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

Michael Lindon, Dae Woong Ham, Martin Tingley, Iavor Bojinov

mlindon@netflix.com | daewoongham@g.harvard.edu | mtingley@netflix.com | ibojinov@hbs.edu

Introduction

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. 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.

Anytime Valid Inference

What is it? Anytime Valid inference replaces classical statistical guarantees that hold only at fixed time/sample-size with time-uniform guarantees that hold **for all** times/sample-sizes.
Confidence interval $\rightarrow$ **confidence sequence**

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

p-value $\rightarrow$ **sequential p-value**

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

These allow experiments to be continuously monitored without invalidating Type I error and coverage guarantees, stopped using data-dependent rules and prevent p-hacking.

Anytime Valid Inference in Linear Models

Assume the linear model $\boldsymbol{Y}_n = \boldsymbol{X}_n\boldsymbol{\beta} + \boldsymbol{Z}_n\boldsymbol{\delta} + \boldsymbol{\varepsilon}_n$. We combine $\boldsymbol{W}_n = [\boldsymbol{X}_n, \boldsymbol{Z}_n]$ and $\boldsymbol{\gamma}_n = [\boldsymbol{\beta}_n, \boldsymbol{\delta}_n]$ to write $\boldsymbol{Y}_n = \boldsymbol{W}_n\boldsymbol{\gamma} + \boldsymbol{\varepsilon}_n$. The coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ are nuisance parameters and we seek inference on $\boldsymbol{\delta} \in \mathbb{R}^d$.

Sequential p-values

For any fixed $\phi > 0$, A **sequential p-value** for $H_0 : \delta = \delta_0$

$$p_n(\boldsymbol{Y}_n; \delta_0) = \sqrt{\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi}} e^{-\frac{1}{2}\frac{\|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2} t_n(\boldsymbol{Y}_n;\delta_0)^2} \tag{3}$$

where

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

These are all available from the classical linear model analysis!
Anytime Valid inference requires nothing more, making it trivial to switch from classical to anytime valid inference.

Confidence Sequences

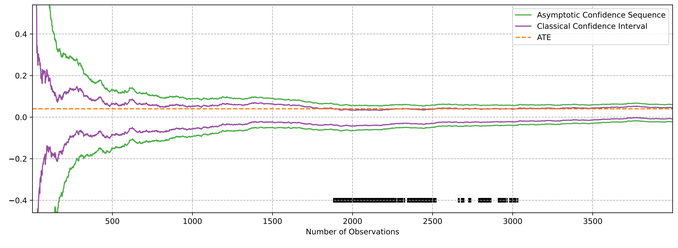
A $1 - \alpha$ **confidence sequence** for δ can be achieved by inverting the sequential test
$C_n(\boldsymbol{Y}_n) = \{\delta : \alpha \leq p_n(\boldsymbol{Y}_n; \delta)\} = (\hat{\delta}_n(\boldsymbol{Y}_n) - r_n(\boldsymbol{Y}_n), \hat{\delta}_n(\boldsymbol{Y}_n) + r_n(\boldsymbol{Y}_n))$ where

$$r_n(\boldsymbol{Y}_n) = \frac{s_n(\boldsymbol{Y}_n)}{\|\tilde{\boldsymbol{Z}}_n\|}\sqrt{\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\|\tilde{\boldsymbol{Z}}_n\|_2^2}\log\left(\frac{\phi + \|\tilde{\boldsymbol{Z}}_n\|_2^2}{\phi\alpha^2}\right)}. \tag{4}$$

Anytime Valid Inference in Randomized Experiments

Our results are neither limited to linear models nor Gaussian outcomes. Fitting a fully interacted linear model is equivalent to the regression adjusted difference in means estimator for the average treatment effect. It is well known that in **randomized** designs, classical linear model analysis gives asymptotically correct inference for the ATE [3]. Anytime-Valid inference is also no exception. Our confidences sequences are nonparametric asymptotic confidence sequences for the average treatment effect for general non-linear models, that is, even when the linear model is misspecified. Consider the following simulation in which true data generating process is **nonlinear**, **non-Gaussian** and has **heterogeneous** treatment effects.

$$\begin{aligned} y_i|\boldsymbol{x}_i, z_i &\sim \text{Bernoulli}(p_i(\boldsymbol{x}_i, z_i)) \\ p_i(\boldsymbol{x}_i, z_i) &= \text{logistic}(-2 + x_{i1}^2 - 0.5\sin(x_{i2}) - 0.3|x_{i3}| + 0.2z_i + 0.1z_ix_{i1}) \\ z_i &\sim \text{Bernoulli}(0.25) \\ \boldsymbol{x}_i &\sim N((1, 2, 3), \boldsymbol{I}_3). \end{aligned} \tag{5}$$



This figure shows the confidence sequence (green) compared to classical confidence intervals (purple). It also shows the failure of classical confidence intervals to cover the ATE at all times, indicated at the base of the figure with black ticks.

Anytime Valid Inference in R

As our confidence sequences and sequential p-values are closed form expressions of the classical estimators and statistics, it is trivial to wrap existing software. The **avsummary** function wraps the original R **summary** function [2].

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

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

Residuals:
    Min      1Q  Median      3Q     Max
-0.7450 -0.2833 -0.1045  0.2951  1.7279

Coefficients:
            Estimate Std. Error t value Seq. p-value    2.5%     97.5%
(Intercept) 0.284857   0.006877 41.4234   5.302e-298 0.26339  0.306325
trt         0.018614   0.013664  1.3622    1.000e+00 -0.02289  0.060115
x1          0.250285   0.006827 36.6621   2.556e-242 0.22896  0.271608
x2          0.002757   0.006904  0.3993    1.000e+00 -0.01879  0.024303
x3         -0.022916   0.006888 -3.3269    2.639e-02 -0.04442 -0.001415
trt:x1      0.016373   0.013987  1.1705    1.000e+00 -0.02612  0.058866
trt:x2      0.014537   0.013558  1.0722    1.000e+00 -0.02664  0.055715
trt:x3     -0.006232   0.013595 -0.4584    1.000e+00 -0.04752  0.035057

Residual standard error: 0.3757 on 3992 degrees of freedom
Multiple R-squared:  0.315,Adjusted R-squared:  0.3138
F-statistic: 262.2 on 7 and 3992 DF,  Seq. p-value: < 2.2e-16
```

This converts all confidence intervals to confidence sequences and all p-values to sequential p-values (including F test)

Conclusion

Regression adjustment is invaluable to tech companies performing A/B tests due to the wealth of existing pre-treatment information on users. This has the potential to dramatically reduce uncertainty on ATEs. This paper addresses the need to perform inference in real-time over streams of experimental data using Anytime Valid inference. These results have broader implications for the reproducibility of research in the applied sciences. It prevents p-hacking in the data collection process. Moreover, our p-values and confidence sequences are closed form expressions of the original estimators and statistics, facilitating reevaluation of published research and wrapping of existing software.

References

1 Michael Lindon, Dae Woong Ham, Martin Tingley, Iavor Bojinov. (2022). Anytime-Valid F-Tests for Faster Sequential Experimentation Through Covariate Adjustment.
2 https://github.com/michaellindon/roshi
3 Winston Lin. "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique." Ann. Appl. Stat. 7 (1) 295 - 318, March 2013.