groups to produce an asymptotically unbiased estimate. The respective statistical models are thought only to estimate some linear approximation to the population response surfaces. Marginally valid standard errors are derived, and the estimator's performance is compared to a classical estimator. Conditions under which the regression-based estimator is preferable are detailed, and demonstrations on real and simulated data are presented.

3.2 Introduction

In the study of randomized controlled trials (RCTs), the average treatment effect (ATE) is a measure of an experimental intervention's global effect on a study population. For a treatment population T and control population C, the ATE is defined as $\tau = \mathbb{E}[T] - \mathbb{E}[C]$ for some measured response that can be continuous or categorical. The parameter τ can be estimated in a multitude of ways, each estimator depending on the sampling framework and model specification. The interpretation of and scope of inference for the ATE parameter will depend on these choices.

Past work has followed two principal strands. The first, earliest investigations of randomized experiments centered around finite, fixed populations, all of whose members would be randomized into either treatment(s) (the number of treatments could exceed one) or control groups; the random assignment furnished the randomness, and inference extended only as far as to these subjects in the trial. The foundation was thereby laid by Neyman, and subsequently developed by Rubin, for the notion of "potential outcomes," whose unbiased estimation represented the first attempt to estimate some ATE (Splawa-Neyman et al., 1990)[1]. In this, earliest exploration of the ATE, the scope of inference was the collection of units examined in the study only. The Neyman framework has since evolved to accommodate a superpopulation from which the experimental units are sampled (Imbens and Rubin, 2007).

More recent literature has aimed to improve the precision of the ATE estimates via regression; whenever signal exists, the conditional variance of the response is reduced, with attendant gains in efficiency. The conventional philosophy behind regression adjustments in RCTs is appealing: not only does the ATE become a parameter of the model, but the random discrepancies in empirical covariate distributions between the

[1]Neyman considered a series of plots in a field, on each of which one of several varieties of fertilizer was applied; he wished to estimate the true average yield of the aggregated plots, even though the individual plots were fertilized with only one variety

treatment and control groups are adjusted away, and the essential difference between treatment and control groups is retained. Some authors (Freedman, 2008) assume the framework in which a true, generating model exists, which could be correctly and completely specified via a regression equation. The estimating regression model in practice, however, is often misspecified, and in this case covariance adjustment can lead to undesirable consequences: in an influential critique, Freedman demonstrates how regression-based ATE estimators can lead to reduced asymptotic precision, and how they can be beset by small-sample bias. Often, fixed-X design is often implicitly assumed, explicitly when inference is restricted to the sample at hand. Elsewhere, also in the name of improving precision of the ATE estimate, knowledge of the population mean of the covariate distribution is assumed (Lin, 2013). In this chapter we will step aside from the finite sample Neyman framework within which Freedman offers his analysis, and we will make fewer assumptions.

In our view, the posited statistical model rarely captures the data generating process, and subjects' covariates ought to be treated as random. We therefore argue for an analysis of RCTs that places minimal assumptions on the population from which data are generated, and assume only that there exists a joint distribution between the covariates, the treatment indicator, and the response[2]. There exist best linear approximations to the regression surfaces, derived through population least squares, and these linear approximations are the targets of inference for the treatment and control regressions. Considered this way, we derive efficient, asymptotically unbiased estimates for the unconditional average difference between these surfaces. Such an approach, with minimal assumptions placed on the data generating mechanism, echoes the work of (Yang and Tsiatis, 2001) and (Tsiatis et al., 2008). In this assumption-

[2]Fixed X is rarely reasonable in the context of RCTs: after patients have entered a clinical trial, nobody seriously presumes that other, putative patients in the target population have the same individual characteristics as the study subjects.

lean framework, we derive an efficient ATE estimator for a more powerful test of the ATE.

In section 2 we describe the assumptions that have underlain much of previous work. In section 3 we define our perspective, define our estimator of the ATE, and compare its performance to an alternate, simple estimator. Section 4 illustrates the comparison on a dataset and investigates the behavior of our estimator via simulation. Section 5 concludes.

3.3 Neyman Framework, Fixed X, True Models

Most pithily, the heart of Neyman's paradigm can be described as a "repeated-sampling randomization-based" method (Rubin, 1990). Of N subjects $\{Y_i\}_{1:N}$, fixed once and for all, n_T are assigned to the treatment group, and the remaining $n_C = N - n_T$ are exposed to the control condition. In subsequent hypothetical realizations of the experiment, another n_T subjects out of the original N are exposed to the treatment, and the remainder to the control. Each of the $\binom{N}{n_T}$ subsets has an equal probability of being the "treated block" in any given experiment. Note that in the thought experiment, the same, fixed n_T number of units are assigned treatment, rather than each of the n subjects being assigned treatment as a Bernoulli trial with probability n_T/n.

To each subject are associated two hypothetical states, one of which is observed in practice[3]. These are called "potential outcomes," and they refer to the (deterministic) response of the subject, had he been subjected to the treatment (or control) condition. Let $Y_i(0)$ be the ith patient's response under the control, and let $Y_i(1)$ be the corresponding response under treatment. The ith patient's unobserved treatment

[3]Of course, with multiple treatments, multiple states will be associated with each subject

effect is defined as $Y_i(1) - Y_i(0)$. The sample-ATE, known as SATE, is defined as

$$\tau^S = \frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)] \tag{3.1}$$

and is estimated (w.l.o.g. let $Y_1, \ldots Y_{n_T}$ be treated) by

$$\hat{\tau}^S = \frac{1}{n_T}\sum_{i=1}^{n_T} Y_i(1) - \frac{1}{n_C}\sum_{i=1}^{n_C} Y_i(0) \tag{3.2}$$

$\hat{\tau}^S$ is an unbiased estimate of τ^S.

In the literature a complementary parameter exists, called the population average treatment effect (PATE). Here the subjects under investigation are thought to have been sample from a superpopulation. The parameter, if the potential outcomes were known, would be computed similarly to the SATE, except the summation in (3.1) would be taken not over the sample in question but over all subjects in the population. In RCTs, where the desired scope of inference extends beyond the sample in question, the PATE is the more logical parameter to estimate. The estimate will be more variable: "sample selection error," defined by $\Delta_S = PATE - SATE$, adds to the uncertainty of the ATE estimate (Imbens, 2004), (Imai et al., 2008).

The attractiveness of this estimator described lies in its simplicity: at its core it is just a difference of means. In the name of simplicity, however, potentially useful subject specific characteristics are sacrificed. It can therefore be desirable to estimate the ATE by way of regression: the intention behind this approach being to make more precise the estimate of the ATE parameter by adjusting for the treated and control units' covariates. The conclusions are sensitive to the assumptions made about the statistical model.

Freedman (Freedman, 2008), responding to its pervasiveness as an estimation tool, specifically considers OLS. He calls the ATE parameter b_{ITT}, where ITT is the

acronym for "intention to treat.[4]" b_{ITT} can be estimated via regression in several ways. In the first, most simple and slightly contrived way, one regresses the response on the treatment indicator only, and takes note of the indicator's coefficient. This is akin to measuring the difference of treated and control means. For testing the equality of b_{ITT} to some value, usually 0, one employs the usual t-tests [5]

One may then proceed to introduce covariates into the regression; the new coefficient of the treatment indicator, $\hat{b}_{ITT}$, is now the estimator of b_{ITT}. Freedman demonstrates that while augmenting the design with covariates can improve the performance of the estimator, it can worsen it as well (standard error is either increased or decreased, depending on the data). What's worse, the nominal standard error of $\hat{b}_{ITT}$, in addition to the estimator itself, can be severely biased. The counterintuitive result arises because, as Freedman writes: "randomization does not justify the assumptions behind the OLS model." That is, the demands the Neyman paradigm places on the nature of the data are not nearly as stringent as those imposed by OLS, with its requirements of homoscedasticity, linearity, and fixed design.

A recent and interesting paper by (Lin, 2013) reacts to Freedman's critique, works in the Neyman paradigm, and reports the conditions under which regression adjustment can give asymptotically valid coverage. His most trenchant point is that, by including a full set of covariate-treatment indicator interactions in the regression model, thereby allowing heterogeneous effects, OLS adjustment cannot worsen asymptotic precision. In his formulation, the covariates, once observed, are fixed, and "random assignment is the sole source of randomness in this model." Another recent paper (Imbens and Wooldridge, 2008) analyzes ATEs under more flexible circumstances,

[4] "Intention to treat" is described as "the effect of assigning everybody to treatment, minus the effect of assigning them to control."

[5] Interestingly, the usual t-tests assume the units to have been randomly sampled, but conclusions are little affected when the assumption does not hold for a difference in means.(Freedman et al., 1998)

allowing covariates to have a distribution and assuming heterogeneous effects. The authors present their useful results "assuming the linear regression model is correctly specified." (Samii and Aronow, 2012) compare the variances of the Neyman based and sandwich based estimators of the variance of the ATE, although the jump between fixed and random covariates is not obvious.

We come to similar conclusions, but after relaxing assumptions of proper specification. We opt for a parallel framework, one which is not hidebound by the assumptions behind OLS. We permit the subjects' covariates to be drawn from a distribution, and though we analyze through OLS, we do not assume that linear relationships hold in the population. We, too, include a full set of covariate-treatment indicator interactions to model heterogeneous effects. The assumption-lean model is described fully in the following section.

3.4 Target of Estimation

In this formulation, nearly all quantities are random. Whereas in the earlier Neyman framework and that adopted by some authors, only the assignment of the n_T treated units is random – but not the subject pool (hence not the covariates), nor the potential responses – now all that will remain fixed is the number of units assigned to treatment, and the number to control.[6] Subjects are not assigned treatment with probability n_T/N. Mathematically, subjects are sampled independently from an infinite population; which subjects are chosen will vary from sample to sample, as will the observed covariates. The subjects of both the treatment and control groups are all assumed to have been sampled at random from the same population – that is, at the population level, the covariate distributions are the same for the two groups, and

[6]As before, the thought experiment requires, in the next realization of the experiment, for the same n_T number of subjects to be assigned treatment, and the remaining n_C – control.

assignment of treatment is independent of covariates. To better define the mathematical target of inference, we include a condensed variant of the exposition in the previous chapter.

Consider for now either the treated or the control population. Let the population of subjects be described by the random variables $X_1, \ldots, X_p, Y$. Their joint distribution $\mathbf{P} = \mathbf{P}\ (dx_1, \ldots, dx_p, dy)$ has a full rank covariance matrix and four moments. $\vec{\boldsymbol{X}} = (1, X_1, \ldots, X_p)'$ is the random vector of the predictor variables. Finally, let $\mu(\vec{\boldsymbol{X}})$ be the conditional mean of Y at $\vec{\boldsymbol{X}}$: $\mu(\vec{\boldsymbol{X}}) = \mathbb{E}\left[Y|\vec{\boldsymbol{X}}\right]$. We relax OLS assumptions, permitting, for example, predictor variables to be omitted, and do not require the true response surface to be linear in the predictors. Indeed, the operating assumption is that it is not. Instead, we work with a conditional mean that can be decomposed into linear and non-linear components.

The linear component is thought of as the *best linear approximation* to the true conditional response surface; its partial slopes are defined by $\boldsymbol{\beta} = \left(\mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right]\right)^{-1}\mathbb{E}\left[\boldsymbol{X}\mu(\boldsymbol{X})\right]$, where the expectation is over the joint distribution of the $\boldsymbol{X}$ and the Y. That $\boldsymbol{\beta}$ will ultimately become a target of inference.

The difference between $\mu(\vec{\boldsymbol{X}})$ and $\boldsymbol{\beta}^T\vec{\boldsymbol{X}}$ is denoted by $\eta(\vec{\boldsymbol{X}})$, which is itself a random variable. Our operating assumption is that $\eta(\vec{\boldsymbol{X}})$ will not be identically equal to zero – that is, that non-linearity will be present in the population. In this chapter, $\boldsymbol{\beta}$ is estimated in the usual least squares fashion: $\hat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1}X^TY$.

The additional results relevant to this chapter are the following:

(a) $N^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges to a random variable with mean 0; $\hat{\boldsymbol{\beta}}$ is an asymptotically unbiased estimator of $\boldsymbol{\beta}$.

(b) In finite samples, $\hat{\boldsymbol{\beta}}$ may be a biased estimator of $\boldsymbol{\beta}$.

With this background presented, we paint in more detail the particulars of how responses might be adjusted for covariates, and what the targets of inference are.

The formulation is more general than in (Yang and Tsiatis, 2001), for example, which considers a baseline measurement of Y (as well as a treatment indicator) as the sole covariates. The treatment and control responses, respectively, can be denoted in the population by

$$T_i = \beta_T^{(0)} + \vec{\boldsymbol{X}}'_{Ti}\boldsymbol{\beta}_T + \eta_T(\vec{\boldsymbol{X}})_i + \epsilon_{Ti} \tag{3.3}$$

and, analagously,

$$C_i = \beta_C^{(0)} + \vec{\boldsymbol{X}}'_{Ci}\boldsymbol{\beta}_C + \eta_C(\vec{\boldsymbol{X}})_i + \epsilon_{Ci} \tag{3.4}$$

The $\beta^{(0)}$ are the respective intercepts at the population level, and the $\boldsymbol{\beta}$ are the respective vectors of population partial slopes. $\vec{\boldsymbol{X}}'_T$ is a random vector of treated units' covariates. Again, because we no longer assume that the response is linear in the covariates, $\beta_T^{(0)} + \vec{\boldsymbol{X}}'_T\boldsymbol{\beta}_T$ should be thought of as the treated group's best linear approximation, at the population level, to $\mathbb{E}\left[T|\vec{\boldsymbol{X}}\right]$. Therefore $\beta_T^{(0)}$ and $\boldsymbol{\beta}_T$ are population parameters derived from population least squares regression and minimize the expected squared distance between the linear surface and the true response surface. The nonlinearity $\eta_T(\vec{\boldsymbol{X}})$ is a random variable that represents the difference between the true conditional mean of T and its best linear approximation in the population. In equations:

$$\eta_T(\vec{\boldsymbol{X}}) = \mathbb{E}\left[T|\vec{\boldsymbol{X}}\right] - (\beta_T^{(0)} + \vec{\boldsymbol{X}}'_T\boldsymbol{\beta}_T) \tag{3.5}$$

Similar facts hold for $\eta_C(\vec{\boldsymbol{X}})$. Certain other assumptions and comments are warranted here.

(a) *Errors.* We place minimal demands on the errors: they should have zero mean; because of iid sampling, they will be independent. Their distributional form is unspecified, and we do not assume normality of errors. Their variances, however,

we allow to differ: denote the treated and control error variances, respectively, by σ_T^2 and σ_C^2.

(b) *Heterogeneity* Note, also, that in the population slopes are not assumed to be the same; we allow for heterogeneous effects. The nonlinearity random variables, too, are allowed to differ between the treatment and control groups.

As detailed in (Buja et al., 2013), the target of estimation – the intercept and slopes – should be estimated, even in the random X setting, by the classical least squares estimators, and we shall do the same.

3.4.1 ATE definition through regression

We are going to re-express the ATE parameter through regression, thereby foreshadowing the proposed estimator. As mentioned in the introduction, and using the notation developed above, the ATE is the difference between the population average of the treated subjects and their control counterparts:

$$\tau = \mathbb{E}[T] - \mathbb{E}[C] \tag{3.6}$$

Subtracting (3.4) from (3.3) and taking expectations, we see that

$$\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right) + \mathbb{E}\left[\vec{\boldsymbol{X}}_T\right]\boldsymbol{\beta}_T - \mathbb{E}\left[\vec{\boldsymbol{X}}_C\right]\boldsymbol{\beta}_C \tag{3.7}$$

Note that the non-linear components $\eta_T(\vec{\boldsymbol{X}})$ and $\eta_C(\vec{\boldsymbol{X}})$ from (3.4) and (3.3) do not appear in the equation above. Simply, they are both equal to zero in expectation over the joint distribution of $\vec{\boldsymbol{X}}$ and Y.[7] It deserves mentioning that the $\boldsymbol{\beta}$ in preceding

[7]This is an interesting point, whose derivation is not central to the discussion, and is therefore deferred to the appendix

equations are derived from the best linear approximations to the response surface, and may differ appreciably therefrom.

The careful reader will remark that we did not simplify fully, as $\mathbb{E}\left[\vec{\boldsymbol{X}}_T\right] = \mathbb{E}\left[\vec{\boldsymbol{X}}_C\right] = \mathbb{E}\left[\vec{\boldsymbol{X}}\right]$, since, according to our assumptions, the treated and control subjects are drawn from the same population. And, indeed, (3.7) can be written as

$$\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right) + \mathbb{E}\left[\vec{\boldsymbol{X}}\right](\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \tag{3.8}$$

We consciously write these two true statement separately. In (3.7), one is tempted to estimate the respective expected values separately by the respective covariate means of treatment and control groups. In (3.8) a single estimate will do, perhaps through a mean of all observed covariates, both treated and control. There will be a difference, in practice, and we wished to emphasize it now.

One more remark: when $\mathbb{E}\left[\vec{\boldsymbol{X}}\right] = \mathbf{0}$, then $\tau = \left(\beta_T^{(0)} - \beta_C^{(0)}\right)$, and the ATE is just the difference between the respective population intercepts. This formulation hints at how we may wish to estimate the ATE from sample regressions.

All the while, we have represented the treatment and control regressions separately, if only to emphasize that the two functional relations of covariates to the responses need bear no relation to one another in order for an ATE to be properly defined, and, later, estimated. A single regression formulation, with interactions, may be more familiar. The response can be written as:

$$Y_i = \beta^{(0)} + \boldsymbol{\beta}^{(T)} I_T + \boldsymbol{\beta}' \vec{\boldsymbol{X}}_i + \boldsymbol{\beta}^{(Int)} I_T \vec{\boldsymbol{X}}_i + \eta(\vec{\boldsymbol{X}})_i + I_T g(\vec{\boldsymbol{X}})_i + \epsilon_i \tag{3.9}$$

where $g(\vec{\boldsymbol{X}})$ is the difference in the treatment and control non-linearity functions. Here I_T is the treatment indicator at the population level; $\boldsymbol{\beta}^{(Int)}$ is the vector in which are collected the differences in coefficients found in the treatment and control

regressions respectively. The linear approximation being the target of estimation, we will restrict our attention to estimating the $\boldsymbol{\beta}$. In equation (3.9) above, $\boldsymbol{\beta}^{(T)}$ is precisely the ATE parameter when the covariate expectation is equal to 0.

3.4.2 ATE estimation

In this section we define two ATE estimators that can be derived from a random-X regression. The first reduces to the most familiar difference in means estimator, while the second borrows information across the treated and control groups. For both estimators, we write the regression-derived expression that is equivalent to the ATE, and then appeal to plug-in MLE estimates for the associated estimator.

(a) Difference in means estimator.

Recall this fact of elementary statistics: that there is one point through which the least squares regression line must pass, and that that point the mean of the predictors and the mean of the response: $\hat{y}|_{x=\bar{x}} = \bar{y}$. So if we substitute $\bar{\vec{\boldsymbol{X}}}_T$ into the treatment regression, the estimated conditional response will be $\bar{T}$, an unbiased estimate of $\mathbb{E}[T]$. In the same way we can find an unbiased estimate of $\mathbb{E}[C]$, and, as a result, of the ATE. One must be very careful when estimating the standard error of this quantity $\left[\hat{\beta}_T^{(0)} + \bar{\vec{\boldsymbol{X}}}_T\hat{\boldsymbol{\beta}}_T\right] - \left[\hat{\beta}_T^{(0)} + \bar{\vec{\boldsymbol{X}}}_T\hat{\boldsymbol{\beta}}_T\right]$, as we do in section 4.3.

What we have done, in effect, by substituting the respective covariate means into the separate regressions, is estimate $\mathbb{E}\left[\vec{\boldsymbol{X}}\right]$ separately in the treatment and the control regression, which is congruent with the decomposition in (3.7). But the winding path leads back to response sample means – to compute them no regressions need to have been run, no covariates measured. The lesson here is that for our purposes, controlling for covariates loses its appeal and effectiveness

if no information is shared between the treatment and the control groups.

(b) A strictly regression derived estimator.

Alternatively, $\mathbb{E}\left[\vec{\boldsymbol{X}}\right]$ can – and in most cases should – be estimated not separately as above, twice, but rather once, by the complete set of the pooled covariates. It should be estimated at the mean of *all* covariates, $\left(n_T\bar{\vec{\boldsymbol{X}}}_T + n_C\bar{\vec{\boldsymbol{X}}}_C\right)/N$. The efficiency gains will be seen in section 4.2. This approach is more congruent with (3.8), so that, substituting the single estimate into (3.7), we find that

$$\hat{\tau}_{\text{regression}} = \left(\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}\right) + \frac{n_T\bar{\vec{\boldsymbol{X}}}_T + n_C\bar{\vec{\boldsymbol{X}}}_C}{N}\left(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C\right)$$

The estimator is invariant to location – a shift of the empirical covariate distribution does not change the value of $\hat{\tau}_{\text{regression}}$, so for the sake of appealing interpretability, we mean center the covariates. Note that we mean-center with respect to the common, pooled mean, so that $\left(\vec{\boldsymbol{X}}_T\right)_i^* = \left(\vec{\boldsymbol{X}}_T\right)_i - \bar{\vec{\boldsymbol{X}}}$, with $\left(\vec{\boldsymbol{X}}_T\right)_i^*$ defined similarly. We thereby estimate the ATE for a covariate distribution with expectation equal to 0. From this we learn that the ATE can be estimated simply, via

$$\hat{\tau}_{\text{regression}} = \left(\hat{\beta}_T^{*(0)} - \hat{\beta}_C^{*(0)}\right) \tag{3.10}$$

Theorem 3.4.1 *$\hat{\tau}_{regression}$ is an asymptotically unbiased estimate of τ.*

Corollary 3.4.2 *$\mathbb{E}\left[\hat{\tau}_{regression}\right] = \tau$ when*

(a) The population response is linear in the covariates, and all covariates have been included in the statistical model, or

(b) $\mathbb{E}\left[T|X\right] = \mathbb{E}\left[C|X\right] + k$, and $n_T = n_C$, where $k \in \mathbb{R}$.

That is, if the treatment and control response functions are offset by a constant, then $\hat{\tau}_{\text{regression}}$ will be unbiased exactly, so long as the treatment and control sample sizes are equal. When they are unequal, the result continues to hold when the units are inversely reweighted. The proofs are deferred to the appendix.

The difference in intercepts (from a mean centered regression) enriches our understanding of the relationship between a single regression with interaction terms, and one without. In a single regression with no interactions, the ATE can be estimated via the least squares regression coefficient of the treatment indicator, which represents the constant gap between the treatment and control response surfaces. It is the difference of intercepts (that is, at $\vec{\boldsymbol{X}} = \mathbf{0}$), but it is also the difference in responses at any arbitrary $\vec{\boldsymbol{X}}$ value, the difference being constant. In a single regression with interaction, the gap between the response surfaces is allowed to vary, and depends on the location of those covariates interacting with the treatment indicator. What then, is the estimated ATE in the regression with interactions? It, too, is the coefficient of the treatment indicator. But how else can the treatment indicator be represented and understood? It, too, is equal to the estimated difference in intercepts. Why intercepts? Intercepts are what are left when the regression is evaluated at 0; and since we are evaluating at the average of the (pooled) mean-centered covariates, we are evaluating at $\mathbf{0}$.

When

$$I_T = \begin{cases} 1 & \text{Treatment is administered} \\ 0 & \text{Control is administered} \end{cases}$$

then in equations, the predicted response, when represented by a single regres-

sion with interactions, looks like

$$\hat{Y}_i = \hat{\beta}^{(0)} + \hat{\beta}^{(T)} I_T + \vec{\boldsymbol{X}}\hat{\boldsymbol{\beta}} + \vec{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{(Int)} I_T \tag{3.11}$$

With the covariates mean centered, substituting in the mean of the mean-centered covariates results in

$$\hat{Y}_i\Big|_{\vec{\boldsymbol{X}}=\bar{\vec{\boldsymbol{X}}}^*} = \hat{\beta}^{(0)} + \hat{\beta}^{(T)} I_T \tag{3.12}$$

for which, as described, $\hat{\beta}^{(T)}$ represents the difference in intercepts. Here, the coefficient of the treatment indicator is precisely equal to $\hat{\tau}_{\text{regression}}$.

Nowhere in the definition of the model were any assumptions made about the nature of the response variables. While a continuous response may have been implicitly assumed, the analysis is not altered if the T_i, C_i are assumed to be count data, or to take on values $0, 1$. When the response is binary, the target of estimation is still $\mathbb{E}[T] - \mathbb{E}[C]$, but these terms can now be rewritten as $P(T) - P(C)$, where $P(T)$ represents the proportion of treatment outcomes in the population that take on the value 1.

One hopes that the estimate $\hat{P}(T) - \hat{C}(V)$ should fall inside $[-1, 1]$. If one estimates $\hat{\tau}$ by the difference in means estimator, then such a desirable outcome is assured. However, $\hat{\tau}_{\text{regression}}$, since it estimates the response Y not at the respective sample means of the covariates $\vec{\boldsymbol{X}}_{iT}$ and $\vec{\boldsymbol{X}}_{iC}$ but at the weighted average $\frac{n_T\bar{\vec{\boldsymbol{X}}}_T + n_C\bar{\vec{\boldsymbol{X}}}_C}{N}$, $\hat{P}(T) - \hat{P}(C)$ is not guaranteed with probability one to be restricted to $[-1, 1]$. The problem arises if there is limited overlap between the observed treatment and control covariates, and the slope coefficients differ appreciably between the two groups. The probability associated with this possibility is small.

3.4.3 Relative performance of ATE estimators

We present in this section the expression for the variances of the difference-in-means and our regression based estimator, as well as for the standard error estimates, and compare the sizes of the variances.

The most familiar expression for $Var[\hat{\tau}_{\text{diff}}]$, of course, is $Var[T]/n_T + Var[C]/n_C$. For the purposes of comparison to $Var[\hat{\tau}_{\text{regression}}]$, the variance can be re-expressed by conditioning on covariates, and then marginalizing over their distribution, so that

Lemma 3.4.3

$$Var\left(\hat{\tau}_{diff}\right) = \left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + \frac{1}{n_T}\left[\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T\right] + \frac{1}{n_C}\left[\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C\right] \tag{3.13}$$

The proof is found in the appendix. The standard deviation of $\hat{\tau}_{\text{diff}}$ should be estimated by

$$\hat{SE}(\hat{\tau}_{\text{diff}}) = \sqrt{\frac{MSE_T}{n_T} + \frac{MSE_C}{n_C} + \frac{1}{n_T}\left(\hat{\boldsymbol{\beta}}_T \hat{\Sigma^{(T)}}_X \hat{\boldsymbol{\beta}}_T\right) + \frac{1}{n_C}\left(\hat{\boldsymbol{\beta}}_C \hat{\Sigma^{(C)}}_X \hat{\boldsymbol{\beta}}_C\right)} \tag{3.14}$$

In the above estimate, MSE_T is the mean square error computed in the treatment regression, defined as usual by $MSE_T = \left(\sum_{i=1}^{n}(T_i - \hat{T}_i)^2\right)/(N-p-1)$, and $\hat{\Sigma}_X$ is the empirical variance-covariance matrix of the complete collection of covariates.

The mean squared error is a scaled estimate of all the variability in the response that is not captured by the linear approximation. So the MSE is composed of two components: the estimate of the variability in the structural errors ϵ, together with the variability of $\eta(\vec{\boldsymbol{X}})$, the random variable measuring the non-linearity in the conditional mean.

$\hat{\tau}_{\text{regression}}$ also admits a clean variance expression:

Lemma 3.4.4

$$Var(\hat{\tau}_{regression}) = \left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + O(N^{-2}) + \frac{1}{N}(\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)'\Sigma_X(\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \tag{3.15}$$

The proof is deferred to the appendix.

The standard deviation of $\hat{\tau}_{\text{regression}}$ should be estimated by

$$SE(\hat{\tau}_{regression}) = \sqrt{\frac{MSE_T}{n_T} + \frac{MSE_C}{n_C} + \frac{1}{N}(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C)'\hat{\Sigma}_X(\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C)} \tag{3.16}$$

The more interesting claim follows: the asymptotic variance of the regression-based estimator dominates the variance of the naive estimator.

Theorem 3.4.5

$$AVar(\hat{\tau}_{diff}) \geq AVar(\hat{\tau}_{regression}) \tag{3.17}$$

The proof is found in the appendix.

To compare the relative asymptotic efficiencies of $\hat{\tau}_{\text{diff}}$ and $\hat{\tau}_{\text{regression}}$, only their respective variances need be compared because $\hat{\tau}_{\text{diff}}$ is a trivially unbiased estimate of τ, and, according to 3.4.1, $\hat{\tau}_{\text{regression}}$ is an asymptotically unbiased estimator of the ATE.

Tsiatis et al. (Tsiatis et al., 2008) also show that the estimator based on the model with interactions – they call it the $ANCOVA_2$ model – is efficient, and compare it with a large class of augmentation estimators. The estimator here can be extracted from the general class of estimators derived in (Zhang et al., 2008). Ours greatly simplifies the corresponding procedure detailed in their section 4, and makes explicit the comparison to the difference in means estimators. (Rosenblum and van der Laan, 2010) arrives at such an estimator through the technique of targeted maximum likelihood. Our variance expressions are clean and explicitly written down, so that the constituent parts of the variance are clearly seen.

The aim here is to describe the nature of the interaction model's efficiency and demonstrate which terms contribute to it. The inequality in 3.4.5 is not strict; and equality between the asymptotic variances can be attained, and is attained iff $\boldsymbol{\beta}_C = -\frac{n_C}{n_T}\boldsymbol{\beta}_T$. When the treatment and control sample sizes are equal, for example, then equality is attained when $\boldsymbol{\beta}_C = -\boldsymbol{\beta}_T$. In this case, when the treatment and control slopes are negative inverses of each other, the regression-based estimate of the ATE is maximally variable. This makes sense: sample estimates of the difference in intercepts are just as likely to be positive as to be negative, with equal probabilities of linearly increasing magnitudes of difference.

Theorem 1 refers, however, to the true variance of the respective estimators, rather than to their estimated variances[8]. The theorem could analogously have been written, and should be seen here for clarity, as

$$\mathbb{E}\left[\hat{Var}(\hat{\tau}_{\text{diff}})\right] \geq \mathbb{E}\left[\hat{Var}(\hat{\tau}_{regression})\right]$$

A remark on the seemingly different estimators of $\hat{\tau}_{\text{diff}}$. Every introductory statistics textbook will teach that

$$Var[\bar{T} - \bar{C}] = \frac{\mathbb{V}\text{ar}\,[T]}{n_T} + \frac{\mathbb{V}\text{ar}\,[C]}{n_C} \tag{3.18}$$

and that it is estimated unbiasedly – for example, for the purpose of hypothesis testing – by

$$\frac{{s^2}_T}{n_T} + \frac{{s^2}_C}{n_C} \tag{3.19}$$

In our chapter, we wrote different expressions for the variance and standard error estimates of $\hat{\tau}_{\text{diff}}$. This was done for ease of comparison. In fact, (3.4.3) and (3.18) are equal, as are (3.14) and (3.19), which are unbiased estimates thereof.

[8]Which means that in a given sample, $\hat{SE}(\hat{\tau}_{regression})$ may exceed $\hat{SE}(\hat{\tau}_{\text{diff}})$

3.4.4 Conditional and marginal estimation

We pause to make explicit the essential difference between conditional and marginal inference in our problem, and to emphasize the role of covariates that are here random. The variance of the difference-in-means estimator is a marginal variance: over all conceivable repetitions of the experiment, as new subjects are sampled and assigned a treatment or a control condition, irrespective of any other measured or unmeasured covariates,

$$Var[\bar{T} - \bar{C}] = \frac{Var[T]}{n_T} + \frac{Var[C]}{n_C}. \tag{3.20}$$

It is estimated, unbiasedly, by ${s^2}_T/n_T + {s^2}_C/n_C$.

Now,as in our problem, measure covariates, and run two separate regressions, so that $\hat{T} = \hat{\beta}_T^{(0)} + \vec{\boldsymbol{X}}_T\hat{\boldsymbol{\beta}}_T$, and $\hat{C} = \hat{\beta}_C^{(0)} + \vec{\boldsymbol{X}}_C\hat{\boldsymbol{\beta}}_C$. From elementary regression, if we estimate the response at the mean of the predictors, then $\hat{T}_i\Big|_{\vec{\boldsymbol{X}}_T=\vec{\boldsymbol{X}}_{\boldsymbol{T}}} = \bar{T}$, and $\hat{C}_i\Big|_{\vec{\boldsymbol{X}}_C=\vec{\boldsymbol{X}}_{\boldsymbol{C}}} = \bar{C}$. Apparently, in estimating the ATE, $\bar{T} - \bar{C} = \hat{T}_i\Big|_{\vec{\boldsymbol{X}}_T} - \hat{C}_i\Big|_{\vec{\boldsymbol{X}}_C}$, so the variance should depend on the the observed covariates! What, then, is the proper variance of $\bar{T} - \bar{C}$? Is it the same as that reported in (3.20)?

It will not be equal, for the simple reason that the classical variance is considered conditional on the observed covariates. To compute, note that $\bar{T}$ is independent of $\bar{C}$, so let us for the moment consider just $Var[\bar{T}]$. $\bar{T}$ was estimated in a regression at a specific covariate value. For ease of exposition, recall the prediction variance from simple regression, where

$$\hat{Var}[\hat{y}|X = x_p] = MSE\left[1 + \frac{1}{n_T} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n_T}(x_i - \bar{x})^2}\right] \tag{3.21}$$

That is to say, at the covariate mean,

$$\hat{Var}[\hat{T}|\vec{\boldsymbol{X}} = \vec{\boldsymbol{X}}_T] = MSE\left[1 + \frac{1}{n_T}\right] \tag{3.22}$$

which, of course, does not uniformly equal $\frac{{s^2}_T}{n_T}$. As a matter of fact, the two estimated

variances will be equal only when the R^2 from the regression exceeds $\frac{p+2}{n_T+1}$, where p is the number of covariates; then the regression based estimated variance will be smaller than that of the marginal, conventional estimated variance. The reason for this discrepancy, for how the relative variances of ostensibly the same statistic depend on the quality of the fit, is simple.

The variance estimated in (3.22) relies on classical regression theory, where the predictors are assumed to be fixed from one realization of the data to the next. Inference is therefore conditional on the covariates; the estimate of the variance of $\bar{T}$ in (3.22) is *conditional* on being estimated at the (here, fixed) mean of the covariates. It is saying: when the mean of the covariates is equal exactly to the mean of the covariates in this sample, what is the variability of the average response? What is unaccounted for is that that selfsame covariate mean is a random quantity, and its variability will contribute to the variability in the average response. This naive regression based estimate (3.22), therefore, artificially deflates the true variance of the response mean. In our analysis we compare two marginal variances, from which an inequality follows that holds for all fits.

3.4.5 Alternative Conditions

3.4.5.1 Distribution of $\boldsymbol{X}$ known

Throughout the discussion and analysis, we have assumed that the underlying distribution of $\boldsymbol{X}$ is unknown. The alternative may present in practice where, for example, covariates like age, weight and income, for which measurements exists in the whole population, are used in the study. In such a case, the variability inherent in estimating $\mathbb{E}[\boldsymbol{X}]$ is removed (only the regression slopes remain to be estimated), with a corresponding diminution of the standard error of the ATE. The precise degree to which the standard error diminishes can be found in the appendix.

3.4.5.2 Treatment Correlated with Covariates

In the preceding discussion, we had assumed that the assignment of treatment (the treatment indicator) was independent of the covariates, with correlation among them presenting itself only in samples. It is conceivable and natural, however, that the decision to administer treatment should depend on the covariates: perhaps, by design and because of cost constraints in the study, the researcher wishes to offer expensive treatment to a higher proportion of those suspected to require it for a shorter duration.

Precisely, suppose that the regression is written as in 3.9, except that $I_T = H(\vec{\boldsymbol{X}})$, either deterministically or stochastically, as when $I_T \sim Bern\left(H\left(\vec{\boldsymbol{X}}\right)\right)$. The treatment indicator is a function of the covariates so the assignment mechanism is different across different strata. In this case, the functional form of $H(\cdot)$ is known, so that $\pi_i = P\left(I_T = 1|\vec{\boldsymbol{X}}\right)$ does not need to be estimated.

With the goal of estimating the ATE, an inverse probability weighting scheme is natural because it can reduce the bias that would result from the differing sampling regimes across strata. Accordingly, reweight the observed response y_i according to

$$y_i^{(T)*} = \frac{y_i^{(T)}}{\pi_i}$$

with π_i defined as above for the treated units, and

$$y_i^{(C)*} = \frac{y_i^{(C)}}{1 - \pi_i}$$

Such a reweighting has been considered by, for example, (Freedman and Berk, 2008), except the functional relationship between the confounders and the treatment indicator was unknown and was consequently estimated via propensity scores. Our future work will extend to cases when this functional relationship needs to be estimated.

One proceeds with the analysis as before, running the two separate treated and control regressions, estimating the (weighted response) at the pooled mean of the covariates, and

taking the difference. Another estimate of the ATE would be

$$\frac{1}{n_T}\sum_{y_i^{(T)*}=1}^{n} -\frac{1}{n_C}\sum_{y_i^{(C)*}=1}^{n} \tag{3.23}$$

, what (Freedman and Berk, 2008) call a weighted contrast, and is the weighted variant of the difference in means estimator considered earlier. The latter is a Horvitz-Thompson type estimator (the formal H-T estimator assumes a finite population from which one samples). The derivations and analysis relating to the weighted scheme are beyond the scope of the current chapter, and will be considered in depth in a forthcoming work.

3.4.5.3 Stratification

The results described in the preceding sections make no assumptions about the nature of the covariates, which may be discrete, continuous, or both. An interesting special case arises when, besides the treatment indicator, the other covariates represent stratum assignment, and interactions are permitted between the treatment indicators and assignment indicators. For example, subjects may be classified by treatment/control, and highest degree of educational attainment (no high school, high school, college, etc.) The result of this pre-stratification is a two-way ANOVA layout, with interactions. In the familiar ANOVA form, the regression model may be described by

$$Y_{ijk} = \mu + s_i + \tau_j + (s\tau)_{ij} + \epsilon_{ijk} \tag{3.24}$$

s_i is the ith stratum, $i = 1, \ldots I$, τ_j is the treatment effect, $j = 0, 1$ (WLOG, let j = 1 when treatment is administered), and $(s\tau)_{ij}$ is the interaction effect. Denote the number of patients in stratum i receiving regime j by K_{ij}.

The difference-in-means estimator is written simply as

$$\bar{\mu} = \bar{Y}_{.1.} - \bar{Y}_{.0.} \tag{3.25}$$

and is unbiased, since $\mathbb{E}[\bar{\mu}] = \mathbb{E}[Y_{\cdot 1}] - \mathbb{E}[Y_{\cdot 0}]$.

Now define the local ATEs, which represent the respective within-stratum ATEs by

$$ATE_i \equiv \theta_i = \mathbb{E}[Y_{i1} - Y_{i0}].$$

The second estimator weights the per-stratum difference-in-means by the proportion of the sample found in each stratum:

$$\tilde{\mu} = \sum_{i=1}^{I} (\bar{Y}_{i1} - \bar{Y}_{i0}) * \hat{p}_i \tag{3.26}$$

where $\hat{p}_i$ is the sample proportion of all subjects in stratum i; it is equivalently written as $\frac{K_{i+}}{K_{++}}$. $\mathbb{E}[\tilde{\mu}] = \sum_{i=1}^{I} p_i \theta_i = \theta$, so it is also unbiased. The estimator is unbiased under randomized assignment and under blocking since in both instances, the proportion of treated cases in a stratum is independent of the mean, and in both cases, $\mathbb{E}[\hat{p}_i] = \theta_i$. As in (Miratrix et al., 2013), which gives an impressive treatment of *post*-stratification in the Neyman framework, the ATE estimate here is assumed to be well-defined – that is, the estimator is computed conditional on the event that each stratum is populated by at least one treated and one control unit. This second estimator just described is precisely $\hat{\tau}_{\text{regression}}$. Our results, in particular Lemma 4.4 and Theorem 4.5 continue to hold. Under slightly modified conditions, (Miratrix et al., 2013) and (Imbens, 2011) show, for example, that its variance is less than that of the difference-in-means estimator, and is higher than the variance resulting from blocking (or pre-stratification) on an order of $O\left(N^{-2}\right)$.

3.5 Illustration on real data

We present a typical application of our regression based ATE estimator on real data. We illustrate the performance of the estimator on data furnished from a classic study discussed in (LaLonde, 1986) and reanalyzed in (Dehejia and Wahba, 1999). The data in question

come from the National Support Work (NSW) Demonstration. A pool of adults with economic and social problems was randomized into two groups. The treated group was offered job training while the control group was not. The intent of the work in (LaLonde, 1986) was to compare ATE estimates from experiments to those from observational studies. He compared the unbiased estimate of the ATE from NSW groups to an estimate drawn by comparing the treated adults to a batch of controls collected from separate comparison groups (PSID-1 and CPS-1 in his paper). Dehejia and Wahba (Dehejia and Wahba, 1999) apply matching techniques for this comparison; relevant for our work are the 185 treated and 260 control male subjects they analyze, and which are available from the original NSW experiment.

The following covariates were adjusted for: age, education (number of years), an indicator for black, indicator for hispanic, indicator for marital status, indicator for high school degree, and earnings in 1974. The response measured was earnings in 1978, after the job training had concluded.

In this experimental context the difference in means is equal to 4709.4 dollars, with a standard error equal to 443.5. The regression based method yields a point estimate of $\hat{\tau} = 4435.2$ dollars, with an SE estimate of 431.9. The gain in SE amounts to 3.1%, this when the R^2 of the regression of reservation price on covariates and their interaction with the treatment indicator was 0.24. A gain of this magnitude is typical for an R^2 of this size. Higher R^2 results in higher SE gains, which is vividly demonstrated in the following section.

3.5.1 Illustration on simulated data

The datasets on RCTs we have encountered have come with an R^2 that doesn't far exceed 0.2. To more vividly illustrate the results obtained in this chapter, we considered the following model. The treated and control groups were defined, respectively, by

$$T = 2X_1 + 3X_2 + Z_T \tag{3.27}$$

$$C = X_1 + X_2 + Z_C \tag{3.28}$$

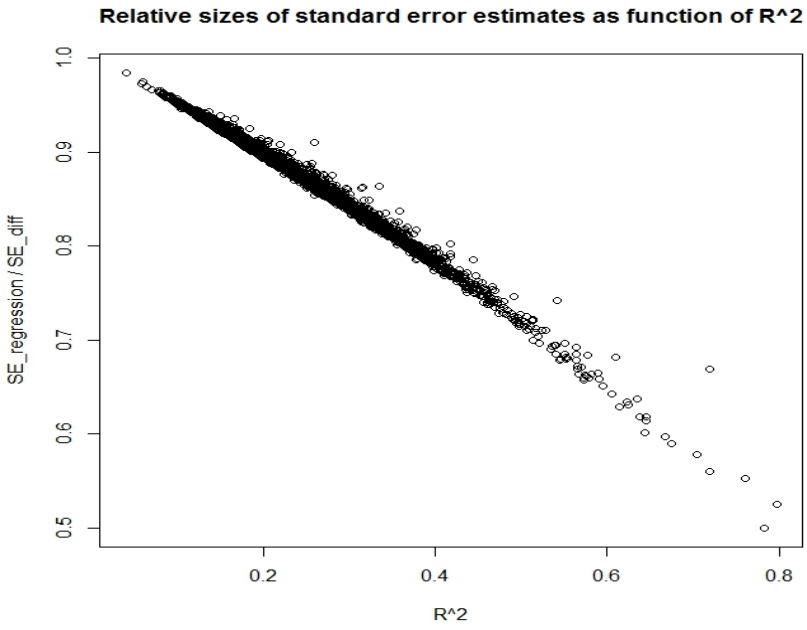
where $X_1 \sim Lognormal(0,1)$, $X_2 \sim Gamma(3,4)$, and $Z_T, Z_C \overset{iid}{\sim} N(0,3)$. Under these conditions, $\mathbb{E}[T] - \mathbb{E}[C] = 2e^{1/2} - 3/2 = 1.797$. We simulated 10,000 times, with 250 tr.eated and 250 control units in each simulation, and recorded the R^2 of the combined regression, as well as the ATE and SE estimates for both the difference-in-means, and for the regression based estimator considered in this chapter. The average R^2 in the 10,000 simulation was 0.75. Accordingly, the average $\hat{SE}(\hat{\tau}_{\text{diff}}) = 0.676$ (with simulation SE = 0.0011), while the average $\hat{SE}(\hat{\tau}_{\text{regression}}) = 0.332$ (with simulation SE = 0.0002). Both estimators were unbiased (up to simulation granularity), with difference-in-mean and regression-based average ATEs equal to 1.798 and 1.796, respectively. Coverage of the true ATE was equal to 0.9473 and 0.949, respectively, when using $\Phi^{-1}(0.975)$ as the multiplier. The regression based estimate naturally leads to a more powerful test. There was nothing particular about the model chosen; similar phenomena are observed for other choices of underlying distribution.

As a final illustration, we show the relationship between the R^2 from the combined model and the respective standard error estimates. $\hat{\tau}_{\text{diff}}$, depending only on the response, does not depend on the quality of the regression fit. $\hat{\tau}_{\text{regression}}$, however, does. 10,000 simulations were again run, except the variance of Z_T, Z_C was dialed from 1 to 100, with attendant decreases in the R^2. The plot of R^2 against $\hat{SE}(\hat{\tau}_{\text{regression}}) / \hat{SE}(\hat{\tau}_{\text{diff}})$ is shown. As R^2 decreases, the estimated standard errors converge. For high R^2, the $\hat{\tau}_{\text{regression}}$ enjoys a dramatically lower standard error.

3.6 Conclusion

This chapter lays the foundation for conducting principled and efficient asymptotic inference on ATEs. After acknowledging the aesthetics but also limitations of the Neyman paradigm, and the unreality of fixed X, we turned our focus to an infinite population, random de-

Figure 3.1: R^2 plotted against $\frac{\hat{SE}\left(\hat{\tau}_{\text{regression}}\right)}{\hat{SE}(\hat{\tau}_{\text{diff}})}$



sign, regression based estimation, where the response surface needn't be linear. Since the regression covariates are seen as random, generated from a distribution, the formulation is a more realistic representation of the practice of random sampling: randomness arises not only from the random assignment of treatment and control to subjects, but also from these subjects' (random) characteristics as well. Despite the added source of variability, the derived standard error, which takes into account these sources of randomness but also adjusts for covariates, is in expectation actually lower than its conventional counterpart.

Bootstrapped confidence intervals can easily be generated and inference conducted for the population ATE. Moreover, the paired bootstrap, mimicking as it does the random X framework, is the natural technique for such intervals. Future work will focus on weighting schemes when the treatment is correlated with covariates, as it would be, for example, in observational studies. In this work we estimated with linear models. We hope to extend the work to GLMs.

3.7 Technical appendix

Derivation of fact in footnote 5: in brief, that $\mathbb{E}\left[\eta_T(\vec{\boldsymbol{X}})\right] = 0$ follows from $\mathbb{E}\left[\eta_T\vec{\boldsymbol{X}}\right] = 0$. $\vec{\boldsymbol{X}}$, as defined, contains an intercept; and since the expectation of the dot product of η_T with a vector of ones must be zero, then $\mathbb{E}\left[\eta_T\vec{\boldsymbol{X}}\right] = 0$ is equivalent to saying that $\mathbb{E}\left[\eta_T\right] = 0$

Proof of 3.4.1

After mean centering, $\hat{\tau}_{\text{regression}} = \left(\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}\right)$. Direct application of the proposition on page 11 in (Buja et al., 2013) shows that the difference of the independent quantities $\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}$ is an unbiased estimate of $\beta_T^{(0)} - \beta_C^{(0)}$, which is equal to τ when $\boldsymbol{\mu} = \mathbf{0}$.

Proof of 3.4.2

(a) When the regression model is correctly specified, then it is an introductory result that the LS estimates are unbiased: $\mathbb{E}\left[\hat{\beta}_T^{(0)}\right] = \beta_T^{(0)}$ and that $\mathbb{E}\left[\hat{\beta}_C^{(0)}\right] = \beta_C^{(0)}$, so $\mathbb{E}\left[\hat{\beta}_T^{(0)} - \hat{\beta}_C^{(0)}\right] = \beta_T^{(0)} - \beta_C^{(0)} = \tau$.

(b) Suppose that the treatment and response surfaces have a constant offset: $n_T = n_C$ and $\mathbb{E}\left[T|X\right] = \mathbb{E}\left[C|X\right] + k$. In the decomposition of $\hat{\tau}_{\text{regression}} - \tau$ in the proof of 3.4.4, the only term which does not generally have expectation $\mathbf{0}$ is the term denoted by R_2, and equal to $\left[\bar{\boldsymbol{X}}_T - \bar{\boldsymbol{X}}_C\right]\left[p_C(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_T) + p_T(\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C)\right]$. It will have expectation 0 when the two bracketed terms are uncorrelated. Exploiting the independence between the treated and control groups, the bracketed terms will be uncorrelated iff

$$p_C Cov\left(\bar{\boldsymbol{X}}_T, \hat{\boldsymbol{\beta}}_T\right) = p_T Cov\left(\bar{\boldsymbol{X}}_C, \hat{\boldsymbol{\beta}}_C\right) \tag{3.29}$$

Inversely weight the observations, giving weight $\frac{1}{n_T}$ to the control observations, and $\frac{1}{n_C}$ to the treatment, so that (3.29) will hold true when $Cov\left(\bar{\boldsymbol{X}}_T, \hat{\boldsymbol{\beta}}_T\right) = Cov\left(\bar{\boldsymbol{X}}_C, \hat{\boldsymbol{\beta}}_C\right)$ When $\boldsymbol{\beta}_C = \boldsymbol{\beta}_T$, then, since the $\bar{\boldsymbol{X}}_T$ and $\bar{\boldsymbol{X}}_C$ are identically distributed, the above equality will hold. $\boldsymbol{\beta}_C = \boldsymbol{\beta}_T$ when there is a constant offset.

Proof of 3.4.3

The conventional estimator of the ATE is $\hat{\tau}_{\text{diff}} = \bar{T} - \bar{C}$. Assume the covariates have

zero mean; then its difference from the true ATE equals

$$\begin{aligned}\hat{\tau}_{\text{diff}} - \tau &= \bar{T} - \bar{C} - \left(\beta_T^0 - \beta_C^0\right) \\ &= \left[\bar{T} - \left(\beta_T^0 + \bar{X}_T\boldsymbol{\beta}_T\right)\right] - \left[\bar{C} - \left(\beta_C^0 + \bar{X}_C\boldsymbol{\beta}_C\right)\right] \\ &+ \bar{X}_T\boldsymbol{\beta}_T - \bar{X}_C\boldsymbol{\beta}_C \end{aligned} \tag{3.30}$$

The two terms – the former the residual means, and the latter a function of the covariates – are independent. Hence

$$\begin{aligned} Var\left(\hat{\tau}_{\text{diff}}\right) &= Var\left\{\left[\bar{T} - \left(\beta_T^0 + \bar{X}_T\boldsymbol{\beta}_T\right)\right] - \left[\bar{C} - \left(\beta_C^0 + \bar{X}_C\boldsymbol{\beta}_C\right)\right]\right\} \\ &+ Var\left\{\bar{X}_T\boldsymbol{\beta}_T - \bar{X}_C\boldsymbol{\beta}_C\right\} \\ &= \left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + \frac{1}{n_T}\left[\boldsymbol{\beta}'_T\Sigma_{X_T}\boldsymbol{\beta}_T\right] + \frac{1}{n_C}\left[\boldsymbol{\beta}'_C\Sigma_{X_C}\boldsymbol{\beta}_C\right] \\ &= \left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + \frac{1}{n_T}\left[\boldsymbol{\beta}'_T\Sigma_X\boldsymbol{\beta}_T\right] + \frac{1}{n_C}\left[\boldsymbol{\beta}'_C\Sigma_X\boldsymbol{\beta}_C\right] \end{aligned}$$

as the covariance matrices of the treatment and control distributions are equal, since the covariates are drawn from the same distribution.

Proof of 3.4.4

As before, we allow for unequal randomization, so that n_T cases receive treatment, and n_C cases receive control; denote the proportions p_T and p_C, respectively, and suppose that $\mathbb{E}[X] = \boldsymbol{\mu}$ and $Var[X] = \Sigma$. Denote the ATE by τ. The ATE in the population, τ, equals $\mathbb{E}[T] - \mathbb{E}[C] = (\beta_T^0 - \beta_C^0) + \boldsymbol{\mu}(\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)$. Then

$$\hat{\tau}_{regression} = \hat{\beta}_T^0 - \hat{\beta}_C^0 + \hat{\boldsymbol{\mu}}(\hat{\boldsymbol{\beta}_T} - \hat{\boldsymbol{\beta}_C})$$

$$
\begin{aligned}
\hat{\tau}_{regression} &= \hat{\beta}_T^0 - \hat{\beta}_C^0 + \left[p_T \bar{\boldsymbol{X}}_T + p_C \bar{\boldsymbol{X}}_C\right] (\hat{\boldsymbol{\beta}_T} - \hat{\boldsymbol{\beta}_C}) \\
&= \bar{T} - \bar{\boldsymbol{X}}_T \hat{\boldsymbol{\beta}}_T - (\bar{C} - \bar{\boldsymbol{X}}_C \hat{\boldsymbol{\beta}}_C) + \left[p_T \bar{\boldsymbol{X}}_T + p_C \bar{\boldsymbol{X}}_C\right] (\hat{\boldsymbol{\beta}}_T - \hat{\boldsymbol{\beta}}_C) \\
&= \bar{T} - \bar{C} - \left(\bar{\boldsymbol{X}}_T - \bar{\boldsymbol{X}}_C\right) \left(p_C \hat{\boldsymbol{\beta}}_T + p_T \hat{\boldsymbol{\beta}}_C\right)
\end{aligned}
$$

The multivariate mean can be taken to equal $\mathbf{0_p}$ WLOG since the problem is one of scale, rather than location. So

$$
\begin{aligned}
\hat{\tau}_{regression} - \tau &= \bar{T} - \bar{C} - \left(\bar{\boldsymbol{X}}_T - \bar{\boldsymbol{X}}_C\right) \left(p_C \hat{\boldsymbol{\beta}_T} + p_T \hat{\boldsymbol{\beta}_C}\right) - \beta_T^0 + \beta_C^0 \\
&= \left[\bar{T} - (\beta_T^0 + \bar{\boldsymbol{X}}_T \boldsymbol{\beta}_T)\right] - \left[\bar{C} - (\beta_C^0 + \bar{\boldsymbol{X}}_C \boldsymbol{\beta}_C)\right] \\
&- \left(\bar{\boldsymbol{X}}_T - \bar{\boldsymbol{X}}_C\right) \left[p_C (\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta}_T) + p_T (\hat{\boldsymbol{\beta}}_C - \boldsymbol{\beta}_C)\right] \\
&+ (p_T \bar{\boldsymbol{X}}_T + p_C \bar{\boldsymbol{X}}_C)(\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \\
&= R_1 + R_2 + R_3 x \qquad (3.31)
\end{aligned}
$$

R_1, R_2, and R_3 are independent: R_1 is a function of the errors, which are independent of the covariates, while R_2 and R_3 lie in the column space of the covariates. R_2 is uncorrelated with R_3 because [we have the correlation between sums and differences of i.i.d variables. Check the math again]. Moreover, each of the terms has expectation $\mathbf{0_p}$: the first, R_1, is a difference of average errors, equal to $\left(\bar{\epsilon}_T + \bar{f}_T\right) - \left(\bar{\epsilon}_C + \bar{f}_C\right)$. The ϵ have expectation 0 by assumption, and the f by construction. R_2 is asymptotically equal to $\mathbf{0}$, for the following reason: the treatment and controls are uncorrelated, and $\mathbb{E}\left[\bar{\boldsymbol{X}}\right] = \mathbf{0}$, so the only component of R_2 not equal for all n to $\mathbf{0}$ in expectation is $p_C \bar{\boldsymbol{X}}_T \hat{\boldsymbol{\beta}}_T - p_T \bar{\boldsymbol{X}}_C \hat{\boldsymbol{\beta}}_C$. We'll now show that $\mathbb{E}\left[\bar{\boldsymbol{X}}_T \hat{\boldsymbol{\beta}}_T\right] \to \mathbf{0}$:

$$
\begin{aligned}
\mathbb{E}\left[\bar{\boldsymbol{X}}_T \hat{\beta}\right] &= \mathbb{E}\left[\bar{\boldsymbol{X}}_T \mathbb{E}\left[\hat{\beta} | \boldsymbol{X}_T\right]\right] \\
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T \left(\boldsymbol{X}_T' \boldsymbol{X}_T\right)^{-1} \boldsymbol{X}_T' \mathbb{E}\left[Y | \boldsymbol{X}_T\right]\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T \left(\boldsymbol{X}_T'\boldsymbol{X}_T\right)^{-1} \boldsymbol{X}_T' \left(\boldsymbol{X}_T\boldsymbol{\beta}_T + \eta_T(\boldsymbol{X}_T)\right)\right] \\
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T \left(\boldsymbol{X}_T'\boldsymbol{X}_T\right)^{-1} \boldsymbol{X}_T'\boldsymbol{X}_T\boldsymbol{\beta}_T + \bar{\boldsymbol{X}}_T \left(\boldsymbol{X}_T'\boldsymbol{X}_T\right)^{-1} \eta_T(\boldsymbol{X}_T)\right] \\
&= \mathbb{E}\left[\bar{\boldsymbol{X}}_T\boldsymbol{\beta}_T\right] + \mathbb{E}\left[\bar{\boldsymbol{X}}_T \left(\boldsymbol{X}_T'\boldsymbol{X}_T\right)^{-1} \eta_T(\boldsymbol{X}_T)\right]
\end{aligned}
$$

The first terms is equal to $\mathbf{0}$ because $\mathbb{E}[\boldsymbol{X}] = \mathbf{0}$ by assumption. The second term is equal to $\mathbf{0}$ because $\eta_T(\boldsymbol{X}_T)$ is uncorrelated with the covariates and itself has expectation zero.

$\mathbb{E}[R_3] = 0$ because $\mathbb{E}[\boldsymbol{X}] = \mathbf{0}$. So

$$
\begin{aligned}
Var(\hat{\tau}_{\text{regression}}) &= \mathbb{E}\left[R_1^2\right] + \mathbb{E}\left[R_2^2\right] + \mathbb{E}\left[R_3^2\right] \\
&= \left\{\left(\mathbb{E}\left[\bar{\epsilon}_T^2\right] + \mathbb{E}\left[\bar{f}_T^2\right]\right) + \left(\mathbb{E}\left[\bar{\epsilon}_C^2\right] + \mathbb{E}\left[\bar{f}_C^2\right]\right)\right\} + O(N^{-2}) \\
&+ (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(p_T^2 \frac{\Sigma_{X_T}}{n_T} + p_C^2 \frac{\Sigma_{X_C}}{n_C}\right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \\
&= \left(\frac{\sigma_T^2}{n_T} + \frac{Var[\eta_T]}{n_T}\right) + \left(\frac{\sigma_C^2}{n_C} + \frac{Var[\eta_C]}{n_C}\right) + O(N^{-2}) \\
&+ (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(p_T \frac{\Sigma_{X_T}}{N} + p_C \frac{\Sigma_{X_C}}{N}\right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \\
&= \left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + O(N^{-2}) + (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(\frac{\Sigma_X}{N}\right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)
\end{aligned}
$$

The last line follows since $\Sigma_{X_T} = \Sigma_{X_C} = \Sigma_X$ – they are all variances of a common distribution. ■

Proof of 3.4.5.1 Suppose now that the distribution of $\boldsymbol{X}$ is known. Its mean can be assumed to be $\mathbf{0}$ WLOG. Then $\tau = \beta_T^0 - \beta_C^0$ and $\hat{\tau}_{regression} = \hat{\beta}_T^0 - \hat{\beta}_C^0$, so that, using a similar rearrangement as before,

$$
\begin{aligned}
\hat{\tau}_{regression} - \tau &= \left(\bar{T} - \hat{\boldsymbol{\beta}}_T\bar{\boldsymbol{X}}_T\right) - \left(\bar{C} - \hat{\boldsymbol{\beta}}_C\bar{\boldsymbol{X}}_C\right) - (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \\
&= \left[\bar{T} - \left(\beta_T^0 + \bar{\boldsymbol{X}}_T\boldsymbol{\beta}_T\right)\right] - \left[\bar{C} - \left(\beta_C^0 + \bar{\boldsymbol{X}}_C\boldsymbol{\beta}_C\right)\right] \\
&+ \bar{\boldsymbol{X}}_T\left(\boldsymbol{\beta}_T - \hat{\boldsymbol{\beta}}_T\right) - \bar{\boldsymbol{X}}_C\left(\boldsymbol{\beta}_C - \hat{\boldsymbol{\beta}}_C\right)
\end{aligned}
$$

$$= \quad R_1 + R_2^* \tag{3.32}$$

Direct comparison of 3.32 with 3.31 will show that the estimated ATE is also asymptotically unbiased, and that its asymptotic variance is decreased by the value of R_3, and some of R_2. With R_3 omitted, the standard error of the regression can just be estimated by $\sqrt{\frac{MSE_T}{n_T} + \frac{MSE_C}{n_C}}$

Proof of 3.4.5

We now verify that the standard error of the proposed estimator dominates the standard error estimator of the conventional ATE. We compare, therefore,

$$\left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + O(N^{-2}) + (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \left(\frac{\Sigma_X}{N}\right) (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)$$

to

$$\left[\frac{\sigma_T^2 + Var[\eta_T]}{n_T} + \frac{\sigma_C^2 + Var[\eta_C]}{n_C}\right] + \frac{1}{n_T} [\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T] + \frac{1}{n_C} [\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C]$$

We easily show that the asymptotic variance of the conventional estimator is higher than that of the regression estimator by comparing the variance components that differ among the two equations, noting that the $O(N^{-2})$ term vanishes.

$$\begin{aligned}
\left(\sqrt{\frac{n_C}{n_T}}\boldsymbol{\beta}_T + \sqrt{\frac{n_T}{n_C}}\boldsymbol{\beta}_C\right)' \Sigma_X \left(\sqrt{\frac{n_C}{n_T}}\boldsymbol{\beta}_T + \sqrt{\frac{n_T}{n_C}}\boldsymbol{\beta}_C\right) \quad &\geq \quad 0 \qquad (3.33)\\
\frac{n_C}{n_T} (\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T) + 2\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_C + \frac{n_T}{n_C} (\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C) \quad &\geq \quad 0\\
\frac{N}{n_T}\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T + \frac{N}{n_C}\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C \geq \boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T - 2\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_C \quad &+ \quad \boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C\\
\frac{1}{n_T} [\boldsymbol{\beta}_T' \Sigma_X \boldsymbol{\beta}_T] + \frac{1}{n_C} [\boldsymbol{\beta}_C' \Sigma_X \boldsymbol{\beta}_C] \geq (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C)' \quad &\left(\tfrac{\Sigma_X}{N}\right) \quad (\boldsymbol{\beta}_T - \boldsymbol{\beta}_C) \blacksquare
\end{aligned}$$

The only non-algebraic step is in the first line, which is true because the LHS is a quadratic form. Equality is attained iff $\boldsymbol{\beta}_C = -\frac{n_C}{n_T}\boldsymbol{\beta}_T$, which can be verified by direct

substitution into (3.33).

Proof of remark on R^2 following equation (3.22):

$Var(\bar{T}) = \frac{SST}{n_T}$, whereas the regression based variance at the covariate mean is estimated by $MSE_T[1 + \frac{1}{n_T}]$, which can be rewritten as $\frac{SST-SSR}{n_T-p-1} \times \left(\frac{n_T+1}{n_T}\right)$ Dividing both expressions by SST leads us to compare $\frac{1}{n_T}$ to $\frac{1-R^2}{n_T-p-1} \times \left(\frac{n_T+1}{n_T}\right)$. Equality is attained when R^2 is equal to $\frac{p+2}{n_T+1}$

4 Calibrated Prediction Intervals

4.1 Abstract

Regression models are often fit even when regression assumptions are violated. Assuming only a joint distribution between $\boldsymbol{X}, Y$, and nothing about the nature of their relationship besides mild regularity conditions, we offer coverage guarantees when their relationship is modeled by a regression. In this light, we describe a procedure for constructing intervals that capture $(1-\alpha)$ of future observations and prove its validity. The procedure is valid marginally over the $\boldsymbol{X}$ distribution, even in the presence of severe misspecification. Several variants of the procedure, calibrated in-sample and via bootstrapped resampling, are proposed and found to exhibit similar behavior.

4.2 Introduction

Classical linear model theory rests on the bedrock of these assumptions: the conditional mean of the response is linear in the predictors; errors are normal, uncorrelated, and homoscedastic; and each future observation will have the same x-coordinates as some observation in the design. In recompense for these assumptions, classical theory delivers guarantees about the conditional value of a future response:

$$P\left(y_{new} \in \hat{y} \pm t_{\alpha/2,n-p-1}\hat{\sigma}\sqrt{1+\vec{\boldsymbol{x}}_i\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\vec{\boldsymbol{x}}_i^T}\right) = 1-\alpha \tag{4.1}$$

In the preceding expression, it is assumed that $y = \boldsymbol{X}\boldsymbol{\beta}+\epsilon, \epsilon \sim N\left(\vec{0}_N, \sigma^2\boldsymbol{I}_{N\times N}\right)$, $\boldsymbol{X}_{N\times(p+1)}$ is the design matrix with N observations, p predictors, and a column of ones prepended, and $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T; \hat{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the usual least squares estimate of $\boldsymbol{\beta}$, and the $\vec{\boldsymbol{x}}_i$ are row vectors of the design matrix, $i = I, \ldots N$. More concisely, it is assumed that the model is first- and second-order correct, and that the design is fixed.

Experience with data teaches that the assumptions underlying regression are easily and commonly violated. Besides replicated laboratory experiments, few designs can honestly claim to be fixed. Whereas in a bivariate regression visual inspection suffices to establish linearity, with added predictor variables it is difficult to discern whether the relationship between a predictor variable – adjusted for all other predictor variables – and the response is linear. It can also be unclear which transformation of the unadjusted $\vec{\boldsymbol{x}}_i$ and y is best suited for the modeling task. And so in practice, the association between the response and some of the predictors can be decidedly nonlinear. Errors, too, are not always well behaved. And the collection of covariates in the probability model might not correspond to the covariate relationship it is intended to model – that is, variables may be omitted.

Researchers use regression, even when its assumptions are violated. Such practice is warranted when the purpose of the analysis is exploratory and the researcher wants to understand general relationships and patterns lying in the data. On the other hand, inferential statistics lose their validity; depending on the nature and degree of the model violations, point estimates will be biased, and intervals will not give desired coverage.

In this chapter we focus our attention on prediction intervals, both for individual

future responses, and also for mean responses. In 4.3 and 4.4, we propose a class of methods which are robust to departures from model assumptions; even in the presence of non-linearity, heteroscedasticity, and random design, our prediction intervals cover at the promised confidence level. In one variant, conditioning on the fitted regression, intervals are nonparametrically calibrated in-sample. An alternate method can properly be called a resampling-based calibration technique: intervals are computed for each resampled dataset, and a functional of these intervals is the one applied to the model. Various shapes for the prediction intervals are considered, but the conclusions are general. In 4.6 we address the problem of covering the mean response.

The prediction interval procedure proposed yields promised coverage that is *marginal*: it ensures that, in probability, a pre-specified proportion $(1 - \alpha)$ of future observations will fall within the constructed interval. We do not condition on the $\boldsymbol{X}$-coordinates and insist that $(1 - \alpha)$ of future observations be covered at a given $\boldsymbol{X} = \vec{\boldsymbol{x}}$, and those insisting on robust conditional coverage will not find their answer here. But the need for marginal coverage does find justification in the world as it is: when the cost of not covering an observation is the same across all observations, the marginal criterion is the appropriate criterion. For example, suppose a researcher wishes to predict tax revenues for the following year in different cities. He constructs a regression based on covariates measured this year, constructs uncertainty intervals, and demands 95% certainty that his predictions for next year will fall within his intervals. It is enough for him to know the expected number of cities which revenues falling short of or exceeding these interval boundaries; the identity of the cities falling outside the interval are immaterial to him. Then a marginal coverage guarantee is exactly what the researcher seeks and finds useful for his aims.

Moreover, as examined in chapter 2, conditional inference in the classical sense remains valid so long as the population response surface is linear in the predictors. But

when the covariates are seen not as fixed, but as generated from a joint distribution, and there are departures from model assumptions, then ancillarity, the argument according to which the distribution of the predictors does not affect inference on the model parameters, ceases to hold. Congruently with the marginally correct standard errors derived in chapter 2, we aim for marginally correct predictions, and do not condition on the realization of the covariates in the sample.

The chapter will be composed of the following sections: we will first detail a theoretical method for marginally correct prediction and prove that it makes good on its coverage guarantees in section 3. In section 4 we present an algorithm for implementing the several methods, and compare their performance with classical regression's in section 5. Section 6 concerns valid coverage for the mean response. Section 7 concludes.

4.3 Marginally correct intervals

We first describe the problem in English, and then symbolically. A sample of predictor-response vectors is observed, and a regression line is fit to the data. We wish to design a procedure by means of which an interval is created around the regression line that guarantees marginal $1-\alpha$ coverage of future responses.

Minimal assumptions on the data generating mechanism are imposed, and assumptions such as linearity, homoscedasticity, etc. are absent. As in (Buja et al., 2013) and chapter 2, allow $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p, Y)$ to be jointly distributed according to joint distribution F, and observe an iid sample $(X_{ij}, Y_i)\, j = 1, \ldots p, i = 1 \ldots n$. Denote the observed sample by $\boldsymbol{S}$. Define the conditional mean $\mu(\boldsymbol{X}) \equiv \mathbb{E}\left[Y|\boldsymbol{X} = \vec{\boldsymbol{x}}\right]$ and assume it exists, as wells as the second moment of Y.

Note that we no longer assume that $\mu(\boldsymbol{X})$ is linear in $\boldsymbol{X}$. Rather, the working

assumption is that the conditional mean is probably non-linear. Since OLS can in this case no longer claim to estimate the population conditional expectation function, we define an alternate target of estimation, similarly to (White, 1980b). Irrespective of the functional form of $\mu(\boldsymbol{X})$, there is a unique hyperplane whose mean square error with respect to $\mu(\boldsymbol{X})$ is minimized. The slope of that hyperplane is the target, and it is computed through a (population) least squares regression. Therefore the population least squares slope, $\boldsymbol{\beta}$, is defined as

$$\boldsymbol{\beta} = \mathbb{E}\left[\boldsymbol{X}\boldsymbol{X}^T\right]^{-1}\mathbb{E}\left[\boldsymbol{X}Y\right] \tag{4.2}$$

which defines the slope of the hyperplane minimizing *expected* squared error loss with respect to the true conditional expectation, and averaged over the joint distribution F. This $\boldsymbol{\beta}$, a population parameter defined through minimization, and the target of estimation, can be estimated, in the usual fashion, via least squares:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^TY$$

(Buja et al., 2013) and chapter 2 derive proper standard error estimates for the $\hat{\boldsymbol{\beta}}$ when they are considered as estimates of $\boldsymbol{\beta}$ defined in (4.2); the source of their randomness includes both the variability of the errors, and the stochastic mechanism behind generating $(\boldsymbol{X}, y)$ pairs during sampling. Unlike classical methods, these standard errors offer valid coverage for $\boldsymbol{\beta}$. We also take into account the randomness deriving from the joint distribution of the $(\boldsymbol{X}, y)$.

4.3.1 Precise Statement

In this section we write down an interval granting valid coverage. Let $\lambda(\boldsymbol{X})$ be a prespecified function of the observations – $\lambda(\boldsymbol{X})$ will define the shape of the interval.

Let $\hat{K}$ be a multiplier estimated from the data, and which will define the interval's width. And let $\hat{\boldsymbol{\beta}}$ be the vector of partial slopes for a hyperplane fit to the data. The statement and proof will hold a vector of slopes computed anyhow, but, in the most familiar case, this vector is estimated through OLS. We will therefore refer to $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ as the "regression line" out of convenience, and may implicitly assume $\boldsymbol{\beta}$ to be $\boldsymbol{\beta}_{OLS}$. Represent the interval by

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} \pm \hat{K}\lambda\left(\boldsymbol{X}\right) \equiv \hat{CI}\left(\boldsymbol{X}, \hat{\boldsymbol{\beta}}, \hat{K}\right)^{1}. \tag{4.3}$$

Such an interval is now sought after that, centered at the regression line, it will have honest marginal coverage: given a level α, it will be necessary for

$$P^{\boldsymbol{S}} {P_F}^{Y,\boldsymbol{X}} \left(Y \in \hat{CI}\left(\boldsymbol{X}, \hat{\boldsymbol{\beta}}, \hat{K}\right)\right) \geq 1 - \alpha \tag{4.4}$$

Note the distributions over which the probabilities are computed. We wish to offer a procedural guarantee: over all realizations of samples, and over the joint distribution of future observations, given a vector of partial slopes, and an appropriately calibrated width of interval, the probability that a future observation will fall within the calibrated interval will be at least $1 - \alpha$. Equivalently, and giving more insight, we seek an interval to guarantee that:

$$\mathbb{E}^{\boldsymbol{S}}\left[{P_F}^{Y,\boldsymbol{X}} \left(Y \in \hat{CI}\left(X, \hat{\boldsymbol{\beta}}, \hat{K}\right) | \boldsymbol{S}\right)\right] \geq 1 - \alpha \tag{4.5}$$

We are able to fulfill this guarantee asymptotically.

Let us now describe why such an interval is attainable, and concretely how to attain it.

[1] In the classical interval, $\hat{K}$ would be equal $t_{\alpha/2,n-p-1}\hat{\sigma}$, while $\lambda(\boldsymbol{X})$ would equal $\sqrt{1 + \vec{x}_i \left(X^T X\right)^{-1} \vec{x}_i^T}$

4.3.2 Parallel Bands

To illustrate the general point, take $\lambda(\boldsymbol{X})$ from (4.3) above to equal 1. Such a choice generates a one-parameter family of band widths $\hat{K}$, with the bands parallel to the regression surface. This is not the "funnel shape" familiar from classical prediction intervals, which widens as points lie farther from the mean of the predictors, and will not give and does not promise conditional coverage.

We claim and will prove that, given a sample, and a hyperplane passing through it, a procedure that with parallel lines marginally captures $1-\alpha$ of the sample data points, will, asymptotically, offer $(1-\alpha)\,\%$ coverage.

For the statement of the main proof, recall that, with $\lambda(\boldsymbol{X}) = 1$,

$$\hat{CI}\left(\boldsymbol{X}, \hat{\boldsymbol{\beta}}, \hat{K}\right) = \boldsymbol{X}\hat{\boldsymbol{\beta}} \pm \hat{K}$$

In the most general formulation, the $\hat{\boldsymbol{\beta}}$ should be thought of simply as the statistic that converges to its limit in the population. Adopting the philosophy from section 2, $\hat{\boldsymbol{\beta}}_{OLS}$ converges to $\boldsymbol{\beta}$ defined in (4.2). But the $\hat{\boldsymbol{\beta}}$ needn't necessarily be the least squares slope. At its most uninspired, it can be taken to be a constant (and can therefore not depend on the sample) – coverage will be guaranteed even in this case. But centering at the regression line will be more efficient, and we therefore proceed assuming that the slopes in the sample and population are derived from sample and population least squares regressions, respectively.

Now, define the population analog to the sample coverage interval and call it $CI(\boldsymbol{X}, \boldsymbol{\beta}, K) = \boldsymbol{X}\boldsymbol{\beta} \pm K$, which is just a hypothetical band around the line defined by the population parameter $\boldsymbol{\beta}$ (the $\boldsymbol{X}$ parameter is therefore redundant). Define also the "coverage function" to be

$$Q_F(\boldsymbol{\beta}, K) = {P_F}^{Y,\boldsymbol{X}}\left(Y \in CI\left(X; \boldsymbol{\beta}, K\right)\right) \tag{4.6}$$

$Q_F(\boldsymbol{X};\boldsymbol{\beta},K)$ measures, given the slope applied to the $\boldsymbol{X}$ in the population and interval width K, how much Y-mass is contained in the strip. And let $K_0(F,\boldsymbol{\beta};\alpha)$ be that interval width derived from

$$\inf_K\{Q_F(\boldsymbol{X};\boldsymbol{\beta},K)\geq 1-\alpha|\boldsymbol{\beta}\} \tag{4.7}$$

K_0 is an oracular constant – if the joint distribution and slope were known, then an interval of width at least K_0 would contain at least the prespecified $1-\alpha$ of Y-mass.

The main theorem is now presented.

Denote by F^n the sampled data's empirical distribution. Suppose K_0 is unique, and F^n comes from a family $\mathcal{F}$ of distributions with bounded second moments. Then

Theorem 4.3.1 $\hat{K}_0\left(F^n,\hat{\boldsymbol{\beta}},\alpha\right)\rightarrow K_0(F,\boldsymbol{\beta};\alpha)$ *in probability.*

First the consequence: a procedure that captures $(1-\alpha)\%$ of the data in the sample, asymptotically captures $(1-\alpha)\%$ of the data in the population.Proof: Metrize the space of distributions with the weak* topology, with metric $\|\cdot\|_*$. We will need the following fact:

Lemma 4.3.2 $K_0(F,\boldsymbol{\beta};\alpha)$ *is continuous over* $\mathcal{F}\times\mathbb{R}^p$

The proof is found in the appendix.

It is standard that $\| F^n-F\|_*\rightarrow 0$ and that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\rightarrow\mathbf{0}$ in $\mathbb{R}^p$. [2] Because, by Lemma 4.3.2, K_0 is a continuous function of the arguments, it follows that $\hat{K}_0$ converges to K_0.

[2]Alternatively, when the slopes are not derived through regression, include as an assumption in the statement of the theorem that $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|\rightarrow\mathbf{0}$ in $\mathbb{R}^p$

4.3.3 Nonparallel bands

Up until now we have considered only bands parallel to the regression line. While offering promised coverage, the shape of the interval ignores leverages, and in general, strays from conditional coverage guarantees. The funnel shape of prediction bands preserves those guarantees under classical assumptions, and we consider such a shape as well. Instead of $\lambda(\boldsymbol{X}) = 1$ as in the previous section, now allow $\lambda(\boldsymbol{X}) = \sqrt{1 + \vec{\boldsymbol{x}}_i \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \vec{\boldsymbol{x}}_i^T}$, as in standard regression.

We emphasize that the substantive results do not change when the shape of the interval does: the procedure still calibrates the bands in-sample to capture 95% of the observed data, and the resultant band intervals will still converge to population intervals capturing 95% of mass. But, when regression assumptions are in fact met, conditional coverage will be improved at locations away from the mean relative to the parallel band method.

Corollary 4.3.3 *When $\hat{\lambda}(\boldsymbol{X})$ converges pointwise to $\lambda(\boldsymbol{X})$, the conclusion of Theorem* (4.3.1) *holds.*

Proof. With the pointwise convergence, the limiting strips in equation (4.14) in the appendix converge, and the rest of the proof follows without change.

4.4 Procedures

We first describe the most natural in-sample calibration technique. We then propose a resampled calibration technique that also achieves promised coverage.and detail it below[3]:

[3]Working name, calibrated resampled prediction (CARP)

In-sample procedure:

Parallel Bands

(a) Fit regression: given the $(\boldsymbol{X}, y)$ sample, find, via least squares, $\hat{\boldsymbol{\beta}}$ and $\hat{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

(b) Calibrate: find such a half-width $\hat{K}(\alpha)$ that $100 * (1-\alpha)\%$ of the responses are bracketed by $\hat{y} \pm \hat{K}(\alpha)$. That is, $(1-\alpha)$ of the responses should lie within the calibrated prediction bands. Procedurally, the width equals $|r_i|_{(1-\alpha)}$, the $1-\alpha$ quantile of all the absolute residuals.

(c) Report $\boldsymbol{X}\boldsymbol{\beta} \pm \hat{K}$ as the prediction interval.

CARP procedure:

(a) Begin with the original data: n data rows, each of them composed of a p-vector of observations and a response $(\vec{\boldsymbol{x}}_i, y_i)_{i=1}^n$, where $\vec{\boldsymbol{x}}_i = (\boldsymbol{X}_{i1}, \ldots \boldsymbol{X}_{ip})$.

(b) Fit regression: find, via least squares, $\hat{\boldsymbol{\beta}}$ and $\hat{y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$.

(c) *Nboot* times, sample n observations with replacement.

(d) Calibrate: In simulation s, find such a half-width $\hat{K}^{(s)}(\alpha)$ that $100 * (1-\alpha)\%$ of the test responses are bracketed by $\hat{y}_{train} \pm \hat{K}^{(s)}(\alpha)$. That is, $(1-\alpha)$ of the test set responses should lie within the calibrated prediction bands. Procedurally, the width equals $|r_i|_{(1-\alpha)}$, the $1-\alpha$ quantile of all the absolute residuals.

(e) Record $Med\left[\hat{K}^{(s)}\right]$, the median of the half-widths from the resampled datasets, as the (constant) half-width.

(f) Publish the regression line obtained from least squares regression applied to the original data, together with the prediction interval whose width was computed in step (e).

Several variants to this procedure exist, which will now be described. Specifically, we consider non-parallel prediction bands.

Resampled nonparallel bands

To construct non-parallel bands, proceed exactly as in the preceding section until and including step (c). To generate the "hourglass" shape of classical prediction bands, normalize the raw residuals by the classical prediction band width: Define $\tilde{r_i}^{(s)} = \frac{r_i^{(s)}}{k(\boldsymbol{x}^{(s)})}$, where $k(\boldsymbol{x})$ is the classical prediction interval

$$\hat{\sigma}\sqrt{1 + \vec{\boldsymbol{x}}_i(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i^T}.$$

Subsequently,

(a) find such a multiplier $\hat{C}(\alpha)$ that

$$\hat{y}_{train} \pm C^{(s)}k(x^{(s)})$$

captures $(1-\alpha)$ of the $nsim \times N$ test responses. That is, $(1-\alpha)$ of the test set responses should lie within the calibrated prediction bands. Procedurally, the width equals $|r_i|_{(1-\alpha)}$, the $1-\alpha$ quantile of all the absolute normalized residuals.

(b) Publish the regression line obtained from least squares regression applied to the original data, together with the prediction interval whose width was computed in (a).

4.5 Performance comparison

In this section we compare the coverage performance of regression prediction intervals, parallel bands calibrated in-sample, and parallel bands derived from resampled

calibration. For sample size equal to 100, 500, 1000, and 10,000, for several model specifications, we report the proportion y_{new} covered by nominally 95% classical intervals and the parallel bands. For each sample size and model specification, we ran the procedure $100,000$ times and recorded the average coverage proportion.

As expected, the parallel bands' coverage approaches the promised 95% as sample size increases. When the model is correctly specified, then in-sample calibration under-performs for small sample sizes, but does converge. In the face of mis-specification, we present two typical examples. A powerful lesson is that even in the face of severe model mis-specification, regression-based prediction intervals at worst have coverage not too far below the promised confidence level (model (b)). Sometimes (model (c)) they over deliver and give confidence limits that are overly conservative. Such wide limits are not practically useful. Coverage guarantees can therefore, asymptotically, either not be met, or else the intervals can be overly wide and therefore less informative. [4]

(a) Correctly specified linear model: $Z_1, Z_2 \overset{iid}{\sim} N(0,1), X_1 \sim N(0,1), X_2 \sim Unif[0,1]$, $X_3 \sim Expo(2), X_4 = X_1 + X_2 + Z_1, X_5 \sim Gamma(2,3)$ and $Y = X_1 + 2X_2 - 3X_3 + 0.1X_4 + X_5 + Z_2$. The $X_i, i = 1, \ldots, 5$ are measured and available, and a regression is fit to them.

(b) Transformed predictors: $Z \sim N(0,1), X_1 \sim N(0,1), X_2 \sim Unif[0,1], X_3 \sim Expo(2), X_4 \sim Norm(0,1), X_5 \sim Gamma(2,3)$ and $Y = X_1^2 + \log X_2 - 3X_3 + 0.1X_1 * X_2 + exp(X_5) + Z$. The regression fits Y to the untransformed $X_1, \ldots, X_5$.

(c) With heteroscedasticity: same as (b), except $Z \sim N(0, 5X_3)$.

[4] These model specifications were chosen to illustrate the general point, but were not the extremes of models considered. We have worked with models, for example, whose classical prediction intervals would cover 93% of the time.

Table 4.1: n = 100

	Classical Coverage	Parallel Band Coverage	CARP coverage
Model 1	0.9505	0.9269	0.9266
Model 2	0.9389	0.9314	0.9311
Model 3	0.977	0.9330	0.9326

Table 4.2: n = 500

	Classical Coverage	Parallel Band Coverage	CARP coverage
Model 1	0.9499	0.9450	0.9449
Model 2	0.9420	0.9463	0.9462
Model 3	0.9959	0.9470	0.9469

Table 4.3: n = 1000

	Classical Coverage	Parallel Band Coverage	CARP coverage
Model 1	0.9503	0.9485	0.9484
Model 2	0.9432	0.9482	0.9481
Model 3	0.9934	0.9484	0.9482

Table 4.4: n = 10000

	Classical Coverage	Parallel Band Coverage	CARP coverage
Model 1	0.9501	0.9493	0.9493
Model 2	0.9439	0.9499	0.9502
Model 3	0.9992	0.9501	0.9498

4.6 Calibrated Intervals for the Mean Response

According to classical regression guarantees, the conditional mean response of the regression can be covered with specified confidence level $(1-\alpha)$. The corresponding statistical procedure creates an interval with half-width $t_{1-\alpha/2,np-1}\hat{\sigma}\sqrt{\vec{\boldsymbol{x}}_i\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\vec{\boldsymbol{x}}_i^T}$, and centers it at the estimated conditional mean response, $\hat{y}$. Here $\hat{\sigma}$ is the RMSE –

computed through $\sqrt{\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n-p-1}}$, and $\vec{\boldsymbol{x}}_i$ is the $(p+1)$ vector at which one wishes to estimate the mean.

(Buja et al., 2013) and chapter 2 elucidate how the violation of the linearity assumption, in concert with stochastically generated predictors, results in overly conservative, or else invalid standard errors for the regression coefficients: they can either over- or under-cover with respect to the nominal confidence level.

The estimated regression line, at any point a linear combination of the coefficients $\hat{\boldsymbol{\beta}}$, should experience a similar departure from nominal coverage. When the conditional response is linear in the predictors, it is the conditional response proper that is being estimated with the conventional statistical procedure. When the conditional response is not linear in the predictors either because of a different functional form, or because some variables are omitted, then the natural target of estimation is the population least squares approximation to the conditional response surface,where the population coefficients are derived through population least squares.

In the following section we show just how far coverage can depart from the nominal level, and we detail an improving remedy.

Example

A basic example illustrates the principle well. Consider the simplest non-linear scenario: $y = x^2$. There is no noise. And consider three distributions for the x:

(a) $x \sim N(0, 1)$

(b) $x \sim |N(0, 1)|$ and

(c) $x \sim e^{N(0,1)}$.

For each scenario, the population least squares slopes $\boldsymbol{\beta}$ were computed, then 100,000 samples of size $n = 50$ were drawn, the conventional confidence interval for the mean computed, and the proportion of points at which the least squares response surface contained – tallied. In none of the scenarios was the nominal coverage guarantee honored, with the third violation particularly egregious:

(a) $x \sim N(0,1)$. Coverage: 79%

(b) $x \sim |N(0,1)|$. Coverage: 75%

(c) $x \sim logN(0,1.)$ Coverage: 20%

When the sample size was quadrupled to $n = 200$, coverage increased slightly to $81\%, 78\%, and 22\%$, respectively.

4.6.1 Calibration

Earlier in this chapter, we calibrated prediction intervals in-sample, captured $(1-\alpha)*$ 100% of the points with our interval, and proved that, asymptotically, the procedure offers valid coverage for future individual observations.

In a similar spirit, we seek a procedure to compute that multiple $\hat{K}$ such that

$$\hat{y} \pm \hat{K} * t_{1-\alpha/2,np-1}\hat{\sigma}\sqrt{\vec{\boldsymbol{x}}_i \left(X^T X\right)^{-1} \vec{\boldsymbol{x}}_i^T}$$

has $(1-\alpha)$ marginal probability of capturing the best linearly approximating surface.

The pairs bootstrap, elucidated in chapter 2, offers a means to approximate K. A comparison of the resultant bootstrap multiples to the "true" multiples found through simulation shows that the population K can be found through bootstrap, but in the severe case of a skewed $\boldsymbol{X}$ distribution, one with high leverage points with low probabilities, a bootstrap will not capture the necessary interval. This makes

sense: the rarely seen leverage point appreciably impacts the slope of the population least squares line; but such points are rarely found in sample, and rarer still in the bootstrap sub-sample.

Procedurally, since the population $\boldsymbol{\beta}$ is known, we repeatedly sampled $n = 50$ observations, computed the regression line and confidence interval for the mean, and searched over a grid of multiples until the dilated confidence intervals covered $\boldsymbol{X\beta}$ with desired probability. The complementary procedure was performed in the sample with 10,000 bootstrap replicates, and the results are here presented.

In scenario (a), the bootstrap multiple was 1.77 ($\pm.1$), while the population multiple was 1.7. Using $K = 1.77$ in the population, 96% coverage is attained.

In scenario (b), the bootstrap multiple was 1.9 ($\pm.4$), and the population multiple was 2.3. Using $K = 1.9$ in the population, 91% coverage is attained.

In scenario (c), the bootstrap multiple was 6.2 (± 3), while the population multiple was 20.1. Using $K = 6.2$ in the population, 64% coverage is attained.

Again, points with high leverage (at the population level) that have a low probability of appearing, and which appear in regions of high nonlinearity, inflate both the interval, and the ratio of the population K to the bootstrap derived $\hat{K}$. There must be a connection to the RAV, which will be investigated in subsequent research.

Maybe we should all become robustniks.

4.7 Conclusion

We have delivered on our promise to provide marginally asymptotically valid prediction intervals that are robust to model misspecification. Nowhere in this chapter is anything assumed about the covariates or response other than some moment and regularity conditions. Simple in-sample calibration suffices, and nothing is gained

from resampling techniques for prediction intervals. Classical statistical formulas are robust against mild departures from assumptions. Grosser violations ought to be caught by the researcher but may not be. Our technique protects the researcher from an unnecessary application of regression techniques and guarantees asymptotically marginal coverage for all reasonable data-generating mechanisms. For coverage of the mean, resampling techniques offer hope of salvaging nominal coverage.

Subsequent work will apply to model selection. While intervals that offer coverage for all conceivably selected models will be untenably large, we believe that work for submodels of a fixed size is manageable.

Appendix

Recall that, given a desired coverage level $1-\alpha$, K_0 is that oracular constant so that $P_F^{Y,X}(Y \in X'\beta \pm K_0|\alpha) = 1-\alpha$. The goal is to prove that $K_0(F,\beta;\alpha)$ is continuous over $\mathfrak{F} \times \mathfrak{R}$. The task cannot be completed directly, so we introduce an auxiliary evaluation function, V, which evaluates a coverage level, and one of whose arguments will be the confidence band width K. Let

$$V(F,\beta,k) = \int_{X\beta \pm k} dF \tag{4.8}$$

and denote the mass of the distribution function over the strip defined by $X\beta \pm k$. Through V, K_α can be equivalently defined as the smallest constant which, for a given α, sets

$$V(F,\beta,K) = 1-\alpha \tag{4.9}$$

To wit, $K_\alpha = \inf_{F,\beta}\{K|V(F,\beta,k) = \alpha\}$. In this way, K_α is defined implicitly as a function of F, β, and α. Notationally, write $K_\alpha(F_n, \beta_n, \alpha)$ as $K_{\alpha,n}$, and we have that $V(F,\beta,K_{\alpha,n}) = 1-\alpha$.

I claim that V is uniformly continuous.

Lemma 4.7.1 *: If K_α is unique then V is uniformly continuous.*

Proof:

Denote by $\mathfrak{X}$ the support of F. Let $X \in \mathfrak{X}$ and take $\epsilon > 0$. Choose δ appropriately. Let $\left|(F_n, \beta_n, \hat{K}) - (F, \beta, K)\right| < \delta$ Then

$$
\begin{aligned}
|V(F_n, \beta_n, K) - V(F, \beta, K)| &= \left| \int_{X\beta_n \pm k_n} dF_n - \int_{X\beta \pm k} dF \right| && (4.10)\\
&\leq \left| \int_{X\beta_n \pm k_n} dF_n - \int_{X\beta_n \pm k_n} dF \right| && (4.11)\\
&+ \left| \int_{X\beta_n \pm k_n} dF - \int_{X\beta \pm k} dF \right| && (4.12)\\
&= \left| \int_{X\beta_n \pm k_n} d(F_n - F) \right| + \left| \int_{X\beta_n \pm k_n - X\beta \pm k} dF \right| && (4.13)
\end{aligned}
$$

We will now show that the two terms in (4.13) tend to 0. [For notational convenience, let S_n represent the strip $X\beta_n \pm k_n$, and S the limiting strip $X\beta \pm k$]. By assumption, $F_n \to F$ weakly, so $F_n(X) \to F(X)$ for all X at which $F(X)$ is continuous. This is equivalent to the formulation that $\forall c, \int c(x)dF_n \to \int c(x)dF$, and, in particular, for $c = \mathbb{1}_{S_n(x)}$. The first term therefore tends to zero. and the integral is evaluated over a compact set.

Second term:

The integral in question can be represented with characteristic functions:

$$\int_\Omega [\mathbb{1}_{S_n(x)} - \mathbb{1}_{S(x)}] dF(x) \tag{4.14}$$

Now $\mathbb{1}_{S_n(x)}$ converges pointwise to $\mathbb{1}_{S(x)}$, and is dominated by, say, 2, so the second term converges to 0 by dominated convergence, and V is uniformly continuous, as desired.

Lemma: Assume that dF_n is bounded away from zero. In the implicit function defined by $V(F, \beta, K_{\alpha,n}) = 1 - \alpha$, K is a strictly decreasing function of α.

Proof: The conclusion follows directly from the statement. Since dF_n is bounded away from 0, $V(F_n, \beta_n, K_{\alpha,n})$ is a strictly increasing function of S_n, the region over which it is integrated. Because β_n is fixed, that region, in turn, is an increasing function of K. As α increases, the RHS decreases, the LHS must decrease for equality to hold, and K must decrease.

Claim: K_α is a continuous function.

Proof.

Consider $V^{-1}(\alpha|F, \beta)$. Since V is uniformly continuous and monotonically increasing in K, the inverse is continous as well. But $V^{-1}(\alpha|F, \beta)$ is equal to $K_\alpha(F, \beta)$, and we are done ■

Bibliography

RA Berk, Brian Kriegler, and Donald Ylvisaker. Counting the homeless in los angeles county. *Probability and Statistics: Essays in Honor of David A. Freedman. Inst. Math. Statist. Collect*, 2:127–141, 2008.

Andreas Buja, Richard A Berk, Lawrence D Brown, Edward I George, E Pitkin, M Traskin, L Zhao, and K Zhang. A conspiracy of random X and model violation against classical inference in linear regression. University of Pennsylvania working paper, Nov 2013.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

David Freedman, Robert Pisani, and Roger Purves. Statistics (3rd edn), 1998.

David A Freedman. Bootstrapping regression models. *The Annals of Statistics*, pages 1218–1228, 1981.

David A Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.

David A Freedman and Richard A Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, 2008.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5(1):81–102, 1978.

Kosuke Imai, Gary King, and Elizabeth A Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171(2):481–502, 2008.

G Imbens and D Rubin. Causal inference: Statistical methods for estimating causal effects in biomedical, social, and behavioral sciences, 2007.

Guido M Imbens. Experimental design for unit and cluster randomized trials. 2011.

Guido M Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. Technical report, National Bureau of Economic Research, 2008.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Robert J LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, pages 604–620, 1986.

Winston Lin. Agnostic notes on regression adjustments to experimental data: reexamining freedmans critique. *The Annals of Applied Statistics*, 7(1):295–318, 2013.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985.

Enno Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *The Annals of Statistics*, pages 255–285, 1993.

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

Dimitris N Politis. Model-free model-fitting and predictive distributions. *Test*, 22(2): 183–221, 2013.

Michael Rosenblum and Mark J van der Laan. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to leverage baseline variables. 2010.

Donald B Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

Cyrus Samii and Peter M Aronow. On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics & Probability Letters*, 82(2):365–370, 2012.

Richard L Schmoyer. Asymptotically valid prediction intervals for linear models. *Technometrics*, 34(4):399–408, 1992.

Jerzy Splawa-Neyman, DM Dabrowska, and TP Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990.

Robert A Stine. Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, 80(392):1026–1031, 1985.

Anastasios A Tsiatis, Marie Davidian, Min Zhang, and Xiaomin Lu. Covariate adjustment for two-sample treatment comparisons in randomized clinical trials: A principled yet flexible approach. *Statistics in Medicine*, 27(23):4658–4677, 2008.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980a.

Halbert White. Using least squares to approximate unknown regression functions. *International Economic Review*, pages 149–170, 1980b.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Li Yang and Anastasios A Tsiatis. Efficiency study of estimators for a treatment effect in a pretest–posttest trial. *The American Statistician*, 55(4):314–321, 2001.

Min Zhang, Anastasios A Tsiatis, and Marie Davidian. Improving efficiency of inferences in randomized clinical trials using auxiliary covariates. *Biometrics*, 64(3): 707–715, 2008.