A Case Weighting Scheme for Primary Outcome Analysis

Tim Lycurgus & Ben B. Hansen

1 Background

In this paper, we introduce a case-weighting scheme, **P**ower-maximizing **W**eighting for **R**epeated-measurements with **D**elayed-effects, constructed with the aim of increasing the power of hypothesis tests for primary outcome analysis in randomized controlled trials (RCTs). The longitudinal nature of many RCTs means researchers often possess multiple observations on individuals but not necessarily the same number of observations on any given indivudal. As a result, the weighting scheme applied can substantially determine the ability to find an effect. Our scheme, referred to as **PWRD** weighting, addresses these issues with repeated measurements by maximizing power for a broad class of interventions.

We illustrate this method on a large IES-funded RCT testing the efficacy of a reading intervention designed to help K-3 students at risk of falling below grade level proficiency. This intervention, which we call RSEG (Reading Support in Early Grades), is a "pull-out" intervention providing targeted instruction to students who test below a certain benchmark rather than the entire classroom. The theory behind RSEG maintains that effects are delayed, i.e. students do not receive an immediate benefit upon joining the intervention but must "test in" first. Furthermore, RSEG maintains that effects are non-uniform and scattered in that only students pulled out are affected; as a result, the treatment is anything but constant when the intervention works in the intended manner.

2 Purpose

The primary objective of **PWRD** weights is constructing a method of weighting that maximizes the likelihood of rejecting a null hypothesis of no effect if a certain theory of intervention is true. In particular, we provide a weighting scheme that increases power in expectation for interventions where students are pulled out to receive targeted instruction.

3 Research Design

In educational settings, students often enter and exit the study at various points so in addition to possessing repeated observations, we frequently see different numbers of observations for different students as well. For example, RSEG examined a reading intervention on K-3 students across four years so the design looked as follows in Table 1 for the first of the four cohorts.

	Grade at Entry	**Year 1**	**Year 2**	**Year 3**	**Year 4**
Cohort 1	**3**	3	-	-	-
	2	2	3	-	-
	1	1	2	3	-
	0	K	1	2	3

Table 1: Progression of Cohort 1 through the four years of the RSEG study.

A natural question in scenarios like RSEG is how much weight to give each observation. Even in the simplest outcome analysis where we only examine outcomes when students exit the study, complications emerge. According to the theory behind RSEG, students are more likely to benefit from the intervention when they participate for more time. As a result, we are less likely to observe an effect in Cohort 1.3 than in Cohort 1.0. It should also be noted that we do not assume students who test in during a given grade would have tested in during the previous grade. Rather, fluctuations in student performance and learning may cause students previously at grade-level reading to fall behind.

Perhaps the simplest way to make use of repeated measurements is to fit a linear model predicting student-year observations from independent variables identifying the cohort and time of follow-up before estimating standard errors of these coefficients with appropriate attention to "clustering" by student or by school; in mixed modeling and general estimating equations literature, this is known as the linear model with working independence structure (Laird, 2004; Fox, 2015). These analyses effectively attach equal weights to each student observation and thus we refer to them as "flat" weights. However, this weighting scheme fails to take into account which observations will best help us detect an effect and some power is lost.

	Grade at Entry	**Year 1**	**Year 2**	**Year 3**	**Year 4**	**Difference**
Cohort 1	**3**	3.33	-	-	-	-
	2	-0.63	-0.38	-	-	0.25
	1	1.44	6.57	2.16	-	0.72
	0	-5.33	3.74	1.18	3.45	8.78

Table 2: Difference in means between treatment and control groups for the first cohort of students.

This is especially true when effects are delayed rather than instantaneous as in RSEG where students only receive the intervention once they "test in". We see this play out in Table 2 for the first cohort of students. While there is some random variation, we generally see larger differences between treatment and control means as students participate in the study for longer. **PWRD** weighting looks to address this issue with repeated measurements and non-instantaneous effects by building around the theory that the longer students participate in a study, the more likely they will be to receive an effect. In other words, the expected size of the effect at time t will be proportional to the percentage of students who directly received the intervention by time t.

3.1 PWRD Weighting

Let X range over the ten measurement opportunities depicted in Table 1. Now take some hypothesis test of no-effect such that each of

$$\mathbb{E}(\Delta_t) = \mathbb{E}(Y_{1t} - Y_{0t}|X = x)$$

for time $t = 1 \dots T$ is zero or negative against the alternative that one or more of these treatment effects is positive. If the following assumptions about the theory of the intervention

hold, **PWRD** weights will maximize power in the resulting test in expectation:

- Individual i receiving the intervention at time j gains a homogenous non-negative effect τ_{ij} at some point between j and when they exit the study. Individuals who do not receive the intervention are unaffected.
- Effect τ_{ij} is retained by individual i throughout the duration of the study, i.e. from $[j, T]$.

The slope of the test statistic will be maximized by weights of the following form:

$$w = \alpha \cdot \Sigma^{-1}\mathbf{p}_0$$

where α represents some positive constant, $\Sigma := \text{Cov}\{(\Delta_{(Z=1,t)} - \Delta_{(Z=0,t)} : t)\}$, and $p_{0t} := \mathbb{P}(\text{student in control would 'test in' to the treatment by time } t|X = x)$.

We do not apply any covariate adjustment under this formulation, but can amend that through the approach outlined in Lin et al. (2013) or through application of a Peters-Belson method (Peters, 1941; Belson, 1956) when estimating Σ. Other adjustments can extend this into an attributable effects setting (Rosenbaum, 2001).

4 Results

We compare the power under our weighting scheme with the power under both "flat" weighting and hierarchical linear models in a simulation study on RSEG data generating effects as follows:

$$Y_{Cijk} = \beta_0 + \beta_1 \text{Grade}_{ijk} + \mu_i + \epsilon_{ijk}$$

$$\mu_i = \gamma_0 + \omega_i.$$

Effects are imposed on treatment observations in a manner following the theory of "pull-out" interventions outlined previously.

In Figure 1, we see preliminary results of the power under three different weighting schemes. It is immediately apparent that **PWRD** weights outperform the other methods. This is unsurprising since **PWRD** weights attach greater importance to student observations most likely to receive an effect, thus leading to more power. This holds for both smaller and larger effect sizes.

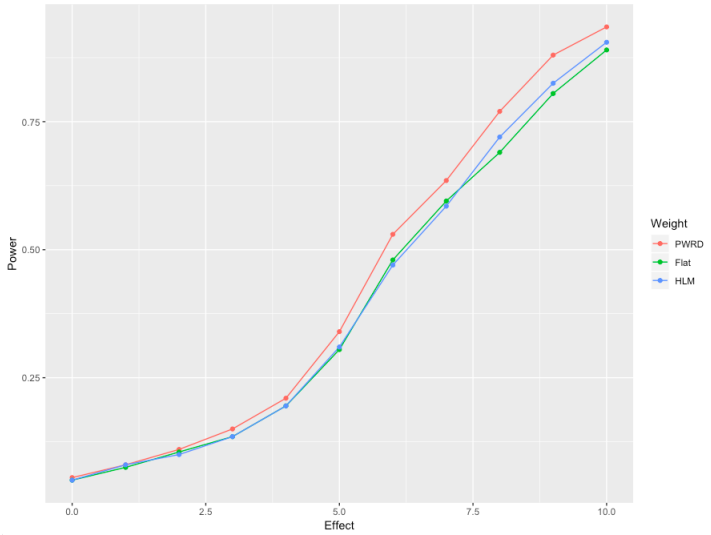

Figure 1: Power for the three weighting schemes across increasing effect sizes.

5 Conclusions

In this paper, we construct a weighting scheme that aims to maximize power in expectation for "pull-out" interventions. When the theory behind these interventions holds, **PWRD** weights provide greater power to detect an effect than many weighting schemes applied in commonly used software. We demonstrate this through a simulation study on a large-scale randomzied trial for a K-3 reading intervention.

References

Belson, W. A. (1956). A technique for studying the effects of a television broadcast. *Applied Statistics*, pages 195–202.

Fox, J. (2015). *Applied regression analysis and generalized linear models*. Sage Publications.

Laird, N. (2004). Analysis of longitudinal and cluster-correlated data. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, pages i–155. JSTOR.

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

Peters, C. C. (1941). A method of matching groups for experiment with no loss of population. *The Journal of Educational Research*, 34(8):606–612.

Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*, 88(1):219–231.