Anytime-Valid Linear Models and Regression Adjusted Causal Inference in Randomized Experiments

Michael Lindon[*], Dae Woong Ham[†], Martin Tingley[*], and Iavor Bojinov[◇]

July 3, 2023

Abstract

Linear models are commonly used in causal inference for the analysis of experimental data. This is motivated by the ability to adjust for confounding variables and to obtain treatment effect estimators of increased precision through variance reduction. There is, however, a replicability crisis in applied research through unknown reporting of the data collection process. In modern A/B tests, there is a demand to perform regression-adjusted inference on experimental data in real-time. Linear models are a viable solution because they can be computed online over streams of data. Together, these motivate modernizing linear model theory by providing "Anytime-Valid" inference. These replace classical fixed-n Type I error and coverage guarantees with time-uniform guarantees, safeguarding applied researchers from p-hacking, allowing experiments to be continuously monitored and stopped using data-dependent rules. Our contributions leverage group invariance principles and modern martingale techniques. We provide sequential t-tests and confidence sequences for regression coefficients of a linear model, in addition to sequential F-tests and confidence sequences for collections of regression coefficients. With an emphasis on experimental data, we are able to relax the linear model assumption in randomized designs. In particular, we provide completely nonparametric confidence sequences for the average treatment effect in randomized experiments, without assuming linearity or Gaussianity. A particular feature of our contributions is their simplicity. Our test statistics and confidence sequences have closed-form expressions of the original classical statistics, meaning they are no harder to use in practice. This means that published results can be revisited and reevaluated, and software libraries which implement linear regression can be easily wrapped.

Keywords: Anytime-Valid Inference, Sequential Testing, Bayes Factors, Confidence Sequences, Sequential p-values, e-processes, Group Invariance, Martingales, A/B Testing

[*] Netflix, 121 Albright Way, Los Gatos, CA 90302. michael.s.lindon@gmail.com, mtingley@netflix.com

[†] Department of Statistics, Harvard, MA. daewoongham@g.harvard.edu

[◇] Harvard School of Business, Harvard, MA. ibojinov@hbs.edu

Contents

1 Introduction

1.1 Linear Models and the Analysis of Experimental Data

Linear models are fundamental to many applied sciences, but perhaps no field employs linear models quite as routinely as causal inference. There is a long history of using linear models for the analysis of experimental data, almost as old as the field of causal inference itself, dating back to the early works on structural equation modelling (Wright, 1921), regression adjustment (Neyman, 1923), and analysis of variance (Fisher, 1925). In the second half of the 20th century, linear models were further employed in the analysis of observational data, through instrumental variable modelling (Wright, 1928; Angrist and Krueger, 1991), regression discontinuity designs (Thistlethwaite and Campbell, 1960) and difference-in-differences (Card and Krueger, 1994).

Linear models are routinely recommended for the analysis of experimental data for two main reasons. The first is to achieve more precise estimates of average treatment effects (ATEs) by regressing on pre-treatment covariates that are correlated with the response. Moreover, in randomized designs, the covariate-adjusted difference in means estimator is robust to the linear model misspecification, providing asymptotically correct inference for ATEs without requiring linearity assumptions Lin (2012); Imbens and Rubin (2015). The second argument is to adjust for imbalances in pre-treatment covariates in quasi- and observational experiments. Consequently linear models are taught to applied researchers (Cook et al., 2002; Huitema, 2011; Howell, 2012) and regression adjustments are widely employed in the social sciences. For example, the institute of education sciences' handbook of procedures and standards reads ”Whenever possible, a reviewer should therefore use covariate-adjusted mean differences to compute effect sizes. These adjusted mean differences can come from various types of models such as multiple regression, analysis of covariance...” (Clearinghouse, 2022). In medicine, the Cochrane handbook for systematic reviews of interventions reads “The preferred statistical approach to accounting for baseline measurements of the outcome variable is to include the baseline outcome measurements as a covariate in a regression model or analysis of covariance (ANCOVA)” (Higgins et al., 2019, Chapter 10.5.2). In large-scale online experiments in technology companies, linear models remain popular due to their scalability and have become an industry standard for providing regression-adjusted estimates of ATEs (Deng et al., 2013).

Our contributions have two motivations. The first motivation is that advances in technology have evolved how modern experiments are designed and analyzed. Consequently, there is a need to evolve classical linear model theory to better suit modern-day practices. The second motivation targets the reproducibility crisis of significant results in applied research. We propose a safe testing analogue that can significantly improve the reliability of published results.

1.2 Modernizing Linear Models with Anytime-Valid Inference

Classical hypothesis testing procedures were designed to obtain the highest statistical power possible at a fixed sample size (henceforth “fixed-n”). This is a consequence of how early experiments were performed in which the outcomes were only available well after the experiment had concluded. Technology has, however, advanced considerably from the agricultural experiments of the early 19’th century and has consequently changed how experiments are performed and analyzed.

The first difference is one of scale. Whereas a single classical experiment might have months of planning, and months of execution, it is not uncommon for large internet companies to have hundreds if not thousands of experiments running concurrently every week. It is not feasible to perform sample size calculations for every experiment, not only because of their number, but also because *accurate* sample size calculations are infeasible - sample size calculations depend on too many unknown model parameters in complex models (Noordzij et al., 2010).

The second difference is that observations typically arrive sequentially, and at extremely high rates. With such a high rate, it is possible for a negative treatment effect to perform a lot of harm very quickly. For this reason it is desirable to be able to detect large effects quickly, while still being able to detect small effects eventually. It is not possible to satisfy both objectives with a fixed-n test, which inevitably under- or overestimates the required sample size. If n is small, then the experiment catches large negative effects early but is often underpowered to detect small effects. If n is large, so that the experiment is powered to detect small negative effects, then there is a high risk of large negative effects being unaddressed for lengthy amounts of time. This is a critical issue for companies that use experiments not to test a hypotheisized positive effect of an intervention, but as quality control gates to mitigate risk when releasing new software (Lindon et al., 2022).

In order to scale modern experimentation, it is necessary to automate much of the experiment lifecycle, orchestrating and analyzing tests algorithmically, whether that be "turning off" badly performing treatments or productionizing winning treatments. This requirement has driven a growing body of literature of "Anytime-Valid" testing which allows experiments to be *continuously monitored*. Anytime-Valid inference extends classical inference by providing *time-uniform* Type-I error and coverage guarantees. In contrast to fixed-n inference, anytime valid inference is valid at all times instead of a single time. The replacement of p-values and confidence intervals with sequential p-values and confidence sequences allows the experiment to be algorithmically managed. For example, data-dependent stopping rules can be used to conclude the test, such as when the confidence sequence excludes zero. More generally, confidence sequences can be used to construct automated best-arm identification algorithms. It also provides flexibility, the researcher can simply stop the experiment when they are satisfied with the inference. In particular, the researcher can perform optional stopping or optional continuation without sacrificing Type-I error guarantees. These methods have been adopted at Microsoft (Waudby-Smith et al., 2022), Amazon (Services, 2023), Adobe (Maharaj et al., 2023), Optimizely (Johari et al., 2021, 2017) and Netflix (Lindon and Malek, 2020; Lindon et al., 2022; Ham et al., 2022; Bibaut et al., 2022).

Linear models remain popular for analyzing large internet experiments (Deng et al., 2013) due to their scalability, versatility but most importantly their ability to provide variance reduction when including pretreatment covariates that are correlated with the response. These are especially fruitful for companies that have a lot of user data prior to assigning treatments. Estimators at time t can be expressed recursively in terms of their values at time $t-1$, making linear models ideal for online computation. Despite their amenability to sequential analysis, there is little literature on how to combine anytime-valid inference with linear models. The purpose of this paper is to combine the benefits of anytime-valid inference and regression-adjustment in a highly scalable way by presenting how to perform anytime-valid inference in linear models.

1.3 Safe Linear Models and the Replication Crisis

The scope of this work is also not limited to large-scale experimentation at internet companies but also helps address the replicability crisis in the applied sciences. It was noted in the introduction that linear modelling, ancova, and regression adjustment are recommended by many official bodies of the applied sciences. Yet it has been widely noted that there is a serious problem with the current state of statistical methodology used in applied research, with many significant findings failing to be replicated (Wasserstein and Lazar, 2016; Benjamin et al., 2018). The open science collaboration presented an effort to replicate the results of 100 papers published in three top-tier psychology journals (Collaboration, 2015). The authors concluded that only 36% of the original statistically significant results could be replicated. In medicine, Ioannidis (2005) outlines many reasons why most published results are in fact false, which we recount below.

One reasons is that the narrow-sighted hunt for statistical power often results in a rigid all-or-nothing statistical inference procedure, from which any deviation can invalidate statistical guarantees. For example, classical fixed-n hypothesis tests are designed to achieve the highest power possible at a fixed sample size. Yet, schedules often shift and resources often change. Consider the very real constraints both on researchers' time and resources. When observations arrive sequentially, it can be incredibly tempting to naively *monitor* experiments, running repeated tests in an effort to finish an experiment early and save on resources, which naturally invalidate Type I error guarantees of classical procedures.

Suppose the results of a long and expensive experiment are "borderline significant". In that case, researchers may fall into the trap of thinking that the test was trivially underpowered and conclude that the appropriate course of action is to continue the experiment and collect more data (Sagarin et al., 2014). In an anonymous survey of 2000 psychologists, approximately 60% of respondents admitted to collecting more data after seeing whether the results were significant, with approximately 20% admitting to stopping data collection after achieving a favourable result (John et al., 2012). The authors estimate the true prevalence to be much higher. The temptation is clear. Given that journals bias toward publishing significant results, collecting more data could transform a submission rejection to an acceptance.

We should also recognize that researchers have both conscious and unconscious biases. Treatments are often conceived with a strong prior belief that it will be effective, and experiments are often funded with this promise. Yet, a researcher intent on disproving the null hypothesis can achieve this result by simply collecting data until this is so, so-called "sampling to a foregone conclusion" (Armitage et al., 1969; Armitage, 1993; Cornfield, 1966).

One does not need to look far to see examples of these problems in published works. (Carney et al., 2010) hypothesized that adopting "power-posing" could significantly increase testosterone, and reported significant increases with merely two-minutes of such poses. These results were unsuccessfully replicated by (Ranehill et al., 2015) and it is speculated that data collection was terminated as soon as a significant result was obtained, as the sample size of 36 is unusually low for this kind of experiment.

The widely observed pressure on researchers to publish can quite innocently compel even the best-intentioned among them to make such mistakes. Recently, "safe" testing procedures using e-variables and e-processes have been proposed as a solution by procedures Grünwald et al. (2021). Our contributions can also be viewed through this perspective, as our test statistics share e-process interpretations. When using anytime valid inference it is fundamentally not possible to falsify significant claims through data-dependent stopping rules in the data collection process. The probability of obtaining a significant result at the α-level is always less than α, regardless of how early or how late the researcher decides to collect data.

This effectively removes one degree of freedom of p-hacking. If results remain significant at the α-level using anytime-valid inference, then the reader can be confident that this is unlikely to be the result of misconduct regarding when to stop data collection. While Benjamin et al. (2018) propose to lower the threshold for significance down from 0.05 to 0.005, a journal reviewer could alternatively consider using the sequential p-values we provide to ensure the quality of published research. Moreover, our sequential p-values and confidence sequences are simple closed form expressions of classical sufficient statistics and estimators, making it easy to easy to revisit published works, perform anytime valid inference, and re-evaluate the statistical significance of their findings.

1.4 Contributions and Paper Outline

The contributions of this paper bring anytime-valid inference to linear models, extending classical theory by providing time-uniform Type-I error and coverage guarantees through modern martingale techniques. We provide (covariate adjusted) sequential t-tests, sequential F-tests, and confidence sequences for regression coefficients. While these results hold for general linear models, we give special attention to ANCOVA - linear models with pre-treatment (nuisance) covariates, main effect terms for treatment groups and optionally interaction terms. A key feature of our contributions is the simplicity of the results. The sequential p-values and confidence sequences we provide are closed-form expressions of ordinary least squares estimators of the coefficients and the residual variance, meaning it is no harder to perform anytime-valid inference than fixed-n inference. If a historical paper performed a fixed-n anaylysis and transparently reported these estimates, then it is easy to retroactively go back and perform the anytime-valid analysis. Consequently, any statistical software that performs a classical linear model analysis can easily be converted to provide an anytime-valid analysis.

The scope of our work is, however, not tied to the linear model assumption. It is well known that the linear regression adjusted average treatment effect estimator is robust to the linear model misspecification in large samples, providing asymptotically correct fixed-n inference for the ATE (Imbens and Rubin, 2015; Lin, 2012). By using the same arguments we are able to extend our results to asymptotic sequential tests and confidence sequences for average treatment effects *even when the linear model is misspecified*. This results in nonparametric, regression-adjusted asymptotic confidence sequences for average treatment effects in randomized experiments through trivial to implement ordinary least squares computations.

They are also safe under optional stopping and continuation. Our construction shares the interpretation as using an e-process. Whatismore, Bayesians will also recognize our test statistic as a Bayes factor. The methodology we present spans multiple testing paradigms and appeals, we hope, to many disparate schools of statistics.

2 Anytime Valid Inference - A Review

Sequential analysis must rely on martingales (Ramdas et al., 2022). For a sequence of iid $N(\delta, \sigma^2)$ random variables, one can test $H_0 : \delta = \delta_0$ vs $H_1 : \delta \neq \delta_0$ using the sequential probability ratio test (SPRT) (Wald, 1945, 1947). The SPRT test statistic is simply the likelihood ratio. This is a non-negative martingale under the null hypothesis, and the probability that it ever exceeds α^{-1} is less than α by Ville's inequality. Extensions to testing composite alternatives are possible by taking mixtures of martingales. Using a $N(0, \sigma^2\phi^{-1})$ mixture on the unknown δ in the alternative yields the following mixture martingale

$$B_n(\mathbf{Y}_n) = \int \frac{p(\mathbf{Y}_n|\delta, H_1)}{p(\mathbf{Y}_n|H_0)} p(\delta|H_1)d\delta = \sqrt{\frac{\phi}{\phi+n}}\, e^{\frac{1}{2}\frac{n}{n+\phi} z_n(\mathbf{Y}_n)^2}, \tag{1}$$

where $\mathbf{Y}_n = (y_1, y_2, \ldots, y_n)$ and $z_n(\mathbf{Y}_n) = \sqrt{n}(\bar{y}_n - \delta_0)/\sigma$ is the classical z-*statistic*. The mixture SPRT test statistic provides a clean way to convert a classical z-*score* to a nonnegative martingale, and hence easily convert a fixed-n to an anytime-valid analysis. For example, a sequential p-value can be obtained be defined by $p_n(\mathbf{Y}_n) = B_n(\mathbf{Y}_n)^{-1}$, which satisfies the definition of a sequential p-value

$$\mathbb{P}[\exists n \in \mathbb{N} : p_n(\mathbf{Y}_n) \leq \alpha] \leq \alpha. \tag{2}$$

The test statistic can also be inverted to construct a *confidence sequence* for δ by defining $C_n(\mathbf{Y}^n) = [\bar{y}_n - r_n(\mathbf{Y}_n), \bar{y}_n + r_n(\mathbf{Y}_n)]$, where

$$r_n(\mathbf{Y}_n) = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{\phi+n}{n} \log\left(\frac{\phi+n}{\phi\alpha^2}\right)} \tag{3}$$

which satisfies the definition of a confidence sequence

$$\mathbb{P}[\forall n \in \mathbb{N} : \delta \in C_n(\mathbf{Y}_n)] \geq 1 - \alpha. \tag{4}$$

This is often referred to as the Robbins confidence sequence (Robbins, 1952, 1970).

At first glance the scope of these results may seem limited due to strong parametric assumptions. In classical fixed-n tests, these can usually be justified by appealing to central limit theorem arguments, and anytime-valid tests are no exception. Waudby-Smith et al. (2021) provide time-uniform analogues of these arguments using strong approximations to sample sums. In particular, they use the parametric Robbins confidence sequence to construct nonparametric *asymptotic confidence sequences*. We recall their definition below.

Definition 2.1. (Asymptotic Confidence Sequence). $(\hat{\delta}_n \pm \tilde{C}_n)_{n=1}^{\infty}$ is two-sided $(1-\alpha)$ asymptotic confidence sequence for a parameter δ if there exists a (typically unknown) two-sided $(1-\alpha)$ nonasymptotic confidence sequence $(\hat{\delta}_n \pm C_n)_{n=1}^{\infty}$ for δ such that

$$\frac{C_n}{\tilde{C}_n} \to 1 \quad a.s. \tag{5}$$

An asymptotic confidence sequence is said to have a rate R_n if $\tilde{C}_n - C_n = O_{a.s.}(R_n)$. Although parametric, the Robbins confidence sequence layed the foundations for the development of nonparametric asymptotic confidence sequences, which have been successfully employed in (Ham et al., 2022; Waudby-Smith et al., 2022).

The contributions of this paper provide significant generalizations to the Robbins confidence sequence. We consider the generalized case of observing a sequence of iid observations

$$y_i|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2 \sim N(\boldsymbol{x}_i'\boldsymbol{\beta} + \boldsymbol{z}_i'\boldsymbol{\delta}, \sigma^2), \tag{6}$$

with $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\delta} \in \mathbb{R}^d$. This has three important generalizations. First, σ^2 is *unknown*. Second, $\boldsymbol{\delta}$ is *multivariate*. Third, there is a *linear regression component*. We are able to provide exact anytime-valid inference for $\boldsymbol{\delta}$, through sequential F-tests and confidence sequences, extending the Robbins confidence sequences to linear models.

The parameterization in equation (6) splits the regression coefficients into those of interest, and those that are nuisance, which includes a variety of interesting models. Motivated by experimental datasets, $\boldsymbol{\delta}$ could denote main effects terms, while pre-treatment covariates as and their centered interactions with treatments can be absorbed by $\boldsymbol{\beta}$, enabling anytime valid ANCOVA. This model be used to estimate the average treatment effect between treatment and control groups (Lin, 2012). Alternatively, $\boldsymbol{\delta}$ could denote the covariate-treatment interaction terms, absorbing pre-treatment covariates and main effects into $\boldsymbol{\beta}$, enabling tests for treatment effect heterogeneity.

Much like the Robbins confidence sequence, this may seem limited in scope at first glance, restricted to linear parametric assumptions. However, much like the Robbins confidence sequence layed the foundations for asymptotic confidence sequences, these confidence sequences lay the foundation for asymptotic regression-adjusted confidence sequences for the average treatment effect in randomized experiments, *even without assuming Gaussianity or linearity*. The main challenge is how to correctly deal with the nuisannce parameters $\boldsymbol{\beta}$ and σ, as we seek Type I error and coverage guarantees for all possible values. For this we appeal to invariance arguments, which will require a modicum of group theory and its applications to statistics (Eaton, 1989; Lehmann and Romano, 2005; Wijsman, 1990).

3 Anytime-Valid Theory for Classical Linear Models

3.1 Notation

We will express equation (6) in matrix notation as $\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{Z}_n\boldsymbol{\delta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{Y}_n = (y_1, y_2, \ldots, y_n)$, $\boldsymbol{X}_n$ is the matrix with rows $(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$, $\boldsymbol{Z}_n$ is the matrix with rows $(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n)$. Let $\boldsymbol{W} = [\boldsymbol{X}, \boldsymbol{Z}]$ and $\boldsymbol{\gamma}' = [\boldsymbol{\beta}', \boldsymbol{\delta}']$. Let $\mathcal{C}(\boldsymbol{A}) = \{\sum_i c_i\boldsymbol{A}_i : c_i \in \mathbb{R}, \boldsymbol{A}_i = \boldsymbol{A}\boldsymbol{e}_i\}$ denote the column space of a given matrix $\boldsymbol{A}$, $\mathcal{C}(\boldsymbol{A})^{\perp}$ the orthogonal complement of $\mathcal{C}(\boldsymbol{A})$, $P_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}'\boldsymbol{A})^{-1}\boldsymbol{A}'$ denote the orthogonal projection operator onto $\mathcal{C}(\boldsymbol{A})$, $r(\boldsymbol{A})$ the rank of $\boldsymbol{A}$ and $\|\boldsymbol{v}\|^2_{\boldsymbol{A}} = \boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}$. We have $r(\boldsymbol{X}) = p$, $r(\boldsymbol{Z}) = d$, $r(\boldsymbol{I}) = n$ and $r(\boldsymbol{W}) = p + d$. Let $s^2(\boldsymbol{Y}) = \boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{W}})\boldsymbol{Y}/(n - d - p)$ denote the usual unbiased estimator of σ^2. Let $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n = \boldsymbol{Z}_n'(\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n = \boldsymbol{Z}_n'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n$. Note that $(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}$ is simply the submatrix of $(\boldsymbol{W}_n\boldsymbol{W}_n)^{-1}$ corresponding to $\boldsymbol{\delta}$. Let $F(d_1, d_2, \mu)$ denote the non-central F-distribution with degrees of freedom d_1 and d_2 with non-centrality parameter μ. Let $\hat{\theta}_n(\boldsymbol{Y}_n)$ denote the maximum likelihood estimator of a parameter θ after n observations. Let $\boldsymbol{\Theta}$ denote the parameter space of $(\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2)$ and $\boldsymbol{\Theta}_0$ denote the subset restricted to $\boldsymbol{\delta} = \boldsymbol{\delta}_0$

3.2 Sequential Covariate Adjusted t-Tests

3.2.1 Nonasymptotic Results

Theorem 3.1. *For an iid sequence of observations $y_i \sim N(\boldsymbol{w}_i\boldsymbol{\gamma}, \sigma^2)$ where $\boldsymbol{w}_i = (\boldsymbol{x}_i, z_i)$ and $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \delta)$ with $\delta = \delta_0$ under the null hypothesis. For any fixed $\phi > 0$ let*

$$B_n(\boldsymbol{Y}_n; \delta_0) = \begin{cases} \sqrt{\frac{\phi}{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}} \frac{\left(1 + \frac{\phi t_n(\boldsymbol{Y}_n;\delta_0)^2}{(n-p-1)(\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2)}\right)^{-\frac{n-p}{2}}}{\left(1 + \frac{t_n(\boldsymbol{Y}_n;\delta_0)^2}{(n-p-1)}\right)^{-\frac{n-p}{2}}} & \boldsymbol{W}_n'\boldsymbol{W}_n \textit{full rank}, \\ 0 & \textit{otherwise}, \end{cases} \tag{7}$$

where $t_n(\boldsymbol{Y}_n; \delta_0) = (\hat{\delta}_n(\boldsymbol{Y}_n) - \delta_0)/se(\hat{\delta}(\boldsymbol{Y}_n))$ is the classical t statistic, $\hat{\delta}_n(\boldsymbol{Y}_n)$ is the OLS estimator of δ, $se(\hat{\delta}(\boldsymbol{Y}_n)) = \sqrt{s_n^2(\boldsymbol{Y}_n)/\|\tilde{\boldsymbol{Z}}_n\|_2^2}$ is the standard error, $\|\tilde{\boldsymbol{Z}}_n\|_2^2 = \boldsymbol{Z}_n'(\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n$ and $s_n^2(\boldsymbol{Y}_n) = \boldsymbol{Y}_n'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_n})\boldsymbol{Y}_n/(n-d-p)$ the usual unbiased estimator of σ^2. Then

$$\mathbb{P}_\theta[\exists n \in \mathbb{N} : B_n(\boldsymbol{Y}_n; \delta_0) \geq \alpha^{-1}] \leq \alpha, \tag{8}$$

for all $\theta \in \Theta_0$,

Theorem 3.1 states that the probability of $B_n(\boldsymbol{Y}_n; \delta_0)$ *ever* exceeding a threshold of α^{-1} when the null hypothesis is true is less than or equal to α, regardless of the values of the nuisance parameters $\boldsymbol{\beta}$ and σ^2. The conditional definition is simply to ensure that enough observations have been observed to enable the inversion of $\boldsymbol{W}_n'\boldsymbol{W}_n$.

Corollary 3.2. *$p_n(\boldsymbol{Y}_n; \delta_0) = 1/B_n(\boldsymbol{Y}_n; \delta_0)$ defines a sequential p-value satisfying*

$$\mathbb{P}_\theta[\exists n \in \mathbb{N} : p_n(\boldsymbol{Y}_n; \delta_0) \leq \alpha] \leq \alpha, \tag{9}$$

for all $\theta \in \boldsymbol{\Theta}_0$

Corollary 3.3. *Let $C_n(\boldsymbol{Y}_n) = \{\delta : B_n(\boldsymbol{Y}_n; \delta) \leq \alpha^{-1}\}$, then $C_n(\boldsymbol{Y}_n)$ defines a confidence sequence satisfying*

$$\mathbb{P}_\theta[\forall n \in \mathbb{N} : \delta \in C_n(\boldsymbol{Y}_n)] \geq 1 - \alpha, \tag{10}$$

for all $\theta \in \boldsymbol{\Theta}$.

3.2.2 Asymptotic Results

While the test-martingale in theorem 3.1 and the sequential p-value in corollary 3.2 are easy to evaluate closed form expressions, the interval $C_n(\boldsymbol{Y}_n)$ in corollary 3.3 requires an additional step. It requires a single root to be found numerically to provide the lower and upper bounds of the interval. Alternatively, we provide the following approximation which can either be used to seed a root-finding algorithm, or just used outright. To be precise about the nature of this approximation, we leverage the concept of asymptotic confidence sequences described in section 2. The following confidence sequence admits a closed form expression and is an asymptotic confidence sequence for the nonasymptotic confidence sequence in corollary 3.3 with an approximation rate is $o_{a.s.}(n^{-1/2})$. In order to guarantee the almost-sure rate we impose some small regularity conditions. Writing observations as $y_i = \boldsymbol{w}_i\gamma + \varepsilon_i$, where we have introduced the notation ε_i for the residuals, we assume that ε_i are zero mean and finite variance, $\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n/n \overset{a.s.}{\to} \boldsymbol{0}$, $\boldsymbol{W}_n'\boldsymbol{W}_n/n \overset{a.s.}{\to} \boldsymbol{A}$, $\boldsymbol{A}$ finite and positive definite.

Theorem 3.4. *Let*

$$\tilde{B}_n(\boldsymbol{Y}_n;\delta_0) = \sqrt{\frac{\phi}{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}} e^{\frac{1}{2}\frac{\|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi+\|\tilde{\boldsymbol{Z}}_n\|_2^2} t_n(\boldsymbol{Y}_n;\delta_0)^2} \tag{11}$$

then

$$\log \tilde{B}(\boldsymbol{Y}_n;\delta_0) = \log B(\boldsymbol{Y}_n;\delta_0) + o_{a.s.}(1) \tag{12}$$

Theorem 3.4 provides an approximate test statistic $\tilde{B}_n(\boldsymbol{Y}_n;\delta_0)$ that converges almost surely to our original test statistic $B_n(\boldsymbol{Y}_n;\delta_0)$, which can be used as an asymptotic sequential test in the following sense

$$\mathbb{P}_\theta[\exists n \in \mathbb{N} : \log \tilde{B}_n(\boldsymbol{Y}_n;\delta_0) \geq -\log\alpha + o_{a.s.}(1)] \leq \alpha, \tag{13}$$

for all $\theta \in \Theta_0$. Consistent with our earlier sequential p-value, an asymptotic sequential p-value can be obtained with $\tilde{p}_n(\boldsymbol{Y}_n) = 1/\tilde{B}_n(\boldsymbol{Y}_n;\delta_0)$. Our original test statistic $B_n(\boldsymbol{Y}_n;\delta_0)$, and consequently $p_n(\boldsymbol{Y}_n)$, already have closed-form expressions and so the reader may wonder why we are going to the trouble of defining these new approximations. The answer is that we can invert these new tests by hand, yielding an asymptotic confidence sequence that *does* have a closed form expression.

Corollary 3.5. *Let* $\tilde{C}_n(\boldsymbol{Y}_n) = \{\delta : \tilde{B}_n(\boldsymbol{Y}_n;\delta) \leq \alpha^{-1}\} = (\hat{\delta}_n(\boldsymbol{Y}_n) - r_n(\boldsymbol{Y}_n), \hat{\delta}_n(\boldsymbol{Y}_n) + r_n(\boldsymbol{Y}_n))$ *where*

$$r_n(\boldsymbol{Y}_n) = \frac{s_n(\boldsymbol{Y}_n)}{\|\tilde{\boldsymbol{Z}}_n\|}\sqrt{\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right)}. \tag{14}$$

$\tilde{C}_n(\boldsymbol{Y}_n)$ *is an asymptotic confidence sequence for* δ *with approximation rate* $o_{a.s.}(n^{-1/2})$.

This can be used to seed a root-finding algorithm to obtain the nonasymptotic confidence sequence, but for almost all intents and purposes it is fine to use as is, which we illustrate with the following example. As a sanity check, it's useful to observe that if we replace $s_n(\boldsymbol{Y}_n)$ with a known residual standard deviation σ, if there were no nuisance covariates $\boldsymbol{\beta}$, and if there were only one group (so that $\boldsymbol{Z}_n = \boldsymbol{1}_n$), then the asymptotic confidence sequence recovers the Robbins confidence sequence described in the introduction.

3.2.3 Example

Consider the following example with

$$y_i = 1 + 2x_{i1} + 3x_{i2} + 4x_{i3} + 2.3z_i + \varepsilon_i, \tag{15}$$

where $x_{ij} \sim N(0,1)$, $\varepsilon \sim N(0,1)$ and $z_i \sim \text{Bernoulli}(1/2)$. Figure 1 illustrates just how quickly the asymptotic- converges to the nonasymptotic confidence sequence, the width of the confidence sequences relative to classical confidence intervals, and how the classical confidence intervals fail to cover the true estimand at all times.

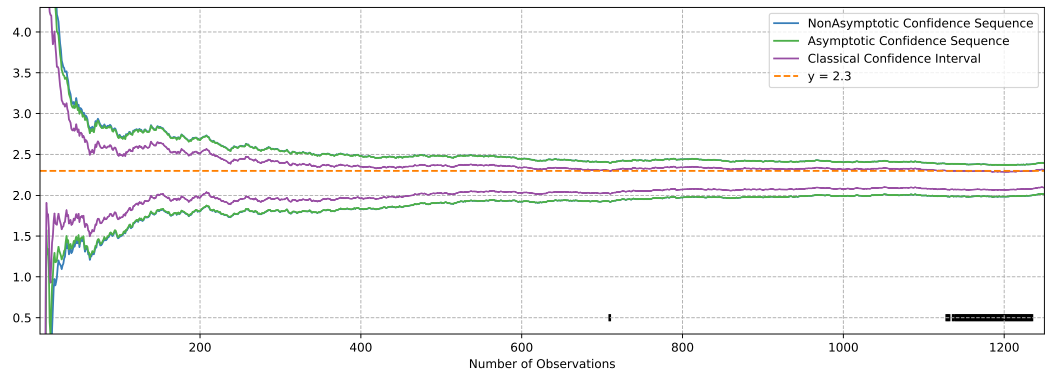

Figure 1: Comparison of nonasymptotic confidence sequence from corollary 3.3 (blue), the asymptotic confidence sequence from corollary 3.5 (green) and classical confidence intervals (purple). Sample sizes where the classical confidence intervals fail to cover the true value $\delta = 2.3$ are indicated with black ticks at the base of the figure. Both confidence sequences were parameterized with $\phi = 1$.

3.3 Sequential F Tests

3.3.1 Nonasymptotic Results

In the preceding sections we focussed on the 1 dimenional case of a single coefficient in a linear model. We now turn out attention to the most general result, namely, sequential F tests. These are the multivariate extensions of the previous sections where we want to test collections of coefficients in a linear model and/or construct confidence sequences. At this point it seems superfluous to state "covariate adjusted" as we are generally testing subsets of covariates in a linear regression model. Much like our handling of the sequential t test, we first provide a nonasymptotic test martingale that has a closed-form expression. Inverting this test martingale yields a confidence sequence for a vector of linear model coefficients, though it is not the most convenient to manipulate. For this reason we provide an asymptotic test martingale that yields a confidence sequence that is much easier to work with - a convex set defined by a quadratic constraint.

Theorem 3.6. *For an iid sequence of observations $y_i \sim N(\boldsymbol{w}_i\boldsymbol{\gamma}, \sigma^2)$ where $\boldsymbol{w}_i = (\boldsymbol{x}_i, z_i)$ and $\boldsymbol{\gamma} = (\boldsymbol{\beta}, \delta)$. For any fixed positive-definie matrix $\boldsymbol{\Phi} \in \mathbb{R}^{d\times d}$, let*

$$B_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) = \begin{cases} \sqrt{\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}}\frac{\left(1+\frac{(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)-\boldsymbol{\delta}_0)'(\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n-\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)-\boldsymbol{\delta}_0)}{s_n^2(\boldsymbol{Y}_n)(n-p-d)}\right)^{-\frac{n-p}{2}}}{\left(1+\frac{(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)-\boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)-\boldsymbol{\delta}_0)}{s_n^2(\boldsymbol{Y}_n)(n-p-d)}\right)^{-\frac{n-p}{2}}} & \boldsymbol{W}_n'\boldsymbol{W}_n \textit{full rank} \\ 0 & \textit{otherwise} \end{cases} \tag{16}$$

Then for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$$\mathbb{P}_\theta[\exists n \in \mathbb{N} : B_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) \geq \alpha^{-1}] \leq \alpha \tag{17}$$

A few remarks are helpful in making sense of this expression. The term $\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n = \boldsymbol{Z}_n'(\boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n = \boldsymbol{Z}_n'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n$ is the component of the information matrix corresponding to $\boldsymbol{\delta}$. Loosely speaking, this is the amount of information containd in the observed sample about $\boldsymbol{\delta}$. Similarly, $\boldsymbol{\Phi}$ can be interpreted as a Gaussian mixture precision (using a mixture $\boldsymbol{\delta} \sim N(\boldsymbol{\delta}_0, \sigma^2\boldsymbol{\Phi})$ in the alternative). The classical F statistic makes an appearance in the denominator as $f_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) = (\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0)ds_n^2(\boldsymbol{Y}_n)$. Let us write $\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n) = (\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n) - \boldsymbol{\delta}_0)/s_n^2(\boldsymbol{Y}_n)$ to make a connection with multivariate t statistics. We define

$$\boldsymbol{t}_n(\boldsymbol{Y}_n) = \frac{\boldsymbol{V}_n'\boldsymbol{Y}_n}{s_n(\boldsymbol{Y}_n)} = \frac{\tilde{\boldsymbol{Z}}_n\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)}{s_n(\boldsymbol{Y}_n)}, \tag{18}$$

where $\boldsymbol{V}_n'$ is a $d \times n$ dimensional matrix satisfying $\boldsymbol{V}_n\boldsymbol{V}_n' = \boldsymbol{P}_{\boldsymbol{W}_n} - \boldsymbol{P}_{\boldsymbol{X}_n}$ obtained from the eigendecomposition, and $\tilde{\boldsymbol{Z}}_n = \boldsymbol{V}_n'\boldsymbol{Z}_n$. The statistic $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ will be important in later sections as it turns out to be a *maximal invariant* test statistic defined in section 5.1. The test martingale can then be restated as

$$B_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) = \begin{cases} \sqrt{\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)}}\frac{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I}_n-\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n')\boldsymbol{t}_n(\boldsymbol{Y}_n)}{(n-p-d)}\right)^{-\frac{n-p}{2}}}{\left(1+\frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'\boldsymbol{t}_n(\boldsymbol{Y}_n)}{(n-p-d)}\right)^{-\frac{n-p}{2}}} & \boldsymbol{W}_n'\boldsymbol{W}_n \textit{full rank} \\ 0 & \textit{otherwise} \end{cases} \tag{19}$$

Corollary 3.7. $p_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) = 1/B_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0)$ *defines a sequential p-value satisfying*

$$\mathbb{P}_\theta[\exists n \in \mathbb{N} : p_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) \leq \alpha] \leq \alpha, \tag{20}$$

for all $\theta \in \boldsymbol{\Theta}_0$

Corollary 3.8. *Let $C_n(\boldsymbol{Y}_n) = \{\boldsymbol{\delta} \in \mathbb{R}^d : B_n(\boldsymbol{Y}_n;\boldsymbol{\delta}) \leq \alpha^{-1}\}$, then $C_n(\boldsymbol{Y}_n)$ defines a confidence sequence satisfying*

$$\mathbb{P}_\theta[\forall n \in \mathbb{N} : \boldsymbol{\delta} \in C_n(\boldsymbol{Y}_n)] \geq 1-\alpha, \tag{21}$$

for all $\theta \in \boldsymbol{\Theta}$.

We remark that these results hold regardless of the nuisance parameters $\boldsymbol{\beta}$ and σ^2

3.3.2 Asymptotic Results

Theorem 3.9.

$$\log B_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) = \log \tilde{B}_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) + o_{a.s.}(1) \tag{22}$$

where

$$\tilde{B}_n(\boldsymbol{Y}_n;\boldsymbol{\delta}_0) = \sqrt{\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})}}e^{\frac{1}{2s_n^2(\boldsymbol{Y}_n)}(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)-\boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}_n{}'\tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}_n{}'\tilde{\boldsymbol{Z}}_n)^{-1}\tilde{\boldsymbol{Z}}_n{}'\tilde{\boldsymbol{Z}}_n(\hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)-\boldsymbol{\delta}_0)} \tag{23}$$

This yields an asymptotic confidence sequence $\tilde{C}_n(\boldsymbol{Y}_n) = \{\boldsymbol{\delta} \in \mathbb{R}^d : \tilde{B}_n(\boldsymbol{Y}_n;\boldsymbol{\delta}) \leq \alpha^{-1}\}$ which is a convex set in $\mathbb{R}^d$ defined by a quadratic constraint. Confidence sequences for functionals of $\boldsymbol{\delta}$ can be computed by maximizing and minimizing them over this convex set.

3.4 Converting Classical to Anytime-Valid (with R Code)

To compute $\tilde{B}_n(\boldsymbol{Y}_n;\delta_0)$, $\tilde{C}_n(\boldsymbol{Y}_n)$ and $\tilde{p}_n(\boldsymbol{Y}_n)$ all we need are a few basic statistics from the classical analysis. If a software library already computes these when performing a classical analysis, then it is trivial to convert this analysis to anytime-avlid by writing a small wrapper. To demonstrate, here are just a few lines of R code that take a convert classical confidence intervals and p-values to anytime valid confidence sequences and sequential p-values.

```
############################## Fit a linear model ##############################
lmfit = lm(outcome ~ . + trt*., data=df)

######################## Extract Sufficient Statistics  ########################
mod = summary(lmfit)
stderrs = mod$coefficients[, ’Std. Error’]
tstats2 = mod$coefficients[, ’t value’]^2
estimates = mod$coefficients[, ’Estimate’]
z2 = (mod$sigma / stderrs) ^ 2

######################### Compute Sequential p-values ##########################
spvals = min(1, sqrt((phi + z2) / phi) * exp(-0.5 * (z2 / (phi + z2)) * tstats2 )

######################## Compute Confidence Sequences ##########################
radii = stderrs * sqrt(log((phi + z2) / (phi * alpha ^2) * (phi + z2) / z2)
lower_cis = estimates - radii
upper_cis = estimates + radii
```

An R package that provides an anytime-valid analogue of the original *summary* function is available (Lindon, 2023). The output is shown below. It converts inferences for all terms in the model to anytime-valid, in addition to F-statistics.

```
> lmfit = lm(outcome ~ . + trt*., data=df)
> avsummary(lmfit)

Call:
lm(formula = y ~ . + trt * ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.00823 -0.56961  0.07473  0.78458  1.92584 

Coefficients:
            Estimate Std. Error t value Seq. p-value   2.5%  97.5%
(Intercept)   0.9163     0.1586   5.776    5.632e-06  0.4021  1.430
X1            1.8723     0.1809  10.352    7.565e-15  1.2921  2.453
X2            2.8503     0.1828  15.590    1.838e-24  2.2643  3.436
X3            3.9056     0.1594  24.499    1.051e-37  3.3891  4.422
trt           2.3224     0.2157  10.769    1.842e-15  1.6393  3.006
X1:trt        2.2998     0.2414   9.527    6.110e-13  1.5404  3.059
X2:trt        3.3117     0.2325  14.243    1.523e-21  2.5786  4.045
X3:trt        0.4801     0.2112   2.273    4.471e-01 -0.1899  1.150

Residual standard error: 1.065 on 92 degrees of freedom
Multiple R-squared:  0.9808,Adjusted R-squared:  0.9793
F-statistic: 671.1 on 7 and 92 DF,  Seq. p-value: < 2.2e-16
```

4 Anytime-Valid Regression-Adjusted Causal Inference

4.1 Classical Asymptotic Regression Adjusted ATE Estimation

In the previous sections we derived confidence sequences for single coefficients of a linear regression model. As motivated in the introduction, linear models remain widely popular among applied researchers and social scientists, particularly those performing causal inference. In practice, however, many of the assumptions of the linear model may break down; the true regression function may not be linear, there may be omitted variables, or the residual variance may not be homoscedastic, so to what extent are these results practically useful beyond the theory? It turns out that the randomization in controlled experiments protects against many of these issues, and that many of these concerns are benign in the large samples (Lin, 2013). As we will show, the confidence sequences presented in equation (14) can be used as an asymptotic confidence sequence for superpopulation average treatment effects when the assignment probabilities are $1/2$ to treatment and control, or when the residual variances are homoscedastic. If either of these fail, the confidence sequence can be trivially rectified by replacing the OLS estimator of the residual variance, $s_n(\boldsymbol{Y}_n)^2$, with the Huber-White estimator.

Suppose a potential outcomes framework. Each experimental unit i has two potential outcomes $y_i^{(1)}$ and $y_i^{(0)}$. Each unit is randomly assigned to the treatment with probability ρ, which we express with a treatment indicator $z_i \sim \text{Bernoulli}(\rho)$, and a potential outcome is observed $y_i^{obs} = z_i y_i^{(1)} + (1 - z_i)y_i^{(0)}$. Our goal is to perform anytime-valid inference for the superpopulation average treatment effect $\delta_{sp} = \mathbb{E}_{sp}[y_i^{(1)} - y_i^{(0)}]$ using a linear regression adjustment typically based on pre-treatment covariates $\boldsymbol{x}_i$ that are believed to be correlated with the potential outcomes. Consider the linear model $y_i = c + z_i\delta + \boldsymbol{x}_i'\boldsymbol{\beta} + \varepsilon_i$. Our aim is to connect inference on δ with inference on δ_{sp}. Although this parameterization can be used in practice it is equivalent without loss of any generality to work with the centered parameterization $y_i = \alpha + (z_i - \rho)\delta + (\boldsymbol{x}_i - \mu_x)'\boldsymbol{\beta} + \varepsilon_i$ where $\boldsymbol{\mu}_x = \mathbb{E}_{sp}[\boldsymbol{x}_i]$. This doesn't change the OLS estimates of $\boldsymbol{\beta}$ or δ and merely simplifies the exposition of the theory. We also note that the following is easily extensible to include interaction terms.

We first recall known fixed-n results that justify the use of linear models, where we closely follow Imbens and Rubin (2015, chapter 7). This serves to introduce some of the notation, but also provides an insightful comparison to the anytime valid results that follow. Let

$$(\alpha^\star, \boldsymbol{\beta}^\star, \delta^\star) = \operatorname{argmin} \mathbb{E}[(y_i - \alpha - (z_i - \rho)\delta - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta})^2] \tag{24}$$

be the population OLS values. It is easily verified that the solutions are

$$\begin{aligned} \alpha^\star &= \rho\mathbb{E}_{sp}[y_i^{(1)}] + (1-\rho)\mathbb{E}_{sp}[y_i^{(0)}] \\ \boldsymbol{\beta}^\star &= \mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)(\boldsymbol{x}_i - \boldsymbol{\mu}_x)']^{-1}\mathbb{E}_{sp}[y_i(\boldsymbol{x}_i - \boldsymbol{\mu}_x)] \\ \delta^\star &= \mathbb{E}_{sp}[y_i^{(1)}] + \mathbb{E}_{sp}[y_i^{(0)}]. \end{aligned} \tag{25}$$

In particular, we have that $\delta^\star = \delta_{sp}$, our causal estimand of interest. Let's write $\boldsymbol{w}_i = [1, \boldsymbol{x}_i - \boldsymbol{\mu}_x, z_i - \rho]$ and $\boldsymbol{\gamma}^\star = (\alpha^\star, \boldsymbol{\beta}^\star, \delta^\star)$. Theory from the literature of M-estimators show that the sample OLS, $\hat{\boldsymbol{\gamma}}_n^{ols}$, has the limiting Gaussian distribution

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}_n^{ols} - \boldsymbol{\gamma}^\star) \xrightarrow{d} N(0, \boldsymbol{\Gamma}^{-1}\boldsymbol{\Delta}\boldsymbol{\Gamma}^{-1}) \tag{26}$$

where

$$\begin{aligned} \boldsymbol{\Gamma} = \mathbb{E}_{sp}[\boldsymbol{w}_i\boldsymbol{w}_i'] &= \mathbb{E}_{sp}\left[\begin{pmatrix} 1 & (\boldsymbol{x}_i - \boldsymbol{\mu}_x)' & (z_i - \rho) \\ (\boldsymbol{x}_i - \boldsymbol{\mu}_x) & (\boldsymbol{x}_i - \boldsymbol{\mu}_x)(\boldsymbol{x}_i - \boldsymbol{\mu}_x)' & (\boldsymbol{x}_i - \boldsymbol{\mu}_x)(z_i - \rho) \\ (z_i - \rho) & (z_i - \rho)(\boldsymbol{x}_i - \boldsymbol{\mu}_x)' & (z_i - \rho)^2 \end{pmatrix}\right], \\ &= \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)(\boldsymbol{x}_i - \boldsymbol{\mu}_x)'] & 0 \\ 0 & 0 & \rho(1-\rho) \end{pmatrix} \end{aligned} \tag{27}$$

and

$$\boldsymbol{\Delta} = \mathbb{V}_{sp}[\boldsymbol{w}_i y_i - \boldsymbol{w}_i\boldsymbol{w}_i\boldsymbol{\gamma}^\star] = \mathbb{E}_{sp}\left[(y_i - \alpha^\star - (z_i - \rho)\delta^\star - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)\boldsymbol{\beta}^\star)^2\boldsymbol{w}_i\boldsymbol{w}_i'\right] \tag{28}$$

Reading of the marginal of the multivariate Gaussian corresponding to δ_n^{ols} provides the primary result of interest.

$$\sqrt{n}(\hat{\delta}_n^{ols} - \delta_{sp}) \xrightarrow{d} N\left(0, \frac{\mathbb{E}_{sp}[(z_i - \rho)^2(y_i - \alpha^\star - (z_i - \rho)\delta_{sp} - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta}^\star)^2]}{\rho^2(1-\rho)^2}\right) \tag{29}$$

In practice, one must use a consistent estimator of the variance term in equation (29). In general, the Huber-White estimator can be used. When the errors are homoscedastic, the standard OLS estimator is consistent. In fact, one can still get away with using the standard OLS estimator even in the presence of heteroscedasticity as long as $\rho = 1/2$ because it converges to a value larger than the true variance, yielding a conservative test and confidence interval (Lin, 2012). We deviate from convention slightly here for reasons that will be apparent later by rewriting equation (29) as

$$\hat{\delta}_n^{ols} - \delta_{sp} \xrightarrow{d} N\left(0, \frac{\sigma_\delta^2}{n\rho(1-\rho)}\right), \tag{30}$$

where $\sigma_\delta^2 = \mathbb{E}_{sp}[(z_i - \rho)^2(y_i - \alpha^\star - (z_i - \rho)\delta_{sp} - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta}^\star)^2]/\rho(1-\rho)$ (this simplifies to $\mathbb{E}_{sp}[(y_i - \alpha^\star - (z_i - \rho)\delta_{sp} - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta}^\star)]$ when the residual variance is homogeneous). Loosely speaking, we interpret this expression as approximately having observed n units of information about δ_{sp}, each unit containing an amount $\sigma_\delta^2/\rho(1-\rho)$ of information. Clearly as $\rho \to 0$ or $\rho \to 1$, observations are arriving almost exclusively from the control or treatment, and little information about δ_{sp} is gained. We now turn our attention to providing the analogous anytime-valid result.

4.2 Anytime-Valid Asymptotic Regression Adjusted ATE Estimation

The essential idea is that using a linear model is fine. If the errors are heteroscedastic then we must be careful to use robust standard errors using the Huber-White estimator, but if the assignment probability to the treatment is $1/2$ or the residuals are homoscedastic then the linear model standard error will suffice. We state this formally in the following theorem. All proofs for this section are contained in appendix section A.9.

Theorem 4.1. *Assume that the covariates $\boldsymbol{x}_i$ are iid and that $\boldsymbol{x}_i$ in addition to the potential outcomes have bounded second moments. In addition, assume that the assignment mechanism is independent of the potential outcomes and the covariates. Let $\hat{\delta}_n(\boldsymbol{Y}_n)$ be the ordinary least squares estimate of the main effect and*

$$r_n(\boldsymbol{Y}_n) = \frac{\hat{\sigma}_{\delta,n}(\boldsymbol{Y}_n)}{\|\tilde{\boldsymbol{Z}}_n\|}\sqrt{\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right)}, \tag{31}$$

where $\hat{\sigma}_{\delta,n}(\boldsymbol{Y}_n)$ is a strongly consistent estimator of the residual variance. Then $C_n(\boldsymbol{Y}_n) := (\hat{\delta}_n(\boldsymbol{Y}_n) - r_n(\boldsymbol{Y}_n), \hat{\delta}_n(\boldsymbol{Y}_n) + r_n(\boldsymbol{Y}_n))$ is an asymptotic confidence sequence for the superpopulation average treatment effect $\delta_{sp} = \mathbb{E}_{sp}[y_i^{(1)} - y_i^{(0)}]$ with rate $o_{a.s.}(\sqrt{\log n/n})$. When the assignment probability to the treatment is $1/2$ or when the residuals have homoscedastic variance, $s_n(\boldsymbol{Y}_n)$ can be used for $\hat{\sigma}_{\delta,n}(\boldsymbol{Y}_n)$, otherwise the Huber-White estimator is required.

This theorem takes the same justification for using linear regression in the classical fixed-n case and extends it to our anytime-valid case. Naturally, the linear model assumes homoscedastic error terms, which is why the standard error provided by the linear model is correct when the errors are homoscedastic. What is less expected is that it is conservative when the errors are heteroscedastic and the treatment assignment probability is $1/2$. Otherwise, robust standard errors must be used. As we are working asymptotically, there is an alternative result which is more reminiscent of equation (30). We provide this alternative for completeness.

Theorem 4.2. *Under the assumptions of theorem 4.1, let*

$$r_n(\boldsymbol{Y}_n) := \sqrt{\frac{\hat{\sigma}^2_{\delta,n}}{n\rho(1-\rho)}}\sqrt{\frac{\phi + n\rho(1-\rho)}{n\rho(1-\rho)}\log\left(\frac{\phi + n\rho(1-\rho)}{\phi\alpha^2}\right)}, \tag{32}$$

then $C_n(\boldsymbol{Y}_n) := (\hat{\delta}_n(\boldsymbol{Y}_n) - r_n(\boldsymbol{Y}_n), \hat{\delta}_n(\boldsymbol{Y}_n) + r_n(\boldsymbol{Y}_n))$ is an asymptotic confidence sequence for the superpopulation average treatment effect $\delta_{sp} = \mathbb{E}_{sp}[y_i^{(1)} - y_i^{(0)}]$ with rate $o_{a.s.}(\sqrt{\log n/n})$.

Theorem 4.2 simply replaces $\|\tilde{\boldsymbol{Z}}_n\|_2^2$ with its limit $n\rho(1-\rho)$ as we are working asymptotically anyway.

This is a fully nonparametric asymptotic regression adjusted confidence sequence for the superpopulation average treatment effect. It assumes neither that the true regression function is linear, nor that there are no omitted variables, nor that the residual variance is homoscedastic. Compare this with the confidence sequence for the mean of iid Gaussians in equation (3). The only difference is that it is centred at $\hat{\delta}_n^{ols}$ instead of the sample mean, the standard error for the sample mean is replaced by the standard error for $\hat{\delta}_n^{ols}$, and n is replaced with $n\rho(1-\rho)$. All of these seem to be appropriate replacements, that we could have perhaps guessed. In order to use this in practice one must use a strongly consistent estimator for σ_δ^2. The Huber-White estimator can be used in all circumstances, such as with heteroscedastic errors and unbalanced assignment probabilities. If the errors are heteroskedastic, but $\rho = 1/2$, then the standard OLS estimator $s_n^2(\boldsymbol{Y}_n)$ converges almost surely to a limit that is greater than σ_δ^2 which, though not optimal, still provides a valid asymptotic confidence sequence. If homoscedasticity can be assumed, the standard OLS estimator s^2 is strongly consistent for arbitrary ρ.

To demonstrate the reach of this result, we provide an example in which the true data generating process is *nonlinear*, *non-Gaussian*, *heterogeneous* residual variance and has *heterogeneous* treatment effects.

4.2.1 Example: Binary outcomes with heterogeneous treatment effects

Consider the following simulation with

$$\begin{aligned}
y_i|\boldsymbol{x}_i, z_i &\sim \text{Bernoulli}(p_i(\boldsymbol{x}_i, z_i)) \\
p_i(\boldsymbol{x}_i, z_i) &= \text{logistic}(-2 + x_{i1}^2 - 0.5\sin(x_{i2}) - 0.3|x_{i3}| + 0.2z_i + 0.1z_ix_{i1}) \\
z_i &\sim \text{Bernoulli}(0.25) \\
\boldsymbol{x}_i &\sim N((1,2,3), \boldsymbol{I}_3).
\end{aligned} \tag{33}$$

The treatment is assigned with probability $1/4$ and there is an interaction term with the treatment and covariate x_{i1}. Expected potential outcomes were computed numerically with $\mathbb{E}[y_i(1)] = 0.32$, $\mathbb{E}[y_i(0)] = 0.28$ and the average treatment effect $\eta = 0.04$ to two decimal places.

Figure 2 shows the application of the asymptotic confidence sequences in comparison to classical confidence intervals for covering the average treatment effect. Both use the same linear model with interaction terms of the main effect with the centered nuisance covariates (Lin, 2012) i.e.

$$\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{Z}_n\delta + \boldsymbol{Z}_n(\boldsymbol{X}_n - \bar{\boldsymbol{X}}_n)\boldsymbol{\psi} + \boldsymbol{\varepsilon}_n, \tag{34}$$

though to be consistent with our earlier notation we absorb the $\boldsymbol{Z}_n(\boldsymbol{X}_n - \bar{\boldsymbol{X}}_n)\boldsymbol{\psi}$ term into the $\boldsymbol{X}_n\boldsymbol{\beta}$ term as we consider them all nuisance parameters when we only care about estimating the ATE. Figure 2 shows that the asymptotic confidence sequence covers the ATE at all sample sizes, whereas the classical confidence interval does not, excluding it for large stretches as indicated by the black ticks at the bottom of the figure.

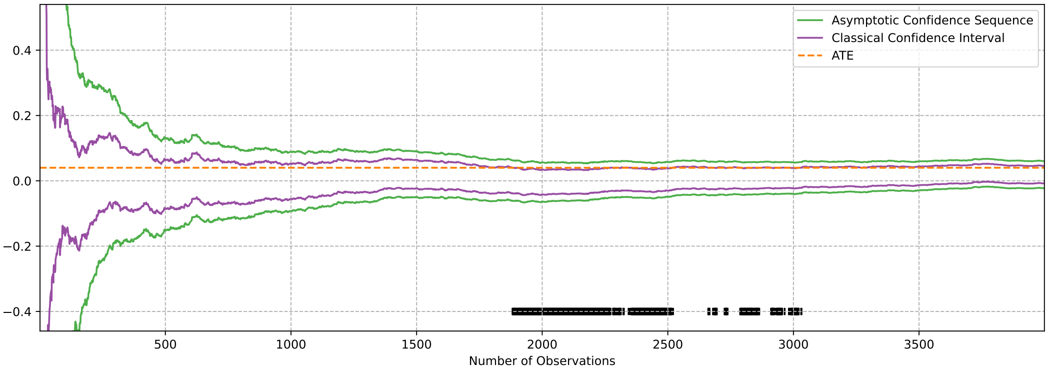

Figure 2: Comparison of the asymptotic confidence sequence from corollary 3.5 (green) with $\phi = 100$ and classical confidence intervals (purple). Sample sizes where the classical confidence intervals fail to cover the average treatment effect of 0.04 are indicated with black ticks at the base of the figure.

5 Group Theory Results

Proving these results will require a modicum of group theory. Applications of group theory to statistics be found in Eaton (1989); Lehmann and Romano (2005); Wijsman (1990). The main challenge in performing anytime-valid inference in linear models is the presence of nuisance parameters - we require time-uniform Type-I error and coverage guarantees for all possible values of the nuisance parameters $\boldsymbol{\beta}$ and σ^2. By way of introduction, we provide a sketch of the argument. Suppose that instead of observing the raw sequence of observations, the statistician is provided instead with a sequence $\boldsymbol{t}_1, \boldsymbol{t}_2, \dots$ of t-statistics. The statistician could perform a sequential probability ratio test using the sequence of t-statistics instead of the raw sequence. The advantage is the sequence of t-statistics depends neither on the nuisance parameters $\boldsymbol{\beta}$ and σ^2 under the null nor under the alternative, and so a Type I error guarantee is obtained regardless of the nuisance parameters. This holds in general for sequences of maximal invariant statistics, and SPRT's constructed in this was are called *invariant SPRT* (Lai, 1981). Another advantage is that the maximal invariant statistics summarizes all information in the sequence, and so the likelihood ratio of the most recent maximal invariant statistic can be used, instead of the whole sequence. For testing a composite alternative, a mixture of invariant SPRT's can be used. This is also equivalent to computing a Bayes-factor/mixture-martingale on the raw sequence of observations when using a special kind of prior/mixture, namely, the *right-Haar* prior/mixture on the nuisances parameters. Optional stopping behaviour of Bayes factors under group invariant models are studied in Hendriksen et al. (2021). Similarly, e-values for group invariant models are studied in Pérez-Ortiz et al. (2022).

5.1 Group Theory Supplement

We deal with the nuisance parameters in the composite null via group invariance arguments. To demonstrate the group invariance structure of the model (6) it necessary to reparameterize in terms of $\boldsymbol{\xi} = \boldsymbol{\delta}/\sigma$, where $\boldsymbol{\xi}$ are the *standardized* coefficients.

Let the parameters be denoted by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma, \boldsymbol{\xi}) \in \boldsymbol{\Theta}$, with $\boldsymbol{\Theta} = \mathbb{R}^p \times \mathbb{R}^+ \times \mathbb{R}^d$. The null parameter space $\boldsymbol{\Theta}_0 = \mathbb{R}^p \times \mathbb{R}^+ \times \{\boldsymbol{0}\}$ and $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta} \setminus \boldsymbol{\Theta}_0$. The model is invariant under the following transformations

$$\begin{aligned} g_{\boldsymbol{\alpha},c} &: \boldsymbol{Y} \mapsto c\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\alpha}, \\ \bar{g}_{\boldsymbol{\alpha},c} &: (\boldsymbol{\beta}, \sigma, \boldsymbol{\xi}) \mapsto (c\boldsymbol{\beta} + \boldsymbol{\alpha}, c\sigma.\boldsymbol{\xi}). \end{aligned} \tag{35}$$

In other words, the transformed observation $g_{\boldsymbol{\alpha},c}(\boldsymbol{Y})$ belongs to the same family of Gaussian linear models with transformed parameters $\bar{g}_{\boldsymbol{\alpha},c}(\boldsymbol{\theta})$. Let the group of transformations that act on the outcome and parameter space be denoted $G = \{g_{\boldsymbol{\alpha},c} : \alpha \in \mathbb{R}^p, c \in \mathbb{R}\}$ and $\bar{G} = \{\bar{g}_{\boldsymbol{\alpha},c} : \alpha \in \mathbb{R}^p, c \in \mathbb{R}\}$ respectively, noting that these are common to both the null and alternative hypotheses and leave $\boldsymbol{\xi}$ unchanged.

The *orbit* of $\boldsymbol{Y}$ is defined as $\mathcal{O}(\boldsymbol{Y}) = \{g(\boldsymbol{Y}), g \in G\}$. A function ϕ is G-invariant if $\lambda(\boldsymbol{Y}) = \lambda(g(\boldsymbol{Y}))$ for all $g \in G$, that is, it is constant on orbits. A *test function* is a function used to reject the null hypothesis when $\lambda(\boldsymbol{Y}) > c$. As an equivalent model is obtained under transformations (35), we should reasonably expect that a test function is G-invariant. For instance, it should not matter if the units of $\boldsymbol{Y}$ are changed or if the component of $\boldsymbol{Y}$ in $\mathcal{C}(\boldsymbol{X})$ is changed given we are testing a hypothesis about the component in $\mathcal{C}(\boldsymbol{W}) \setminus \mathcal{C}(\boldsymbol{X})$. It is helpful to regard all elements of an orbit as carrying the same amount of evidence against the null hypothesis. It is known that the likelihood ratio test statistic and the Bayes factor resulting from the use of the right-Haar prior are G-invariant (Hendriksen et al., 2021).

Definition 5.1. A *maximal invariant* function M is a function that is constant on orbits and takes distinct values on each orbit, that is, $M(\boldsymbol{Y}_1) = M(\boldsymbol{Y}_2)$ implies $\boldsymbol{Y}_1 = g(\boldsymbol{Y}_2)$ for some $g \in G$.

A maximal invariant statistic is simply a maximal invariant function of the data.

Lemma 5.2. *A test function $\lambda(\boldsymbol{Y})$ is invariant if and only if it is a function of a maximal invariant statistic.*

The proof is given in appendix A.2. Both likelihood ratio test statistic and our test statistic $B_n(\boldsymbol{Y}_n)$ in equation (19) can be written in terms of

$$\boldsymbol{t}(\boldsymbol{Y}) = \frac{\boldsymbol{V}'\boldsymbol{Y}}{s(\boldsymbol{Y})} = \frac{\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s(\boldsymbol{Y})}, \tag{36}$$

where $\boldsymbol{V}'$ is a $d \times n$ dimensional matrix satisfying $\boldsymbol{V}\boldsymbol{V}' = \boldsymbol{P}_{\boldsymbol{W}} - \boldsymbol{P}_{\boldsymbol{X}}$ obtained from the eigen-decomposition and $\tilde{\boldsymbol{Z}} = \boldsymbol{V}'\boldsymbol{Z}$. It is related to the f-statistic by $f(\boldsymbol{Y}) = \boldsymbol{t}(\boldsymbol{Y})'\boldsymbol{t}(\boldsymbol{Y})/d$.

Proposition 5.3. *The statistic $\boldsymbol{t}(\boldsymbol{Y})$ is a maximal invariant statistic under G.*

The proof is given in appendix A.2. By noting that $\boldsymbol{V}_d'\boldsymbol{Y} = \tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y}) \sim N(\tilde{\boldsymbol{Z}}\sigma\boldsymbol{\xi}, \sigma^2\boldsymbol{I}_d)$ and $s^2(\boldsymbol{Y}) \sim \chi^2_{n-p-d}$, it follows that

$$\boldsymbol{t}(\boldsymbol{Y})_i | \boldsymbol{\xi} \sim t^{nc}_{n-p-d}((\tilde{\boldsymbol{Z}}\boldsymbol{\xi})_i),$$

namely, the i'th component of $\boldsymbol{t}(\boldsymbol{Y})$ is an independent noncentral t distribution with $n-p-d$ degrees of freedom and noncentrality parameter $(\tilde{\boldsymbol{Z}}\boldsymbol{\xi})_i$.

Critically, the distribution of $\boldsymbol{t}(\boldsymbol{Y})$, and hence the distribution of an invariant test statistic $\lambda(\boldsymbol{Y}) = h(\boldsymbol{t}(\boldsymbol{Y}))$, is independent of the nuisance parameters $(\boldsymbol{\beta}, \sigma^2)$ and depends *only* on $\boldsymbol{\xi}$. Under the null hypothesis, the distribution of the i'th component of $\boldsymbol{t}(\boldsymbol{Y})$ is

$$\boldsymbol{t}(\boldsymbol{Y})_i \sim t_{n-p-d},$$

namely, independent t-statistics with $n-p-d$ degrees of freedom. This is how composite null hypotheses can be simplified down to simple null hypotheses via group invariance arguments.

5.2 Sequential F-Tests via Group Invariant Bayes Factors

The main challenge we face in developing sequential F-tests is that the null is composite. We require our test statistic be a test martingale *for all* $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, that is, for all values of the nuisance parameters. The previous section 5.1 showed that a composite null can be reduced down to a simple null when we restrict ourselves to consider test statistics that are invariant under G, as the distribution of the maximal invariant depends only on $\boldsymbol{\xi}$ in both the null and the alternative. We could therefore consider the following test statistic

$$B_n(\boldsymbol{Y}_n) := \frac{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \dots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \dots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)} = \frac{\int p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \dots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{\xi}, H_1) p(\boldsymbol{\xi}|H_1) d\boldsymbol{\xi}}{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \dots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)}, \tag{37}$$

where $\boldsymbol{\xi}|H_1 \sim N(0, \boldsymbol{\Phi}^{-1})$, instead of the Bayes factor based on the original sample $\boldsymbol{Y}_n$. In other words we could consider the sequence $\{\boldsymbol{t}_i(\boldsymbol{Y}_i)\}_{i=1}^{\infty}$ instead of the original sequence $\{\boldsymbol{Y}_i\}_{i=1}^{\infty}$. The n-fold product density is identical for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and it follows that $B_n(\boldsymbol{Y}_n)$ is a nonnegative supermatringale for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}[A_n|\sigma(t_1, \dots, t_{n-1})] &= A_{n-1}\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{p(t_n|t_1, \dots, t_{n-1}, H_1)}{p(t_n|t_1, \dots, t_{n-1}, H_0)}|\sigma(t_1, \dots, t_{n-1})\right] \\ &= A_{n-1} \end{aligned} \tag{38}$$

an α-level sequential test can be constructed by rejecting as soon as this statistic exceeds α^{-1} via Ville's inequality (Ville, 1939). However, this would require evaluating a complicated product density. Fortunately, it simplifies due to the properties of maximal invariants.

Theorem 5.4.

$$\frac{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)} = \frac{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \dots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \dots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)}. \tag{39}$$

The proof is given in appendix A.4. The consequence of this theorem is that we only need to evaluate the density of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ up to a normalizing constant under H_1 and H_0, instead of the n-fold product density. This is provided in the following theorem.

Theorem 5.5. *The distribution of the maximal invariant statistic under H_0 and H_1 is*

$$\boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{\beta}, \sigma, H_0 \sim t_{n-p-d}(\boldsymbol{0}, \boldsymbol{I}_d), \tag{40}$$

$$\boldsymbol{t}_n(\boldsymbol{Y}_n)|\boldsymbol{\beta}, \sigma, H_1 \sim t_{n-p-d}(\boldsymbol{0}, \boldsymbol{I}_d + \tilde{\boldsymbol{Z}}_n \Phi^{-1} \tilde{\boldsymbol{Z}}_n), \tag{41}$$

which are both independent of the nuisance parameters $\boldsymbol{\beta}$ and σ. The statistic $B_n(\boldsymbol{Y}_n)$ can be expressed as the likelihood ratio based on $\boldsymbol{t}_n(\boldsymbol{Y}_n)$

$$B_n(\boldsymbol{Y}_n) = \frac{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_1)}{p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_0)} \tag{42}$$

The proof is given in appendix A.5. Equation (42) establishes the densities of the maximal invariant statistic required to evaluate (39) and finally reveals the origins of our test statistic $B_n(\boldsymbol{Y}_n)$. Hence, we have established that $B_n(\boldsymbol{Y}_n)$ is a nonnegative supermartingale for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ which proves theorem 3.6.

The proof our main result in theorem 3.6 is already complete, but it may interest the reader that this was not our original path in finding $B_n(\boldsymbol{Y}_n)$. Instead, consider the following theorem.

Theorem 5.6. *Let $\boldsymbol{\xi}|\sigma, H_1 \sim N(\boldsymbol{0}, \boldsymbol{\Phi}^{-1})$ and $p(\boldsymbol{\beta}, \sigma|H_1) = p(\boldsymbol{\beta}, \sigma|H_0) \propto 1/\sigma$ be the right-Haar prior. Then*

$$\begin{aligned} B_n(\boldsymbol{Y}_n) &= \frac{\int p(\boldsymbol{Y}_n|\boldsymbol{\beta}, \boldsymbol{\xi}, \sigma, H_1) p(\boldsymbol{\xi}|\sigma, H_1) p(\boldsymbol{\beta}, \sigma|H_1) d\boldsymbol{\beta} d\boldsymbol{\xi} d\sigma}{\int p(\boldsymbol{Y}_n|\boldsymbol{\beta}, \sigma, H_0) p(\boldsymbol{\beta}, \sigma|H_0) d\boldsymbol{\beta} d\boldsymbol{\xi} d\sigma} \\ &= \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n)^{\frac{1}{2}}} \frac{\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)'(\boldsymbol{I} - \tilde{\boldsymbol{Z}}_n(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}_n' \tilde{\boldsymbol{Z}}_n)^{-1} \tilde{\boldsymbol{Z}}_n') \boldsymbol{t}_n(\boldsymbol{Y}_n)}{n-p-d}\right)^{-\frac{n-p}{2}}}{\left(1 + \frac{\boldsymbol{t}_n(\boldsymbol{Y}_n)' \boldsymbol{t}_n(\boldsymbol{Y}_n)}{n-p-d}\right)^{-\frac{n-p}{2}}} \end{aligned} \tag{43}$$

The proof is given in appendix A.6 by direct computation. The expression in equation (43) is equal to (16) except that we now express it in terms of the maximal invariant statistic $\boldsymbol{t}_n(\boldsymbol{Y}_n) = \tilde{\boldsymbol{Z}}_n \hat{\boldsymbol{\delta}}_n(\boldsymbol{Y}_n)/s_n(\boldsymbol{Y}_n)$. The data only enters the Bayes factor through the maximal invariant statistic $\boldsymbol{t}_n(\boldsymbol{Y}_n)$. By lemma 5.2 the Bayes factor is invariant. The distribution of the Bayes factor is completely specified under the null and only depends on $\boldsymbol{\xi}$ under the alternative as the data enters only through the maximal invariant test statistic $\boldsymbol{t}_n(\boldsymbol{Y}_n)$. That the distribution of the Bayes factor is independent of the nuisance parameters when using the right-Haar prior is to be expected from the general result of Dass and Berger (2003, Theorem 1). A decision maker who observes the full sequence $\boldsymbol{Y}_n = (y_1, y_2, \ldots, y_n)$ up to time n has no additional information than another decision maker who is provided with $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ at time n. This is the case because the Bayes factor based on $\boldsymbol{Y}_n$ is equal to the Bayes factor based on $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ by equation (42). This means we don't lose anything by working with $\{\boldsymbol{t}_i(\boldsymbol{Y}_i)\}_{i=1}^{\infty}$ instead of $\{y_i\}_{i=1}^{\infty}$. This result is expected when using the right-Haar prior from the general result of Berger et al. (1998, Theorem 2.1).

The construction of our test martingale through Theorem 5.6 matches the original idea of obtaining a martingale through the method of mixtures. A multivariate Gaussian mixture is used for $\boldsymbol{\xi}$ and the right-Haar mixture is used for the nuisance parameters. Although we borrow these ideas from Bayesian analysis, we do not want to confuse the reader. We have not developed a Bayesian procedure. It is true that the Bayes factor that we compute could be used to compute posterior probabilities over H_1 and H_0, but our goal has always been to provide a procedure for anytime-valid inference, that is, to provide *frequentist* α-level sequential tests and $1-\alpha$ confidence sequences. These guarantees are unaffected by the choice of mixture/prior. These guarantees hold for all values of the nuisance parameters, and not just under the Bayes marginal distributions of H_1 and H_0. It may be helpful to simply regard $B_n(\boldsymbol{Y}_n)$ as a test statistic of which we study the frequentist properties.

6 Discussion

This paper was motivated by a demand to perform regression-adjusted causal inference in real-time over streams of experimental data. At tech companies, regression adjustment has the ability to dramatically reduce uncertainty on treatment effects because of the wealth of pre-treatment covariate information they have on users. For example, companies often have pre-treatment measurements for each user which are highly correlated with their post-treatment outcome. In addition to the demand for regression adjustment, there is also the demand to continuously monitor experiments. This allows experiments to be orchestrated algorithmically, such as stopping immediately when the treatment effect is known with confidence to be negative, or when the best treatment group has been identified. Removing the human element allows tech companies to scale their experimentation operations.

To this end we provided sequential t-tests, F-tests, and confidence sequences for collections of regression coefficients in linear models, modernizing classical inference to anytime-valid inference. Moreover, the scope of our results is not limited to linear model assumptions. Much like randomization justifies the use of OLS estimators of treatment effects in large samples, our confidence sequences yield asymptotic nonparametric confidence sequences for the average treatment effect, even when the linear model is misspecified.

Although it was not our primary motivation, our results have interesting implications for the use of linear models in applied research such as in the social sciences. As our results are anytime-valid, it is not possible for an applied researcher (intentionally or unintentionally) to obtain statistically significant results through optional stopping or optional continuation. Although there are many ways to p-hack, anytime-valid inference removes at least 1 degree of freedom by making it impossible to sample to a foregone conclusion. Journal reviewers and editors could use our sequential p-values and confidence sequences to screen or gain further confidence in submissions, helping to improve the overall reproducibility of published research. Moreover, our sequential p-values and confidence sequences require nothing more than the same statistics and estimates from the classical analysis, making it no harder to implement. Published research can be revisited and reevaluated, and software libraries easily converted.

References

Angrist, J. D. and A. B. Krueger (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics 106*(4), 979–1014.

Armitage, P. (1993). *Interim Analyses in Clinical Trials*. Multiple Comparisons, Selection and Applications in Biometry. CRC Press.

Armitage, P., C. K. McPherson, and B. C. Rowe (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General) 132*(2), 235–244.

Benjamin, D. J., J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, D. Cesarini, C. D. Chambers, M. Clyde, T. D. Cook, P. De Boeck, Z. Dienes, A. Dreber, K. Easwaran, C. Efferson, E. Fehr, F. Fidler, A. P. Field, M. Forster, E. I. George, R. Gonzalez, S. Goodman, E. Green, D. P. Green, A. G. Greenwald, J. D. Hadfield, L. V. Hedges, L. Held, T. Hua Ho, H. Hoijtink, D. J. Hruschka, K. Imai, G. Imbens, J. P. A. Ioannidis, M. Jeon, J. H. Jones, M. Kirchler, D. Laibson, J. List, R. Little, A. Lupia, E. Machery, S. E. Maxwell, M. McCarthy, D. A. Moore, S. L. Morgan, M. Munafó, S. Nakagawa, B. Nyhan, T. H. Parker, L. Pericchi, M. Perugini, J. Rouder, J. Rousseau, V. Savalei, F. D. Schönbrodt, T. Sellke, B. Sinclair, D. Tingley, T. Van Zandt, S. Vazire, D. J. Watts, C. Winship, R. L. Wolpert, Y. Xie, C. Young, J. Zinman, and V. E. Johnson (2018, Jan). Redefine statistical significance. *Nature Human Behaviour 2*(1), 6–10.

Berger, J. O., L. R. Pericchi, and J. A. Varshavsky (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 60*(3), 307–321.

Bibaut, A., N. Kallus, and M. Lindon (2022). Near-optimal non-parametric sequential tests and confidence sequences with possibly dependent observations.

Card, D. and A. B. Krueger (1994, September). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *American Economic Review 84*(4), 772–793.

Carney, D. R., A. J. Cuddy, and A. J. Yap (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological science 21*(10), 1363–1368.

Clearinghouse, W. W. (2022). Procedures and standards handbook, version 5.0. Retrieved from `https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf`.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science 349*(6251), aac4716.

Cook, T. D., D. T. Campbell, and W. Shadish (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin Boston, MA.

Cornfield, J. (1966). A bayesian test of some classical hypotheses, with applications to sequential clinical trials. *Journal of the American Statistical Association 61*(315), 577–594.

Dass, S. C. and J. O. Berger (2003). Unified conditional frequentist and bayesian testing of composite hypotheses. *Scandinavian Journal of Statistics 30*(1), 193–210.

Deng, A., Y. Xu, R. Kohavi, and T. Walker (2013). Improving the sensitivity of online controlled experiments by utilizing pre-experiment data. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, New York, NY, USA, pp. 123–132. Association for Computing Machinery.

Eaton, M. L. (1989). Group invariance applications in statistics. *Regional Conference Series in Probability and Statistics 1*, i–133.

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh Oliver & Boyd.

Grünwald, P., R. de Heide, and W. Koolen (2021). Safe testing.

Ham, D. W., I. Bojinov, M. Lindon, and M. Tingley (2022). Design-based confidence sequences for anytime-valid causal inference.

Hendriksen, A., R. de Heide, and P. Grünwald (2021). Optional Stopping with Bayes Factors: A Categorization and Extension of Folklore Results, with an Application to Invariant Situations. *Bayesian Analysis 16*(3), 961 – 989.

Higgins, J. P., J. Thomas, J. Chandler, M. Cumpston, T. Li, M. J. Page, and V. A. Welch (2019). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Huitema, B. (2011). *The analysis of covariance and alternatives: Statistical methods for experiments, quasi-experiments, and single-case studies*. John Wiley & Sons.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine 2*(8), e124.

Johari, R., P. Koomen, L. Pekelis, and D. Walsh (2017). Peeking at a/b tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, New York, NY, USA, pp. 1517–1525. Association for Computing Machinery.

Johari, R., P. Koomen, L. Pekelis, and D. Walsh (2021). Always valid inference: Continuous monitoring of a/b tests. *Operations Research*.

John, L. K., G. Loewenstein, and D. Prelec (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science 23*(5), 524–532.

Lai, T. L. (1981). Asymptotic Optimality of Invariant Sequential Probability Ratio Tests. *The Annals of Statistics 9*(2), 318 – 333.

Lehmann, E. L. and J. P. Romano (2005). *Testing statistical hypotheses* (Third ed.). Springer Texts in Statistics. New York: Springer.

Lin, W. (2012, August). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *arXiv e-prints*, arXiv:1208.2301.

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics 7*(1), 295 – 318.

Lindon, M. (2023). Roshi. GitHub repository.

Lindon, M. and A. Malek (2020). Anytime-valid inference for multinomial count data.

Lindon, M., C. Sanden, and V. Shirikian (2022). Rapid regression detection in software deployments through sequential testing. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, New York, NY, USA, pp. 3336–3346. Association for Computing Machinery.

Maharaj, A., R. Sinha, D. Arbour, I. Waudby-Smith, S. Z. Liu, M. Sinha, R. Addanki, A. Ramdas, M. Garg, and V. Swaminathan (2023, apr). Anytime-valid confidence sequences in an enterprise a/b testing platform. In *Companion Proceedings of the ACM Web Conference 2023*. ACM.

Morrow, G. and W. Philipp (1982). An almost sure invariance principle for hilbert space valued martingales. *Transactions of the American Mathematical Society 273*(1), 231–251.

Neyman, J. (1923). edited and translated by dorota m. dabrowska and terrence p. speed (1990). on the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science 5*(4), 465–472.

Noordzij, M., G. Tripepi, F. W. Dekker, C. Zoccali, M. W. Tanck, and K. J. Jager (2010, 01). Sample size calculations: basic principles and common pitfalls. *Nephrology Dialysis Transplantation 25*(5), 1388–1393.

Pérez-Ortiz, M. F., T. Lardy, R. de Heide, and P. Grünwald (2022). E-statistics, group invariance and anytime valid testing.

Ramdas, A., J. Ruf, M. Larsson, and W. Koolen (2022). Admissible anytime-valid sequential inference must rely on nonnegative martingales.

Ranehill, E., A. Dreber, M. Johannesson, S. Leiberg, S. Sul, and R. A. Weber (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological science 26*(5), 653–656.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society 58*(5), 527 – 535.

Robbins, H. (1970). Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics 41*(5), 1397–1409.

Sagarin, B. J., J. K. Ambler, and E. M. Lee (2014). An ethical approach to peeking at data. *Perspectives on Psychological Science 9*(3), 293–304.

Services, A. W. (2023). Calculating expected results for cloudwatch anomaly detection. `https://docs.aws.amazon.com/AmazonCloudWatch/latest/monitoring/CloudWatch-Evidently-calculate-results.html`. Accessed on May 10, 2023.

Strassen, V. (1964). An invariance principle for the law of the iterated logarithm. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 3*(3), 211–226.

Strassen, V. (1967). Almost sure behavior of sums of independent random variables and martingales. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 3, pp. 315. Univ of California Press.

Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology 51*, 309–317.

Ville, J. (1939). *Étude critique de la notion de collectif*.

Wald, A. (1945, 06). Sequential tests of statistical hypotheses. *Ann. Math. Statist. 16*(2), 117–186.

Wald, A. (1947). *Sequential analysis*. J. Wiley & sons, Incorporated.

Wasserstein, R. L. and N. A. Lazar (2016). The asa statement on p-values: Context, process, and purpose. *The American Statistician 70*(2), 129–133.

Waudby-Smith, I., D. Arbour, R. Sinha, E. H. Kennedy, and A. Ramdas (2021). Time-uniform central limit theory, asymptotic confidence sequences, and anytime-valid causal inference.

Waudby-Smith, I., L. Wu, A. Ramdas, N. Karampatziakis, and P. Mineiro (2022). Anytime-valid off-policy inference for contextual bandits.

Wijsman, R. (1990). *Invariant Measures on Groups and Their Use in Statistics*. IMS Lecture Notes. Institute of Mathematical Statistics.

Wright, P. (1928). *The Tariff on Animal and Vegetable Oils*. Investigations in international commercial policies. Macmillan.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research 20*(7), 557–585.

A Appendix

A.1 Review: The classical fixed-n F-Test

As this section concerns fixed-n case, we drop the n superscript from $\boldsymbol{Y}_n$ to simplify the exposition. All defitions are contained in the notation section 3.1. An α-level test of $H_0 : \boldsymbol{\delta} = 0$, without loss of generality, can be obtained by examining the likelihood ratio test statistic

$$\Lambda(\boldsymbol{Y}) := \frac{\sup\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1} p(\boldsymbol{Y}|\boldsymbol{\theta})}{\sup\limits_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} p(\boldsymbol{Y}|\boldsymbol{\theta})} \tag{44}$$

and rejecting the null hypothesis when $\Lambda(\boldsymbol{Y}) > c_\alpha$ for some constant $c_\alpha > 0$ suitably chosen to provide a Type-I error probability of at most α. The following lemma recalls the classical likelihood ratio test construction of the F-test.

Theorem A.1.

$$\Lambda(\boldsymbol{Y}) = 1 + \frac{d}{n-p-d} f(\boldsymbol{Y}) \tag{45}$$

where the f-statistic is defined as

$$f(\boldsymbol{Y}) = \frac{\frac{\boldsymbol{Y}'(\boldsymbol{P_W}-\boldsymbol{P_X})\boldsymbol{Y}}{d}}{\frac{\boldsymbol{Y}'(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}}{n-p-d}} = \frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{ds^2(\boldsymbol{Y})} = \frac{\boldsymbol{t}(\boldsymbol{Y})'\boldsymbol{t}(\boldsymbol{Y})}{d} \tag{46}$$

Then $\Lambda(\boldsymbol{Y}) > c_\alpha \iff f(\boldsymbol{Y}) > f_\alpha$ for some $f_\alpha > 0$. The distributions of the f-statistic under H_1 and H_0 are

$$\begin{aligned} f(\boldsymbol{Y})|\boldsymbol{\beta},\boldsymbol{\delta},\sigma^2, H_1 &\sim F(d, n-p-d, \|\tilde{\boldsymbol{Z}}\boldsymbol{\delta}\|_2^2/\sigma^2) \\ f(\boldsymbol{Y})|\boldsymbol{\beta},\sigma^2, H_0 &\sim F(d, n-p-d, 0) \end{aligned} \tag{47}$$

Rejecting when $f(Y) > f_\alpha$, with f_α denoting the $1-\alpha$ quantile $F(d, n-p-d, 0)$ yields a fixed-n test with Type-I error probability α.

Note that the f statistic can be written in terms of the maximal invariant statistic $\boldsymbol{t}(\boldsymbol{Y})$. In the case of $d = 1$, when there is only a single main effect, then f can be identified as the square of the usual t-statistic ($t \sim t_{n-p-1} \Rightarrow t^2 \sim F(1, n-p-1)$). A p-value can be calculated by computing $\mathbb{P}[f \geq f(Y)]$ under the null $F(d, n-p-d, 0)$ distribution.

Proof. Starting with the denominator in (44), consider first expressing the quadratic form in the Guassian likelihood as a component in $\mathcal{C}(\boldsymbol{X})$ and a component in $\mathcal{C}(\boldsymbol{X})^\perp$.

$$\begin{aligned} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 &= \|\boldsymbol{P_X}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P_X})(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\|_2^2 \\ &= \|\boldsymbol{P_X}\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}\|_2^2 \end{aligned} \tag{48}$$

This is minimized by setting $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})\boldsymbol{X}'\boldsymbol{Y}$, which sets the first term to zero. The likelihood is then maximized by setting $\hat{\sigma^2} = \|(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}\|_2^2/n$. It follows that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} p(\boldsymbol{Y}|\boldsymbol{\theta}) = \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\|(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}\|_2^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}} \tag{49}$$

Now consider the numerator in (44), expressing the quadratic form in the Gaussian likelihood as a component in $\mathcal{C}(\boldsymbol{W})$ a component in $\mathcal{C}(\boldsymbol{W})^{\perp}$.

$$\begin{aligned}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\delta}\|_2^2 = \|\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{\gamma}\|_2^2 &= \|\boldsymbol{P_W}(\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})(\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{\gamma})\|_2^2 \\ &= \|\boldsymbol{P_W}(\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2\end{aligned} \tag{50}$$

where $\boldsymbol{W} = [\boldsymbol{X}, \boldsymbol{Z}]$ and $\boldsymbol{\gamma}' = (\boldsymbol{\beta}', \boldsymbol{\delta}')$. Applying the same reasoning as before, this is minimized by setting $\hat{\gamma} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'Y$, which sets the first term to zero. The likelihood is then maximized by setting $\hat{\sigma}^2 = \|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2/n$. It follows that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_1} p(\boldsymbol{Y}|\boldsymbol{\theta}) = \left(\frac{n}{2\pi}\right)^{\frac{n}{2}} \left(\frac{1}{\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2}\right)^{\frac{n}{2}} e^{-\frac{n}{2}} \tag{51}$$

and therefore

$$\Lambda(\boldsymbol{Y}) = \left(\frac{\|(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2}{\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2}\right)^{\frac{n}{2}}. \tag{52}$$

However, the vector in the numerator be expressed as a component in $\mathcal{C}(\boldsymbol{W})$ and a component in $\mathcal{C}(\boldsymbol{W})^{\perp}$.

$$\begin{aligned}\|(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 &= \|\boldsymbol{P_W}(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})(\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 \\ &= \|(\boldsymbol{P_W}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2 \\ &= \|(\boldsymbol{P_W}-\boldsymbol{P_X})\boldsymbol{Y}\|_2^2 + \|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2\end{aligned} \tag{53}$$

and so the likelihood ratio can be written in terms of the f-statistic as

$$\Lambda(\boldsymbol{Y}) = 1 + \frac{d}{n-p-d} f(\boldsymbol{Y}). \tag{54}$$

To show $f(\boldsymbol{Y})$ can be expressed in terms of $\hat{\boldsymbol{\delta}}(\boldsymbol{Y})$ as in equation (46), note simply that $(\boldsymbol{P_W}-\boldsymbol{P_X})\boldsymbol{Y} = (\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{P_W}\boldsymbol{Y} = (\boldsymbol{I}-\boldsymbol{P_X})(\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{Y}) + \boldsymbol{Z}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})) = (\boldsymbol{I}-\boldsymbol{P_X})\boldsymbol{Z}\hat{\boldsymbol{\delta}}(\boldsymbol{Y}) = \tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})$. □

A test of the null hypothesis $H_0 : \boldsymbol{\delta} = \boldsymbol{\delta}_0$ can easily be obtained from a hypothesis test of $\boldsymbol{\delta} = \boldsymbol{0}$ by replacing $\boldsymbol{Y}$ with $\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{\delta}_0$. In this case, the f- statistic becomes

$$f(\boldsymbol{Y};\boldsymbol{\delta}_0) = \frac{\frac{(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)'(\boldsymbol{P_W}-\boldsymbol{P_X})(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)}{d}}{\frac{(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)'(\boldsymbol{I}-\boldsymbol{P_W})(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{\delta}_0)}{n-p-d}} = \frac{(\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta}_0)'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta}_0)}{ds^2(\boldsymbol{Y})} \tag{55}$$

By finding the set of null-values that would not be rejected by this test one obtains a confidence set for the vector $\boldsymbol{\delta}$.

Corollary A.2. *A $1-\alpha$ confidence set for $\boldsymbol{\delta}$ is provided by*

$$\mathcal{C}_\alpha(\boldsymbol{Y}) := \{\boldsymbol{\delta} : (\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}(\boldsymbol{Y})-\boldsymbol{\delta}) \leq ds^2(\boldsymbol{Y})f_\alpha\}, \tag{56}$$

A.2 Proof of Lemma 5.2 - Invariant test functions and maximal invariant test statistics

($\Rightarrow$) Assume $\lambda(\boldsymbol{Y}) = h(M(\boldsymbol{Y}))$ where M is a maximal invariant. For all $g \in G$, $\lambda(g(\boldsymbol{Y})) = h(M(g(\boldsymbol{Y})) = h(M(\boldsymbol{Y})) = \lambda(\boldsymbol{Y})$ and therefore ϕ is an invariant function.
($\Leftarrow$) Assume ϕ is invariant, then ϕ is a constant on orbits. The maximal invariant is also constant on orbits and takes a unique value on each orbit, by definition, hence their exists a surjective function that maps the values taken by the maximal invariant on orbits to the values taken by ϕ on orbits.

A.3 Proof of Proposition 5.3 - $t(Y)$ is a maximal invariant

We proceed by proving the contrapositive, namely, if $\boldsymbol{Y}_1 \neq g(\boldsymbol{Y}_2)$ for any $g \in G$ then $\boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)$. Write

$$\boldsymbol{Y}_1 = \boldsymbol{P_X}\boldsymbol{Y}_1 + (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}_1$$
$$\boldsymbol{Y}_2 = \boldsymbol{P_X}\boldsymbol{Y}_2 + (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}_2$$

If $\boldsymbol{Y}_1 \neq g(\boldsymbol{Y}_2)$ for any $g \in G$ then we know

1. $\boldsymbol{Y}_1 \neq c\boldsymbol{Y}_2$ for any $c \in \mathbb{R}$
2. $(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}_1 \neq (\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}_2$

The latter must be true because if both vectors only differed by their component in $\mathcal{C}(\boldsymbol{X})$, then one could easily be expressed in terms of the other plus an appopriate term $\boldsymbol{X}\boldsymbol{\alpha}^\star$ for some $\boldsymbol{\alpha}^\star$. It must be the components in $\mathcal{C}(\boldsymbol{X})^\perp$ that are different. Let's take the component of each vector in $\mathcal{C}(\boldsymbol{X})^\perp$ and further decompose it into a component in $\mathcal{C}(\boldsymbol{W})$ and a component in $\mathcal{C}(\boldsymbol{W})^\perp$,

$$(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}_i = (\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_i + (\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_i, \tag{57}$$

for $i \in \{1, 2\}$. There are now three cases to consider

Case 1: $(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_1 \neq (\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_2$ and $(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_1 = (\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_2$.
Clearly $s^2(\boldsymbol{Y}_1) = s^2(\boldsymbol{Y}_2)$, but $\boldsymbol{Y}_1'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_1 \neq \boldsymbol{Y}_2'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_2$, which implies $\boldsymbol{t}(\boldsymbol{Y}_1)'\boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)'\boldsymbol{t}(\boldsymbol{Y}_2)$ which implies $\boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)$.

Case 2: $(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_1 = (\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_2$ and $(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_1 \neq (\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_2$.
Clearly $s^2(\boldsymbol{Y}_1) \neq s^2(\boldsymbol{Y}_2)$, which implies $\boldsymbol{t}(\boldsymbol{Y}_1)'\boldsymbol{t}(\boldsymbol{Y}_1) = s^2(\boldsymbol{Y}_2)\boldsymbol{t}(\boldsymbol{Y}_2)'\boldsymbol{t}(\boldsymbol{Y}_2)/s^2(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)'\boldsymbol{t}(\boldsymbol{Y}_2) \Rightarrow \boldsymbol{t}(\boldsymbol{Y}_1) \neq \boldsymbol{t}(\boldsymbol{Y}_2)$

Case 3: $(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_1 \neq (\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}_2$ and $(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_1 \neq (\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_2$.
Clearly $s^2(\boldsymbol{Y}_1) \neq s^2(\boldsymbol{Y}_2)$. Proof by contradiction. If $\boldsymbol{t}(\boldsymbol{Y}_1) = \boldsymbol{t}(\boldsymbol{Y}_2)$ then

$$\boldsymbol{Y}_1'(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_1 = \frac{s^2(\boldsymbol{Y}_1)}{s^2(\boldsymbol{Y}_2)}\boldsymbol{Y}_2(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_2$$

which would imply $(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_1 = (s(\boldsymbol{Y}_1)/s(\boldsymbol{Y}_2))(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}_2$, but this is a contradiction because $\boldsymbol{Y}_1 \neq c\boldsymbol{Y}_2$ for any c.

A.4 Proof of Theorem 5.4 - Likelihood ratios of sequences of maximal invariants

We first require a simple lemma

Lemma A.3. *Let* $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ *the maximal invariant statistic defined in equation* (36)*. Then for* $i <= n$ $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ *can be written as a function of* $\boldsymbol{t}_n(\boldsymbol{Y}_n)$*.*

Proof. Knowledge of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ implies knowledge of $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ for all $i < n$ also. To see this note that $\boldsymbol{Y}_i = \boldsymbol{P}_{in}\boldsymbol{Y}_n$ where $\boldsymbol{P}_{in}$ is the projection from $\mathbb{R}^n$ to $\mathbb{R}^i$ obtained by retaining only the first i elements of the vector $\boldsymbol{Y}_n$. Then $\boldsymbol{t}_i(\boldsymbol{Y}_i) = \boldsymbol{t}_i(\boldsymbol{P}_{in}\boldsymbol{Y}_n)$, which we write $\boldsymbol{t}_i(\boldsymbol{Y}_i) = \boldsymbol{u}_{in}(\boldsymbol{Y}_n)$. Each function $\boldsymbol{u}_{in}$ is i) a function of $\boldsymbol{Y}_n$ that is also ii) invariant under transformations $\boldsymbol{Y}_n \to c\boldsymbol{Y}_n + \boldsymbol{X}_n\boldsymbol{\alpha}$. It follows from lemma 5.2 that each $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ can be written as a function of the maximal invariant $\boldsymbol{t}_n(\boldsymbol{Y}_n)$. □

We can now state the proof of theorem 5.4

Proof. Lemma A.3 implies that each $\boldsymbol{t}_i(\boldsymbol{Y}_i)$ can be written as functions of $\boldsymbol{t}_n(\boldsymbol{Y}_n)$ for $i \leq n$. This implies $p(\boldsymbol{t}_1(\boldsymbol{Y}_1), \ldots, \boldsymbol{t}_n(\boldsymbol{Y}_n)|H_i) = p(\boldsymbol{t}_n(\boldsymbol{Y}_n)|H_i)$, □

A.5 Proof of Theorem 5.5 - $B(Y_n)$ as a ratio of multivariate t densities

Starting with the model under H_0

$$\begin{aligned} \boldsymbol{t}(\boldsymbol{Y}) = (\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y} &= \tilde{Z}\hat{\boldsymbol{\delta}}(\boldsymbol{Y}) \\ \boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_0 &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I) \\ \Rightarrow \boldsymbol{V}_d'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_0 &\sim N_d(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_d) \\ \Rightarrow \boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_0 &\sim t_{n-p-d}(\boldsymbol{0}, \boldsymbol{I}_d) \end{aligned}$$

Therefore the density is

$$p(\boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_0) = \frac{\Gamma(\frac{n-p}{2})}{\Gamma(\frac{n-p-d}{2})} \frac{1}{(n-p-d)^{\frac{d}{2}} \pi^{\frac{d}{2}}} \left(1 + \frac{\boldsymbol{t}(\boldsymbol{Y})' \boldsymbol{t}(\boldsymbol{Y})}{n-p-d}\right)^{\frac{n-p}{2}} \tag{58}$$

Now considering the model under H_1

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_1 &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2(I + \boldsymbol{Z}\Phi^{-1}\boldsymbol{Z})) \\ \Rightarrow \boldsymbol{V}_d'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_1 &\sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{V}_d'(I + \boldsymbol{Z}\Phi^{-1}\boldsymbol{Z}')\boldsymbol{V}_d) \\ \Rightarrow \boldsymbol{V}_d'(\boldsymbol{P_W} - \boldsymbol{P_X})\boldsymbol{Y}|\boldsymbol{\beta}, \sigma^2, H_1 &\sim N(\boldsymbol{0}, \sigma^2(\boldsymbol{V}_d'\boldsymbol{V}_d + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}}')) \\ \Rightarrow \boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_1 &\sim t_{n-p-d}(\boldsymbol{0}, (\boldsymbol{I}_d + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}}')), \end{aligned}$$

where the last line follows from $\boldsymbol{I}_d = \boldsymbol{V}_d'\boldsymbol{V}_d$. From the Sherman-Morrison-Woodbury Identity

$$(\boldsymbol{I}_d + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}})^{-1} = \boldsymbol{I}_d - \tilde{\boldsymbol{Z}}(\Phi + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'.$$

By the matrix determinant lemma

$$\frac{1}{\det(\boldsymbol{I} + \tilde{\boldsymbol{Z}}\Phi^{-1}\tilde{\boldsymbol{Z}}')} = \frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})}$$

Therefore the density is

$$\begin{aligned} p(\boldsymbol{t}(\boldsymbol{Y})|\boldsymbol{\beta}, \sigma^2, H_0) = & \frac{\Gamma(\frac{n-p}{2})}{\Gamma(\frac{n-p-d}{2})} \frac{1}{(n-p-d)^{\frac{d}{2}} \pi^{\frac{d}{2}}} \sqrt{\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})}} \\ & \left(1 + \frac{\boldsymbol{t}(\boldsymbol{Y})(I - \tilde{\boldsymbol{Z}}(\Phi + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}')\boldsymbol{t}(\boldsymbol{Y})}{n-p-d}\right)^{\frac{n-p}{2}} \end{aligned} \tag{59}$$

A.6 Bayes Factor and Haar Mixture Martingale Interpretation of $B_n(\boldsymbol{Y}_n)$

Proof of Theorem 5.6.

Proof. First decompose the quadratic form in the likelihood into two components, one in $\mathcal{C}(\boldsymbol{W})$ and the other in $\mathcal{C}(\boldsymbol{W})^{\perp}$. Then, further subdivide the component in $\mathcal{C}(\boldsymbol{W})$ into two subcomponents, one in $\mathcal{C}(\boldsymbol{X})$ and $\mathcal{C}(\boldsymbol{X})^{\perp}$. This helps to isolate terms in $\boldsymbol{\beta}, \boldsymbol{\delta}$ and σ^2 to make computing the marginals easier.

$$\begin{aligned}
\|\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma}\|_2^2 &= \|\boldsymbol{P_W}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P_W})(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 \\
&= \|\boldsymbol{P_W}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}\|_2^2 \\
&= \|\boldsymbol{P_X}\boldsymbol{P_W}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(I - \boldsymbol{P_X})\boldsymbol{P_W}(\boldsymbol{Y} - \boldsymbol{W}\boldsymbol{\gamma})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}\|_2^2 \qquad (60) \\
&= \|X\hat{\boldsymbol{\beta}} + \boldsymbol{P_X}\boldsymbol{Z}\hat{\boldsymbol{\delta}} - X\boldsymbol{\beta} - \boldsymbol{P_X}\boldsymbol{Z}\boldsymbol{\delta})\|_2^2 + \|(I - \boldsymbol{P_X})\boldsymbol{Z}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}\|_2^2 \\
&= \|\boldsymbol{X}(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}(\boldsymbol{Y}, \boldsymbol{\delta}))\|_2^2 + \|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{P_W})\boldsymbol{Y}\|_2^2
\end{aligned}$$

where $\tilde{\boldsymbol{\beta}}(\boldsymbol{Y}, \boldsymbol{\delta})) = \hat{\boldsymbol{\beta}} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Z}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta}))$.

Step 1) Compute $p(\boldsymbol{Y}|H_1)$
Let's proceed first by computing the marginal under H_1

$$p(\boldsymbol{Y}|H_1) = \int\int\int p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, H_1)p(\boldsymbol{\delta}|\sigma^2, H_1)p(\boldsymbol{\beta}, \sigma^2|H_1)d\boldsymbol{\beta}d\boldsymbol{\delta}d\sigma^2. \qquad (61)$$

We can handle these three marginalizations in three consecutive steps.

Step 1)i) Compute $p(\boldsymbol{Y}|\boldsymbol{\delta}, \sigma^2, H_1)$
Handling the marginalization for $\boldsymbol{\beta}$ first gives

$$\begin{aligned}
p(\boldsymbol{Y}|\boldsymbol{\delta}, \sigma^2, H_1) &= \int p(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\delta}, \sigma^2, H_1)p(\boldsymbol{\beta}|H_1)d\beta \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\left(\|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})\|_2^2+\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2\right)} \int e^{-\frac{1}{2\sigma^2}\|\boldsymbol{X}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}(\boldsymbol{Y},\boldsymbol{\delta}))\|_2} d\boldsymbol{\beta} \qquad (62) \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})\|_2^2} e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2},
\end{aligned}$$

where the last line follows from recognizing the integrand as the kernel of a multivariate Gaussian in $\boldsymbol{\beta}$ with precision matrix $\boldsymbol{X}'\boldsymbol{X}/\sigma^2$.

Step 1)ii) Compute $p(\boldsymbol{Y}|\sigma^2, H_1)$
We now move onto performing the marginalization with respect to $\boldsymbol{\delta} \sim N(\boldsymbol{\delta}_0, \sigma^2\boldsymbol{\Phi}^{-1})$. Before doing this we complete the square in the following sense

$$\|\tilde{\boldsymbol{Z}}(\hat{\boldsymbol{\delta}} - \boldsymbol{\delta})\|_2^2 + \boldsymbol{\delta}'\boldsymbol{\Phi}\boldsymbol{\delta} =(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}})'(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})(\boldsymbol{\delta} - \tilde{\boldsymbol{\delta}}) + \hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}} \qquad (63)$$

where $\tilde{\boldsymbol{\delta}} = (\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\delta}$ is the posterior mean and $(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})/\sigma^2$ the posterior precision. Performing the marginalization then yields

$$\begin{aligned} p(\boldsymbol{Y}|\sigma^2, H_1) &= \int p(\boldsymbol{Y}|\boldsymbol{\delta}, \sigma^2, H_1)p(\boldsymbol{\delta}|\sigma^2, H_1)d\boldsymbol{\delta} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2} e^{-\frac{1}{2\sigma^2}\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}} \\ &\quad \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{d}{2}} \det(\boldsymbol{\Phi})^{\frac{1}{2}} \int e^{-\frac{1}{2\sigma^2}(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})'(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})(\boldsymbol{\delta}-\tilde{\boldsymbol{\delta}})} d\boldsymbol{\delta} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \left(\frac{\det(\boldsymbol{\Phi})}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})}\right)^{\frac{1}{2}} \\ &\quad e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2} e^{-\frac{1}{2\sigma^2}\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}} \end{aligned} \tag{64}$$

where the last step is achieved by recognizing the kernel of a d-dimensional multivariate Gaussian density in $\boldsymbol{\delta}$ with precision $\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}$.

Step 1)iii) Compute $p(\boldsymbol{Y}|H_1)$

Now we perform the final marginalization over σ^2

$$\begin{aligned} p(\boldsymbol{Y}|H_1) &= \int p(\boldsymbol{Y}|\sigma^2, H_1)p(\sigma^2|H_1)d\sigma^2 \\ &= \left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}} \\ &\quad \int \left(\frac{1}{\sigma^2}\right)^{\frac{n-p}{2}+1} e^{-\frac{1}{2\sigma^2}\left(\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2+\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}\right)} d\sigma^2 \\ &= \left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}} \\ &\quad \Gamma\left(\frac{n-p}{2}\right) \left(\frac{\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2}{2} + \frac{\hat{\boldsymbol{\delta}}'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}}{2}\right)^{-\frac{n-p}{2}} \end{aligned} \tag{65}$$

where the last line follows from recognizing the kernel of an Inverse Gamma. Tidying the expression up yields

$$\begin{aligned} p(\boldsymbol{Y}|H_1) = &\left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}} \Gamma\left(\frac{n-p}{2}\right) \\ &\left(\frac{s^2(\boldsymbol{Y})}{2}\right)^{-\frac{n-p}{2}} (n-p-d)^{-\frac{n-p}{2}} \left(1 + \frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}-\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi}+\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}} \end{aligned} \tag{66}$$

The derivation for $p(\boldsymbol{Y}|H_0)$ proceeds similarly as $p(\boldsymbol{Y}|H_1)$, except for the marginalization over $\boldsymbol{\delta}$. Performing the marginalization with respect to $\boldsymbol{\beta}$ first yields

$$\begin{aligned} p(\boldsymbol{Y}|\sigma^2, H_0) &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\hat{\boldsymbol{\delta}}'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}} e^{-\frac{1}{2\sigma^2}\|(\boldsymbol{I}-\boldsymbol{P_W})\boldsymbol{Y}\|_2^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} e^{-\frac{s^2(\boldsymbol{Y})(n-p-d)}{2\sigma^2}\left(1+\frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)} \end{aligned} \tag{67}$$

Performing the marginalization finally with respect to σ^2 yields

$$\begin{aligned} p(\boldsymbol{Y}|H_0) = &\left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} \left(\frac{1}{\det(\boldsymbol{X}'\boldsymbol{X})}\right)^{\frac{1}{2}} \Gamma\left(\frac{n-p}{2}\right) \\ &\left(\frac{s^2(\boldsymbol{Y})}{2}\right)^{-\frac{n-p}{2}} (n-p-d)^{-\frac{n-p}{2}} \left(1 + \frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}} \end{aligned} \tag{68}$$

The Bayes factor (or ”likelihood ratio mixture”) is given by

$$\frac{p(\boldsymbol{Y}|H_1)}{p(\boldsymbol{Y}|H_0)} = \frac{\det(\boldsymbol{\Phi})^{\frac{1}{2}}}{\det(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{\frac{1}{2}}} \frac{\left(1 + \frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'(\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}} - \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}(\boldsymbol{\Phi} + \tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})^{-1}\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}})\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}}}{\left(1 + \frac{\hat{\boldsymbol{\delta}}(\boldsymbol{Y})'\tilde{\boldsymbol{Z}}'\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\delta}}(\boldsymbol{Y})}{s^2(\boldsymbol{Y})(n-p-d)}\right)^{-\frac{n-p}{2}}} \tag{69}$$

□

A.7 Proof of Theorem 3.4 - Asymptotic sequential t-tests

The first step requires the following lemma

Lemma A.4.

$$-\frac{\nu+1}{2}\log\left(1+\frac{t^2}{\nu}\right)=-\frac{1}{2}t^2+o(\frac{1}{\nu^{1-\epsilon}}) \tag{70}$$

for $0<\epsilon<1$ *as* $\nu\to\infty$

Proof. The Maclaurin series for the logarithmic part is

$$\log\left(1+\frac{t^2}{\nu}\right)=\sum_{n=1}^{\infty}(-1)^{n+1}\frac{1}{n}\left(\frac{t^2}{\nu}\right)^n=\frac{t^2}{\nu}+\sum_{n=2}^{\infty}(-1)^{n+1}\frac{1}{n}\left(\frac{t^2}{\nu}\right)^n.$$

Note this only holds when $t^2/\nu<1$, otherwise the series diverges, but this is trivially satisfied for large ν. For a large enough ν

$$\begin{aligned}-\frac{\nu+1}{2}\log\left(1+\frac{t^2}{\nu}\right)&=-\frac{1}{2}\frac{\nu+1}{\nu}t^2-\frac{1}{2}\sum_{n=2}^{\infty}(-1)^{n+1}\frac{v+1}{n}\left(\frac{t^2}{\nu}\right)^n\\&=-\frac{1}{2}t^2-\frac{1}{2\nu}t^2-\frac{1}{2}\sum_{n=2}^{\infty}(-1)^{n+1}\frac{1}{n}\left(\frac{\nu+1}{\nu^n}\right)t^{2n}\\&=-\frac{1}{2}t^2+r_v\end{aligned}$$

where $r_v=o\left(\frac{1}{\nu^{1-\epsilon}}\right)$. Readers may be familiar with this result from the observation that a t_ν density looks increasingly like a Gaussian as $\nu\to\infty$. □

Lemma A.5. *(Strong consistency of OLS estimator) Suppose*

1. $\boldsymbol{Y}_i=\boldsymbol{w}_i'\boldsymbol{\gamma}+\varepsilon_i\quad i=1,2,\dots,$
2. $\boldsymbol{\varepsilon}_n\overset{i.i.d.}{\sim}F[0,\sigma^2]$*, where* $F[0,\sigma^2]$ *denotes a distribution with mean zero and variance* $0<\sigma^2<\infty$.
3. $\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n/n\overset{a.s.}{\to}\mathbf{0}$,
4. $\boldsymbol{W}_n'\boldsymbol{W}_n/n\overset{a.s.}{\to}\boldsymbol{A}$, $\boldsymbol{A}$ *finite and positive definite,*

then $\hat{\gamma}_n=(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\boldsymbol{W}_n'\boldsymbol{Y}_n\overset{a.s.}{\to}\gamma$*, and* $\hat{\sigma}_n^2\overset{a.s.}{\to}\sigma^2$.

Proof. Because $\boldsymbol{A}$ is positive definite, $(\boldsymbol{W}_n'\boldsymbol{W}_n/n)^{-1}$ exists almost surely for all sufficiently large n. By definition of the estimator $\hat{\boldsymbol{\gamma}}_n=\boldsymbol{\gamma}+(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon}$. The continuous mapping theorem, combined with the third and fourth assumptions, $\hat{\boldsymbol{\gamma}}_n=\boldsymbol{\gamma}+(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon}\overset{a.s.}{\to}\boldsymbol{\gamma}+\boldsymbol{A}^{-1}\mathbf{0}=\boldsymbol{\gamma}$. For strong consistency of σ^2, $\hat{\sigma}_n^2=(\boldsymbol{\varepsilon}_n'\boldsymbol{\varepsilon}_n-\boldsymbol{\varepsilon}_n'\boldsymbol{W}_n(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n)/(n-p-d)$. From the third and fourth assumptions, combined with the continuous mapping theorem, $\boldsymbol{\varepsilon}_n'\boldsymbol{W}_n(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}\boldsymbol{W}_n'\boldsymbol{\varepsilon}_n/(n-p-d)\overset{a.s.}{\to}\mathbf{0}$ while $\boldsymbol{\varepsilon}_n'\boldsymbol{\varepsilon}_n/(n-p-d)\overset{a.s.}{\to}\sigma^2$ by the strong law. □

Lemma A.6. *Under the null hypothesis and assumptions of lemma A.5*

$$\frac{t_n(\boldsymbol{Y}_n)}{\sqrt{n}}=\frac{\|\tilde{\boldsymbol{Z}}_n\|_2}{\sqrt{n}}\frac{\hat{\delta}_n(\boldsymbol{Y}_n)-\delta_0}{\hat{\sigma}_n(\boldsymbol{Y}_n)}=o_{a.s.}(1),$$

where the almost sure statement is with respect to the measure under the null.

Proof. Note that although $\boldsymbol{Z}_n$ is stochastic, $\sqrt{n} = \|\mathbf{1}_n\|_2 \geq \|\boldsymbol{Z}_n\| = \|\boldsymbol{P}_{\boldsymbol{X}_n}\boldsymbol{Z}_n\|_2 + \|(I - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n\|_2 \geq \|(I - \boldsymbol{P}_{\boldsymbol{X}_n})\boldsymbol{Z}_n\|_2 = \|\tilde{\boldsymbol{Z}}_n\|_2$ and so the first term $\|\tilde{\boldsymbol{Z}}_n\|_2/\sqrt{n} \leq 1$. For the second term, by lemma A.5 $\hat{\delta}_n$ and $\hat{\sigma}_n^2$ are strongly consistent, and under the null hypothesis $\hat{\delta}_n \overset{a.s.}{\to} \delta_0$. It follows that the second term converges almost surely to zero. □

We can now provide the proof for theorem 3.4.

Proof. To simplify notation let $a_n = \phi/(\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2)$ and simply write t_n instead of $t_n(\mathbf{Y}_n; \delta_0)$. Note that $a_n \leq \phi/(\phi + n)$ for all n.

$$\begin{aligned}|\log B_n(\mathbf{Y}_n;\delta_0) - \log \tilde{B}_n(\mathbf{Y}_n;\delta_0)| = &-\frac{n-p}{2}\log\left(1 + \frac{a_n t_n^2}{(n-p-1)}\right)\\ &+ \frac{n-p}{2}\log\left(1 + \frac{t_n^2}{(n-p-1)}\right)\\ &+ \frac{1}{2}(a_n - 1)t_n^2.\end{aligned} \tag{71}$$

By lemma A.6, $t_n^2/(n-p-1) < 1$ almost surely for all sufficiently large n, which allows the logarithms to be expressed in terms of their series expansion in lemma A.4

$$\begin{aligned}|\log B_n(\mathbf{Y}_n;\delta_0) - \log \tilde{B}_n(\mathbf{Y}_n;\delta_0)| &= -\frac{(a_n - 1)}{2(n-p-1)}t_n^2 - \frac{1}{2}\sum_{i=2}^{\infty}(-1)^{i+1}\frac{1}{i}\left(\frac{n-p}{(n-p-1)^i}\right)(a_n^i - 1)t_n^{2i}\\ &:= d_n(\mathbf{Y}_n)\end{aligned} \tag{72}$$

The difference term is $o_{a.s.}(1)$ by lemma A.6. It follows that

$$\begin{aligned}\alpha =&\mathbb{P}_{H_0}[\exists n \in \mathbb{N} : \log B_n(\mathbf{Y}_n;\delta_0) > -\log(\alpha)]\\ =&\mathbb{P}_{H_0}[\exists n \in \mathbb{N} : o(1) + \log \tilde{B}_n(\mathbf{Y}_n;\delta_0) > -\log(\alpha)]\end{aligned}$$

□

A.8 Proof of Corollary 3.5 - Asymptotic confidence sequences for regression coefficients

Proof. From theorem 3.4,

$$\begin{aligned}\alpha =&\mathbb{P}[\exists n \in \mathbb{N} : d_n(\mathbf{Y}_n) + \log \tilde{B}_n(\mathbf{Y}_n;\delta_0) > -\log(\alpha)]\\ =&\mathbb{P}\left[\exists n \in \mathbb{N} : t_n(\mathbf{Y}_n;\delta_0)^2 > \frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) - 2\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}d_n(\mathbf{Y}_n)\right]\\ =&\mathbb{P}\left[\exists n \in \mathbb{N} : (\hat{\delta}(\mathbf{Y}_n) - \delta_0)^2 > \frac{s_n(\mathbf{Y}_n)^2}{\|\tilde{\boldsymbol{Z}}_n\|^2}\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) - 2\left(\frac{s_n(\mathbf{Y}_n)^2}{\|\tilde{\boldsymbol{Z}}_n\|^2}\right)\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\right)d_n(\mathbf{Y}_n)\right].\end{aligned}$$

Applying the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ yields

$$1 - \alpha =\mathbb{P}\left[\forall n \in \mathbb{N} : |\hat{\delta}(\mathbf{Y}_n) - \delta_0| \leq r_n(\mathbf{Y}_n) + e_n(\mathbf{Y}_n)\right],$$

where $r_n(\mathbf{Y}_n)$ is defined as in equation (14) and

$$e_n(\mathbf{Y}_n) = \frac{s_n(\mathbf{Y}_n)}{\|\tilde{\boldsymbol{Z}}_n\|}\sqrt{2\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\right)|d_n(\mathbf{Y}_n)|} = o_{a.s.}(n^{-\frac{1}{2}})$$

The order of $e_n(\mathbf{Y}_n)$ follows because $d_n(\boldsymbol{Y}) = o_{a.s.}(1)$ and $\|\tilde{\boldsymbol{Z}}_n\|_2 = o_{a.s.}(\sqrt{n})$ □

A.9 Proofs for Section 4

The difference between the sample ols and the population ols can be seen to contain a sum, $\hat{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star = \boldsymbol{W}_n'\boldsymbol{W}_n \sum_{i=1}^n \boldsymbol{w}_i y_i - \boldsymbol{w}_i \gamma^\star$. Focusing on the summand, we notice that the terms are iid and zero mean.

Proposition A.7.

$$\mathbb{E}_{sp}[\boldsymbol{w}_i y_i - \boldsymbol{w}_i \boldsymbol{w}_i \boldsymbol{\gamma}^\star] = 0$$

Proof.

$$\mathbb{E}_{sp}[\boldsymbol{w}_i(y_i - \boldsymbol{w}_i'\boldsymbol{\gamma}^\star)] = \mathbb{E}_{sp}\left[\begin{pmatrix} 1 \\ \boldsymbol{x}_i - \boldsymbol{\mu}_x \\ z_i - \rho \end{pmatrix}(y_i - \alpha^\star - (z_i - \rho)\delta^\star - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta}^\star)\right]$$

Let's compute each element in turn, starting with the first

1.

$$\begin{aligned}
&\mathbb{E}_{sp}[y_i - \alpha^\star - (z_i - \rho)\delta^\star - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta}^\star)] \\
&= \mathbb{E}_{sp}[y_i] - \alpha^\star - 0 - 0 \\
&= \rho\mathbb{E}_{sp}[y_i^{(1)}] + (1-\rho)\mathbb{E}_{sp}[y_i^{(0)}] - \alpha^\star = 0
\end{aligned}$$

Line 2 follows from $\mathbb{E}_{sp}[z_i - \rho] = 0$, $\mathbb{E}_{sp}[\boldsymbol{x}_i - \mu_x] = 0$, and the final line from the definition of $\alpha^\star$.

2.

$$\begin{aligned}
&\mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)(y_i - \alpha^\star - (z_i - \rho)\delta^\star - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta}^\star)] \\
&= \mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)y_i] - 0 - 0 - \mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)(\boldsymbol{x}_i - \boldsymbol{\mu}_x)']\boldsymbol{\beta}^\star \\
&= 0
\end{aligned}$$

The first equality colds because $\mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)\alpha^\star] = 0$ and $\mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)(z_i - \rho)] = 0$ by independence of z_i and $\boldsymbol{x}_i$. The last equality holds from the definition of $\boldsymbol{\beta}^\star$

3.

$$\begin{aligned}
&\mathbb{E}_{sp}[(z_i - \rho)(y_i - \alpha^\star - (z_i - \rho)\delta^\star - (\boldsymbol{x}_i - \boldsymbol{\mu}_x)'\boldsymbol{\beta}^\star)] \\
&= \mathbb{E}_{sp}[(z_i - \rho)y_i] - 0 - \mathbb{E}_{sp}[(z_i - \rho)^2]\delta^\star - 0 \\
&= \rho(1-\rho)\mathbb{E}_{sp}[y_i^{(1)}] + \rho(1-\rho)\mathbb{E}_{sp}[y_i^{(0)}] - \rho(1-\rho)\delta^\star \\
&= 0
\end{aligned}$$

The first line follows because $\mathbb{E}_{sp}[(z_i - \rho)\alpha^\star] = 0$ and $\mathbb{E}_{sp}[(\boldsymbol{x}_i - \boldsymbol{\mu}_x)(z_i - \rho)] = 0$ by independence of z_i and $\boldsymbol{x}_i$. The last line follows from the definition of $\delta^\star$.

□

This proposition establishes that $\sum_i \boldsymbol{w}_i y_i - \boldsymbol{w}_i \boldsymbol{w}_i' \gamma^\star$ is a sum of iid random variables with zero mean and covariance given by $\boldsymbol{\Delta}$ in equation (28). Almost sure approximations of sums of random variables by sums of Gaussian random variables are known as *strong approximation theorems* or *strong invariance principles* (Strassen, 1964, 1967). We leverage the following multivariate strong approximation due to Morrow and Philipp (1982)

Lemma A.8. *Assume $\boldsymbol{w}_i' y_i$ has bounded second moments, then*

$$\sum_{i=1}^{n} \boldsymbol{w}_i y_i - \boldsymbol{w}_i \boldsymbol{w}_i' \boldsymbol{\gamma}^\star = \sum_{i=1}^{n} \boldsymbol{g}_i + \boldsymbol{r}_n \tag{73}$$

where $\boldsymbol{g}_i \sim N(0, \boldsymbol{\Delta})$, and $\boldsymbol{r}_n = o_{a.s.}(n^{3/8} \log n)$

For each n, left-multiplying by $(\boldsymbol{W}_n' \boldsymbol{W}_n)^{-1}$ yields

$$\hat{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star = (\boldsymbol{W}_n' \boldsymbol{W}_n)^{-1} \sum_{i=1}^{n} \boldsymbol{g}_i + (\boldsymbol{W}_n' \boldsymbol{W}_n)^{-1} \boldsymbol{r}_n \tag{74}$$

This is beginning to look like what we expect, except that the fixed-n asymptotics in equation (26) depend not on the sample information matrix $\boldsymbol{W}_n' \boldsymbol{W}_n$, but on it's expected value $n\boldsymbol{\Gamma}$. We can make this switch using the following lemma

Lemma A.9.

$$(\boldsymbol{W}_n' \boldsymbol{W}_n)^{-1} = \frac{1}{n} \boldsymbol{\Gamma}^{-1} + o_{a.s.}\left(\sqrt{\frac{2 \log \log n}{n^3}}\right) \tag{75}$$

Proof. $\boldsymbol{W}_n' \boldsymbol{W}_n - n\mathbb{E}_{sp}[\boldsymbol{w}_i \boldsymbol{w}_i'] = \sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i' - \boldsymbol{\Gamma}$ is a sum of zero mean random variables. Appealing elementwise to the law of iterated logarithm we have $\sum_{i=1}^n \boldsymbol{w}_i \boldsymbol{w}_i' = n\boldsymbol{\Gamma} + e_n$ where $e_n = o_{a.s.}(\sqrt{2n \log \log n})$. In particular this implies that $\|\tilde{\boldsymbol{Z}}_n\|_2^2 = n\rho(1-\rho) + o_{a.s.}(\sqrt{2n \log \log n})$. To pass this through the inverse we have from the Sherman-Morrison-Woodbury identity

$$(\boldsymbol{W}_n' \boldsymbol{W}_n)^{-1} = (n\boldsymbol{\Gamma} + e_n)^{-1} = \frac{1}{n}\boldsymbol{\Gamma}^{-1} + \frac{1}{n}\boldsymbol{\Gamma}^{-1}\left(\boldsymbol{e}_n^{-1} + \frac{1}{n}\boldsymbol{\Gamma}^{-1}\right)^{-1} \frac{1}{n}\boldsymbol{\Gamma}^{-1}$$

□

Combining equation 74 and lemma A.9 yields the following almost sure approximation

$$\hat{\gamma}_n^{ols} - \boldsymbol{\gamma}^\star = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{g}_i + o_{a.s.}\left(\frac{n^{3/8} \log n}{n}\right), \tag{76}$$

where $\boldsymbol{g}_i \sim N(0, \boldsymbol{\Gamma}^{-1} \boldsymbol{V} \boldsymbol{\Gamma}^{-1})$. If we inspect element corresponding to δ, i.e. left multiplying by $\boldsymbol{e}_1' = (0, \boldsymbol{0}, 1)'$, we obtain

$$\hat{\delta}_n^{ols} - \delta_{sp} = \frac{1}{n} \sum_{i=1}^{n} g_i + o_{a.s.}\left(\frac{n^{3/8} \log n}{n}\right), \tag{77}$$

where $\boldsymbol{g}_i \sim N(0, \sigma_\delta^2/(\rho(1-\rho))$. Compare this with equation (30). The fixed-n result in equation (30) gave the asymptotic distribution of $\hat{\delta}_n^{ols} - \delta_{sp}$ as a normal with mean 0 and variance $\sigma_\delta^2/n\rho(1-\rho)$. Our new result in equation (77) provides a strong approximation to $\hat{\delta}_n^{ols} - \delta_{sp}$ as a sum of n independent Gaussians with mean 0 and variance $\sigma_\delta^2/\rho(1-\rho)$. If we can provide a time-uniform bound to $\frac{1}{n}\sum_{i=1}^n g_i$, then we can get an approximate time uniform bound to $\hat{\delta}_n^{ols} - \delta_{sp}$ with approximation rate $o_{a.s.}\left(\frac{n^{3/8} \log n}{n}\right)$. This is established in the following lemma.

Lemma A.10. *Let $\{g_i\}_{i=1}^\infty$ be a sequence of iid mean 0 σ^2/c subGaussian random variables, then for any prespecified $\phi > 0$*

$$\mathbb{P}\left[\forall n \in \mathbb{N} : \frac{1}{n}|\sum_{i=1}^{n} g_i| \leq \sqrt{\frac{\sigma^2}{nc}} \sqrt{\frac{\phi + nc}{nc} \log\left(\frac{\phi + nc}{\phi \alpha^2}\right)}\right] \geq 1 - \alpha \tag{78}$$

Proof. Let

$$M_n(\mu) := e^{\frac{c}{\sigma^2}\left(\sum_{i=1}^n y_i\mu - \frac{1}{2}\mu^2\right)} = e^{\frac{c}{\sigma^2}\left(y_i\mu - \frac{1}{2}\mu^2\right)} M_{n-1}(\mu) \tag{79}$$

Taking the conditional expectation gives

$$\mathbb{E}[M_n(\mu)|\mathcal{F}_{n-1}] = \mathbb{E}\left[e^{\frac{c}{\sigma^2}y_i\mu}|\mathcal{F}_{n-1}\right] e^{-\frac{c}{2\sigma^2}\mu^2} M_{n-1}(\mu) \leq M_{n-1}(\mu) \tag{80}$$

where the last inequality follows from the definition of a σ^2/c-subGaussian random variable, with equality for $N(0, \sigma^2/c)$ random variables. This establishes that M_n is a nonnegative supermartingale. Mixtures of martingales remain mixtures, and so

$$M_n := \int M_n(\mu) dF^{\phi}(\mu) = \left(\frac{\phi}{\phi + nc}\right)^{\frac{1}{2}} e^{\frac{1}{2}\frac{nc}{\phi+nc}\left(\frac{\frac{1}{n}\sum y_i}{\sqrt{\frac{\sigma^2}{nc}}}\right)^2}, \tag{81}$$

where F^{ϕ} is a Gaussian measure with mean 0 and variance $\sigma^2\phi^{-1}$, is also a nonnegative supermartingale. From Ville's inequality for nonnegative supermartingales (Ville, 1939)

$$\mathbb{P}[\forall n \in \mathbb{N} : \log M_n \leq -\log \alpha] > 1 - \alpha.$$

Rearranging for $((1/n)\sum y_i)^2$ yields the desired result. □

Lemma A.10 in combination with the strong approximation in equation (77) yields the desired asymptotic confidence sequence for the superpopulation average treatment effect

$$\mathbb{P}\left[\forall n \in \mathbb{N} : |\hat{\delta}_n^{ols} - \delta_{sp}| \leq \sqrt{\frac{\sigma_\delta^2}{n\rho(1-\rho)}}\sqrt{\frac{\phi + n\rho(1-\rho)}{n\rho(1-\rho)}\log\left(\frac{\phi + n\rho(1-\rho)}{\phi\alpha^2}\right)} + r_n\right] \geq 1 - \alpha \tag{82}$$

where $r_n = o_{a.s.}\left(\frac{n^{3/8}\log n}{n}\right)$.

Equation (82) is trivial to implement. All you need the OLS estimate $\hat{\delta}_n^{ols}$, the sample size n, and the propensity score ρ. So what of the earlier confidence sequence presented in equation (14)? Could this have also been used to estimate δ_{sp}? Note that the only difference is $\|\tilde{\boldsymbol{Z}}_n\|_2^2$ is used in place of $n\rho(1-\rho)$, but what exactly is the difference? Recall that $1/\|\tilde{\boldsymbol{Z}}_n\|_2^2$ is simply the diagonal element of the $(\boldsymbol{W}_n'\boldsymbol{W}_n)^{-1}$ matrix corresponding to the variance of the $\hat{\delta}_n^{ols}$ estimator. In the limit, this is simply the diagonal element of $(n\boldsymbol{\Gamma})^{-1}$, which is equally to $1/n\rho(1-\rho)$. This follows more formally from lemma A.9

$$\frac{1}{n\rho(1-\rho)} = \frac{1}{\|\tilde{\boldsymbol{Z}}_n\|_2^2} + e_n,$$

where $e_n = o_{a.s.}\left(\frac{\sqrt{2\log\log n}}{n^3}\right)$. Also $n\rho(1-\rho) = \|\tilde{\boldsymbol{Z}}_n\|_2^2 + o_{a.s.}(\|\tilde{\boldsymbol{Z}}_n\|_2^2)$. Therefore

$$\begin{aligned}
&\sqrt{\frac{\sigma_\delta^2}{n\rho(1-\rho)}}\sqrt{\frac{\phi + n\rho(1-\rho)}{n\rho(1-\rho)}\log\left(\frac{\phi + n\rho(1-\rho)}{\phi\alpha^2}\right)} + o_{a.s.}(\sqrt{\log n/n}) \\
&= \sqrt{\frac{\sigma_\delta^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}(1+e_n)}\sqrt{\left(\frac{\phi}{\|\tilde{\boldsymbol{Z}}_2\|_2^2} + 1 + e_n\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2 + o_{a.s.}(\|\tilde{\boldsymbol{Z}}_n\|_2^2)}{\phi\alpha^2}\right)} + o_{a.s.}(\sqrt{\log n/n})
\end{aligned}$$

Beginning with the logarithmic component

$$\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2 + o_{a.s.}(\|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) = \log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\left(1 + o_{a.s.}(1)\right)\right)$$
$$= \log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) + o_{a.s.}(1),$$

because $\log(1 + o_{a.s.}(1)) = o_{a.s.}(1)$. Considering now the second square-root term

$$\sqrt{\left(\frac{\phi}{\|\tilde{\boldsymbol{Z}}_2\|_2^2} + 1 + e_n\right)\left(\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) + o_{a.s.}(1)\right)}$$
$$= \sqrt{\left(\frac{\phi}{\|\tilde{\boldsymbol{Z}}_2\|_2^2} + 1\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) + o_{a.s.}\left(\frac{1}{n}\right) + o_{a.s.}\left(\log n\frac{\sqrt{2\log\log n}}{n^3}\right) + o_{a.s.}\left(\frac{\sqrt{2\log\log n}}{n^3}\right)}$$
$$= \sqrt{\left(\frac{\phi}{\|\tilde{\boldsymbol{Z}}_2\|_2^2} + 1\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) + o_{a.s.}\left(\frac{1}{n}\right)}$$

Lastly, combining the two square root terms

$$= \sqrt{\frac{\sigma_\delta^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\left((1 + e_n)\left(\left(\frac{\phi}{\|\tilde{\boldsymbol{Z}}_2\|_2^2} + 1\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) + o_{a.s.}\left(\frac{1}{n}\right)\right)\right)}$$
$$= \sqrt{\frac{\sigma_\delta^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\left(\left(\frac{\phi}{\|\tilde{\boldsymbol{Z}}_2\|_2^2} + 1\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) + o_{a.s.}\left(\frac{\log n\sqrt{2\log\log n}}{n^3}\right) + o_{a.s.}\left(\frac{1}{n}\right) + o_{a.s.}\left(\frac{\sqrt{2\log\log n}}{n^4}\right)\right)}$$
$$= \sqrt{\frac{\sigma_\delta^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\left(\left(\frac{\phi}{\|\tilde{\boldsymbol{Z}}_2\|_2^2} + 1\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right) + o_{a.s.}\left(\frac{1}{n}\right)\right)}$$
$$= \sqrt{\frac{\sigma_\delta^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_2\|_2^2}{\|\tilde{\boldsymbol{Z}}_2\|_2^2}\right)\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right)} + o_{a.s.}\left(\frac{1}{n}\right)$$

Combining $o_{a.s.}\left(\frac{1}{n}\right)$ with the original $o_{a.s.}\left(\frac{n^{3/8}\log n}{n}\right)$ we see that the dominating order remains $o_{a.s.}\left(\frac{n^{3/8}\log n}{n}\right)$. This allows us to formally interchange $n\rho(1-\rho)$ and $\|\tilde{\boldsymbol{Z}}_n\|_2^2$.

So far, the variance term σ_δ^2 is considered known. Suppose we have a strongly consistent estimator, that is, $\hat{\sigma}_\delta^2 \to \sigma_\delta^2$ almost surely. We can write $\hat{\sigma}_\delta^2 = \sigma_\delta^2 + o_{a.s.}(\sigma_\delta^2)$, then

$$\sqrt{\frac{\sigma_\delta^2}{n\rho(1-\rho)}}h_n + r_n = \sqrt{\frac{\hat{\sigma}_\delta^2}{n\rho(1-\rho)}}h_n + \sqrt{\frac{o_{a.s.}(\sigma_\delta^2)}{n\rho(1-\rho)}}h_n + r_n, \tag{83}$$

where

$$h_n = \sqrt{\frac{\phi + n\rho(1-\rho)}{n\rho(1-\rho)}\log\left(\frac{\phi + n\rho(1-\rho)}{\phi\alpha^2}\right)} = O(\sqrt{\log n}).$$

It follows that the second term in equation (83) is $o_{a.s.}(\sqrt{\log n/n})$, which goes to zero slower than $o_{a.s.}\left(\frac{n^{3/8}\log n}{n}\right)$. Therefore the overall approximation rate is $o_{a.s}(\sqrt{\log n/n})$. Therefore the approximation goes to zero faster than the rate at which the confidence sequence width shrinks, yielding an asymptotic confidence sequence.