Testing for Treatment Effects in Randomized Control Trials: The Effect of Differing Cluster Sizes

Andrew V. Carter and Douglas G. Steigerwald

April 15, 2019

Abstract

We consider a common situation where we have binary responses to compare between two samples, but the observations are not independent. The dependence structure is such that there are a number of independent clusters, but that the observations within each cluster are dependent. Sandwich-type variance estimators that are based entirely on the variability between the statistics for each cluster are robust to this kind of dependence. We establish a set of sufficient conditions to imply that the variance estimators are consistent, and therefore there is a test statistic that has a standard normal distribution in the limit. The conditions require that the number of clusters goes to infinity and that there is enough homogeneity that no one cluster dominates the calculation.

1 Introduction

Assessing the effect of a group level treatment on a binary outcome depends on correctly accounting for the correlation of observations within groups. Existing approaches focus on models of the individual outcomes within groups, either indirectly through a latent variable specification or directly through a linear probability model. These approaches have drawbacks: the range of correlations allowed by the latent model is restricted and does not always capture the correlations in the observed binary outcomes, while the binary error in the linear probability model does not satisfy the conditions under which the asymptotic normality of the t-test is derived. We suggest a test statistic based on group-level outcomes and establish the asymptotic null distribution of the statistic. We also derive a measure of heterogeneity across groups that indicates when the asymptotic theory provides a good approximation.

To determine the effectiveness of a treatment administered to all individuals within a group, researchers have typically focused on individual outcomes. One popular solution to the clustering problem is to introduce a latent continuous outcome that has arbitrary within

cluster correlation. This closely parallels the model for an observed continuous outcome. Unfortunately, it is difficult to match the correlation in the latent variable to the correlation in the observed binary outcomes. Indeed, each specified latent model places restrictions on the correlations in the resultant binary outcomes, which may be violated by the observed outcomes.

A second approach is to directly model the individual outcomes with a linear probability model and use a cluster-robust t-statistic. The conditions under which asymptotic normality is obtained for the cluster-robust t-statistic do not explicitly preclude binary outcomes, but do include fourth moment restrictions on the errors. The binary errors in the linear probability model do not satisfy these conditions, indicating that the normal approximation for the t-statistic may not work well in this setting. Specifically, because the normal approximation relies on the estimation of the variance within each cluster, if all clusters have a large number of observations then the normal approximation is likely to hold. If, however, there are a small or moderate number of observations in the clusters, then the normal approximation will not hold even as the number of clusters grows without bound.

We propose an alternative approach that models the number of successes within each cluster rather than the individual outcomes. This setting arises naturally when a treatment is experimentally applied to individual observations that are grouped into clusters, such as a classroom or a village, but also corresponds to panel data, where observations on an individual over time form a cluster. Within this setting the parameter of interest is the difference in the response probability between the treatment and control groups and because of the experimental nature of the data, no other covariates are needed. The cluster level outcomes correspond to the general estimating equations (GEE) framework utilized by Liang and Zeger (1986).[1] In order to perform proper inference for this parameter, it is necessary to have a variance estimator that is robust to different dependence structures within clusters. Liang and Zeger (1986) present variance estimators, the simplest of which is a sandwich estimator that does not assume a particular correlation structure. Their discussion implies that the variance estimator adapts to the within-cluster correlation as long as the separate clusters are independent of each other. This discussion seems to rely on homogeneity across clusters, that is, all of the treatment (control) clusters are the same size and have the same within-cluster correlation structure. We formally establish consistency of the variance estimator both under cluster homogeneity and cluster heterogeneity and correspondingly show that the test statistic converges to a standard normal distribution. For the latter case, we determine how variation in cluster sizes and within-cluster correlation structures affects the finite-sample performance of the estimator.

The essential element for consistency is that the number of clusters needs to grow to

[1] Freedman (2006) raises concerns about GEE estimators by asking: If the estimators maximize a criterion function that is not the likelihood, then what meaning do these estimators have? Our response is to set out sufficient conditions for consistent estimation of the parameter of interest and thus imply that these estimators can provide reasonable inference in this problem.

infinity. There are further conditions that control the heterogeneity between clusters so that no one cluster dominates the estimate. For example, if the clusters are of vastly different sizes then we have, effectively, a smaller number of clusters. The concept of an effective number of clusters has previously been introduced for a regression model with a continuous dependent variable and nearly normal errors by Carter, Schnepel, and Steigerwald (2017).[2]

To verify the conditions for asymptotic normality in empirical settings, we consider two general models of cluster correlation. When defining the possible cluster correlation structures, we do not characterize the dependence by a simple covariance because that seems unlikely to describe a realistic class of models with binary outcomes. Unlike a normal observation, where the covariance describes the entire dependence structure, if the observations are only 0 or 1 the covariance does not naturally translate to the probability distribution of all n-tuples. To describe a realistic set of correlation patterns, we consider two models. The first, which captures the type of correlation found in panel data, is a model where each cluster consists of observations of a first-order Markov chain. A panel study that monitored large samples of independent subjects and recorded a binary response at regular time intervals over the course of the study for each subject could be modeled by this Markov chain model. The second model, which captures the type of correlation found in grouped data, is a hierarchical model where the probability of success is randomly assigned to each cluster, and then the probability of success is constant among the conditionally independent observations within the cluster.

From simulations we show that finite sample inference in the case of homogeneous clusters requires at least 100 clusters to have accurate coverage probabilities for confidence intervals and size control for the test statistic. With smaller numbers of clusters the variation in the variance estimator renders the normal approximation inaccurate, echoing the findings in Kauermann and Carroll (2001). When the clusters are heterogeneous, then it is the effective number of clusters, rather than the number of clusters, that provides a guide to inference.

Extensions to other simple inference tasks that involve a linear combination of estimated probabilities are easy to formulate. A confidence interval for a single population proportion estimate using these variance estimates would also be asymptotically valid under the conditions stated.

Section 2 describes the data structure, presents a test statistic, and contains the main result that establishes the asymptotic distribution of the statistic. Sections 3 and 4 establish when our conditions hold in two plausibly realistic models of panel data and clustered data. We have an example of how our procedure and results work out in a particular data example in Section 5.

[2] They present conditions, on the within-cluster correlation structure, the design, and the sizes of the clusters, under which a cluster(correlation)-robust variance estimator is consistent and the t-statistic converges to a standard normal distribution. Djogbenou et al. (2018) derive the asymptotic behavior of cluster-robust tests based on the bootstrap.

2 Model

This paper considers a model for binary response data which are sampled from G clusters where G_0 of those clusters are under the control condition and G_1 are measured under the treatment condition. These clusters are assumed to be independent of each other, but the sizes of the clusters n_g are arbitrary and the responses within a single cluster may be correlated.

It would seem natural to work directly with the observed data y_{ig}, the binomial outcome for individual i assigned to group g. The quantity of interest is then the difference in the response probabilities $p_1 - p_0$, where $p_1 = \mathbb{P}(y_{ig} = 1|T_g = 1)$ and T_g is the indicator for the treatment condition. To parallel the correlation structure for the continuous outcomes case, the observed data is related to a latent outcome y^*_{ig} through $y_{ig} = \{y^*_{ig} \geq 0\}$, where

$$y^*_{ig} = \alpha + \beta T_g + u_{ig}.$$

The correlation that arises within a group is captured by $\Omega_g = \mathbb{E}[u_g u_g^{\mathsf{T}}]$.

A main drawback of this approach is the fact that it is hard to determine a unique structure for Ω_g from the observed correlations in y_{ig}. Put another way, the latent model might not be able to capture the observed correlation pattern. This may seem surprising, as Ω_g is generally unrestricted. But, as Chaganty and Joe (2004) demonstrate the range of potential correlations in the observed values implied by Ω_g is restricted by α.

One could instead use the linear probability model

$$y_{ig} = \alpha + \beta T_g + u_{ig}.$$

This is precisely the form of the model analyzed in Carter, Schnepel, and Steigerwald (CSS), but their analysis assumes that the error is"nearly" normal, in that it satisifies a fourth-order moment condition. In addition to the standard drawbacks of the linear probability model that arise from the implication that $\mathbb{P}(y_{ig} = 1|T_g) = \alpha + \beta T_g$, when y_{ig} is binary then so to is the error and binary errors do not generally satisfy the fourth-order moment condition in CSS.

Our approach is to avoid specification of a latent model and work directly with y_{ig} through the number of successes in cluster g, which we denote Y_{jg} for $j = 0$ or 1 depending on whether that cluster is in the treatment or control group. We use the general estimating equations (GEE) estimators proposed in Liang and Zeger (1986), which mimic the MLE for i.i.d. data, and are simply the proportion of successes under treatment or control

$$\widehat{p}_j = \frac{1}{n_j} \sum_{g=1}^{G_j} Y_{jg} \tag{1}$$

where $n_j = \sum_g n_{jg}$ is the total number of observations in the control or treatment group (for 0 or 1 respectively.) The variance of each estimator is

$$\widehat{V}_j = \frac{1}{n_j^2} \sum_{g=1}^{G_j} \left(Y_{jg} - n_{jg}\widehat{p}_j\right)^2 . \tag{2}$$

The variance estimator allows for arbitrary correlations within clusters while also allowing n_g to vary.

2.1 Main Result

These estimators are generally asymptotically normal.

Lemma 1 *If Conditions 1–4 below are satisfied, then*

$$Z = \frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{V_1 + V_0}} \rightsquigarrow \mathcal{N}(0,1).$$

The proof is straightforward (see Appendix A.1). We are interested in establishing that this is still true using estimated variances.

Theorem 1 *Under the model described, if Conditions 1–4 are satisfied then*

$$\frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\widehat{V}_1 + \widehat{V}_0}} \rightsquigarrow \mathcal{N}(0,1).$$

The proof is in Appendix A.2. This theorem is a specific version of the consistency of the variance estimators.

The fundamental condition is that we have a large number of clusters.

Condition 1 *The number of clusters* $G_j \to \infty$.

While this condition is not necessary for consistency of $\widehat{p}_j$ (Lemma 1), for the variance estimators in (2) to be consistent, it is crucial.

Because the robust variance estimators we propose are functions only of the sum of the observations within each cluster, the characterization of the dependence within a cluster affects only the variance of those sums. Therefore, our assumptions on the within-cluster correlation structure require a bound on the amount of variation in the cluster sums.

We also need a condition to ensure that while the sizes of the clusters are not equal, the variability in their sizes is under control.

Condition 2 *The empirical coefficient of variation of the n_{jg}'s is negligible as $G_j \to \infty$*

$$\sum_{g=1}^{G_j} \left(\frac{n_{jg}}{n_j} - \frac{1}{G_j} \right)^2 \to 0.$$

for $j = 0$ and 1.

Then we need a condition on the distribution of the Y_{jg} which will be

Condition 3 *The kurtosis of each Y_{jg} is bounded*

$$\mathrm{Var}\left([Y_{jg} - n_{jg} p_j]^2 \right) \leq \kappa \ [\mathrm{Var}(Y_{jg})]^2 .$$

for all g and j.

If the binary outcomes are actually independent then Lemma C.2 implies that κ must be at least $2 + 1/(np_j(1-p_j))$ which requires that the p_j are bounded away from 1 and 0. This condition is similar to conditions that would be needed to establish asymptotic normality. It is likely possible that we could weaken this condition by truncating Y_{jg} and discarding negligible sets, but this only helps in unusual cases.

The last condition is a strengthening of Condition 2 in some situations where the variance of Y_{jg} is large. It requires that not only the sample sizes but also the variances of the Y_{jg} are not overly varied.

Condition 4

$$\sum_{g=1}^{G_j} \left(\frac{\mathrm{Var}(Y_{jg})}{n_j^2 V_j} - \frac{1}{G_j} \right)^2 \to 0,$$

for $j = 0$ or 1.

This is a measure of heterogeneity of the variances in that if all the $\mathrm{Var}(Y_{jg})$ are the same then this is exactly 0.

However, in cases where the clusters are of different sizes, it is unlikely that the variances will be the same because they are a function of the n_g. For instance, if the binary outcomes are nearly independent then $\mathrm{Var}(Y_{jg})$ are all approximately $c\, n_{jg} p_j(1-p_j)$. In this case, the $V_j \approx c\, p_j(1-p_j)/n_j$ so that Condition 4 follows immediately from Condition 2. For models where there is stronger correlation within the clusters, if the $\mathrm{Var}(Y_g)$ are all like $c\, n_g^2$ instead of $c\, n_g$, then Condition 4 is a stronger requirement than Condition 2.

From the proof of Theorem 1, the convergence of the test statistic depends on the variance of $\hat{V}_i$ being negligible relative to the size of the error which is chiefly controlled by a combination of Conditions 1 and 4. The quantities of interest are

$$\frac{1}{G_j} \left(1 + \frac{1}{G_j} \sum_{g=1}^{G_j} \frac{\left[\mathrm{Var}(Y_{jg}) - \bar{V}_j \right]^2}{\bar{V}_j^2} \right) \tag{3}$$

where $\bar{V}_j$ is the average value of the variances $\bar{V}_j = n_j^2 V_j / G_j$. Our typical condition will depend on (3) going to 0 in both the control and treatment group.

This condition suggests that the magnitude of (3) can serve as a guide to how closely a data set corresponds to the conditions under which the test statistic is asymptotically normal. We will define the effective number of clusters for the treatment (control) group as

$$G_j^* = G_j \left(1 + \frac{1}{G_j} \sum_{g=1}^{G_j} \frac{\left[\mathrm{Var}(Y_{jg}) - \bar{V}_j \right]^2}{\bar{V}_j^2} \right)^{-1} . \tag{4}$$

This measure adjusts the observed number of clusters downward to account for the degree of heterogeneity in the sample, where the heterogeneity arises both from clusters of different sizes and differing correlations across clusters. The effective number of clusters defined in (4) is similar to the quantity introduced by Carter, Schnepel, and Steigerwald for linear models of individual-level observations. For convergence of the test statistic, it is sufficient that G_1^* and G_0^* are unbounded as $n \to \infty$.

Remark We can conceive of situations where it is expensive to obtain control groups so that the effective number of clusters is bounded for the control group, but the variance estimate is consistent because $V_0/(V_1 + V_0) \to 0$. This would be a situation with many observations in a few control clusters where we could essentially consider p_0 known. Then the test statistic depends almost entirely on the variance of the estimate of the treatment proportion.

3 A Model for Panel Data

With panel data the cluster correlations typically follow a pattern of temporal decay. If there is weak dependence among the observations within each cluster, then the required conditions for asymptotic convergence can be simplified. In particular, we will see that Condition 2 on the cluster sizes is sufficient to imply that Condition 4 is also true. This implication follows when $\mathrm{Var}(Y_g) = cn_g$ in each cluster. A completely independent model would have $c = p(1-p)$.

We model the within cluster, or panel, correlation as a two-state Markov chain. Let $\mathbf{y}_{tg}$ denote the observation for period t in cluster g. Each cluster is independent of other clusters and $Y_g = \sum_{t=1}^{n_g} \mathbf{y}_{tg}$.

For this discussion, we suppress the cluster index on individual observations to keep the notation simple. Stationarity will be assumed, and implies $\mathbb{E}(\mathbf{y}_t) = p$ for all t. The Markov process implies $\mathbb{E}\left(f(\mathbf{y}_3) \mid \mathbf{y}_1, \mathbf{y}_2\right) = \mathbb{E}\left(f(\mathbf{y}_3) \mid \mathbf{y}_2\right)$. We will parameterize the correlation between successive observations

$$\mathbb{E}\left[(\mathbf{y}_{t+1} - p) \mid \mathbf{y}_t\right] = \alpha(\mathbf{y}_t - p),$$

or equivalently

$$\mathrm{Cov}(\mathbf{y}_t, \mathbf{y}_{t+1}) = \alpha p(1-p).$$

This implies

$$\mathbb{E}\left[(\mathbf{y}_{t+s} - p) \mid \mathbf{y}_t\right] = \alpha^s(\mathbf{y}_t - p). \tag{5}$$

To relate the parameters of the model to Conditions 3 and 4 observe first that

$$\mathrm{Var}(Y_g) = n_g p(1-p)\left[\frac{1+\alpha}{1-\alpha}\right] \tag{6}$$

implying that

$$G^* = \frac{1}{G}\left(1 + \frac{1}{G}\sum_g \frac{(n_g - n/G)^2}{(n/G)^2}\right)^{-1} \tag{7}$$

so for the Markov model Condition 2 implies Condition 4.

For Condition 3 we establish that

$$\begin{aligned}\mathrm{Var}\left((Y_g - np)^2\right) &= 3n^2p^2(1-p)^2\left(\frac{1+\alpha}{1-\alpha}\right)^2 - n^2p^2(1-p)^2\left[\frac{1+\alpha}{1-\alpha}\right]^2 + O(n) \\ &\leq 2n^2p^2(1-p)^2\left(\frac{1+\alpha}{1-\alpha}\right)^2. \end{aligned} \tag{8}$$

Hence Condition 3 holds with $\kappa = 2$. We establish (6) and (8) in Appendix B.

4 A Model for Data Grouped into Clusters

With data clustered into groups the cluster correlations typically do not follow a pattern of temporal decay. To capture this pattern, we use a Beta-Binomial model in which the response probability for cluster g is a random variable p_g from a Beta distribution with expectation p and variance $\gamma p(1-p)$. The number of successes in cluster g, given the draw of p_g, is $Y_g \mid p_g \sim \mathrm{Binomial}(n_g, p_g)$.

In this setting,

$$\begin{aligned}\mathbb{E}(Y_g) &= n_g p \\ \mathrm{Var}(Y_g) &= \left[n_g(1-\gamma) + n_g^2\gamma\right] p(1-p)\end{aligned}$$

There are two sources of randomness that enter the variance: first that Y_g is binomial and second that p_g is a random variable. Depending on the value of γ, the variance of Y_g is between $n_g p(1-p)$ and $n_g^2 p(1-p)$.

Condition 4 is more interesting in this model because the variance can be of order n_g^2. The variance of either estimator is

$$V = \mathrm{Var}(\hat{p}) = (1-\gamma)np(1-p) + \gamma p(1-p)\sum_g n_g^2.$$

Condition 4 therefore requires the bound

$$\sum_{g=1}^{G}\left(\frac{n_g(1-\gamma)+n_g^2\gamma}{(1-\gamma)n+\gamma\sum_g n_g^2}-\frac{1}{G}\right)^2 \leq \sum_{g=1}^{G}\left(\frac{n_g^2}{\sum_g n_g^2}-\frac{1}{G}\right)^2$$

which is achieved for $\gamma = 1$. This bound is of the right order for large clusters as long as $\gamma > 0$. Thus, the effective number of clusters is

$$G_j^* = G_j\left(1+\frac{1}{G_j}\sum_{g=1}^{G_j}\frac{(n_{jg}^2-\nu_j)^2}{\nu_j^2}\right)^{-1}$$

where $\nu_j = \sum_g n_{jg}^2/G_j$.

4.1 Kurtosis Condition

To establish that this model satisfies Condition 3, we need a bound on $\mathrm{Var}\left([Y_g - n_g p]^2\right)$. This variance can be calculated conditioning on the group level proportions p_g using a standard decomposition of the variance

$$\mathrm{Var}\left([Y_g - n_g p]^2\right) = \mathbb{E}\,\mathrm{Var}([Y_g - n_g p]^2 \mid p_g) + \mathrm{Var}\left(\mathbb{E}\left([Y_g - n_g p]^2 \mid p_g\right)\right).$$

Because $Y_g \sim \mathrm{Binomial}(n_g, p_g)$, with $p_g \in [0,1]$,

$$\mathbb{E}\,\mathrm{Var}([Y_g - n_g p]^2 \mid p_g) \leq n_g^3 + n_g^2 + n_g. \tag{9}$$

The second term is the variance of

$$\mathbb{E}\left([Y_g - n_g p]^2 \mid p_g\right) = n_g p_g(1-p_g) + n_g^2(p-p_g)^2.$$

For larger clusters, the largest term will therefore be

$$\mathrm{Var}\left(n_g^2(p-p_g)^2\right) = n_g^4\,\mathrm{Var}([p-p_g]^2) \leq 3n_g^4\gamma^2 p(1-p)$$

which implies that κ must be at least

$$\frac{n_g^4\,\mathrm{Var}([p-p_g]^2)}{[\mathrm{Var}(Y_g)]^2} \leq \frac{3}{p(1-p)}. \tag{10}$$

Therefore, Condition 3 holds for some κ as long as p is bounded away from 0 and 1. We establish (9) and (10) in Appendix C.

5 Data Example

We consider an example from Ochi and Prentice (1984) that is also considered in McCulloch (1994). The original experiment measured the survival rates among 32 rat litters, 16 of those litters were in a control group and 16 were in a treatment group.

If we assume that the all of the outcomes are independent, then a standard independent sample binomial model finds that $\hat{p}_0 = 0.899$ and $\hat{p}_1 = 0.772$ with a 95% confidence interval for the difference

$$\hat{p}_1 - \hat{p}_0 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}} = [-0.209,\ -0.043]\,.$$

If we wanted a Z statistic for testing if the difference is significant then it would be

$$\frac{\hat{p}_1 - \hat{p}_0}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_0}\right)}} = -2.98$$

which is statistically significant.

We are not satisfied with this result because it seems very likely that the observations are highly correlated within each litter. Our cluster-robust estimator is robust to this source of correlation. It estimates the probabilities the same way but uses $\hat{V} = n^{-2}\sum(Y_g - n_g\hat{p})^2$ to estimate the variance in the control and treatment group. We find $\hat{V}_1 = 0.00449$ and $\hat{V}_0 = 0.00067$. This results in a 95% confidence interval of

$$\hat{p}_1 - \hat{p}_0 \pm 1.96\sqrt{\hat{V}_1 + \hat{V}_0} = [-0.267,\ 0.014]\,,$$

and a Z statistic

$$\frac{\hat{p}_1 - \hat{p}_0}{\sqrt{V_0 + V_1}} = -1.759$$

which might be significant for a one-sided test. It certainly calls into doubt our original conclusion. The robust procedure results in an interval which is 70% larger than under the independence assumption.

If we assume that the observations are all independent, then $n_1 = 158$ and $n_0 = 145$. By most common standards, this implies that the central limit theorem is appropriate. In contrast, the estimators of the variance $\hat{V}_j$ are based on a calculation from 16 clusters each. In fact, because the litters are of different sizes, we should calculate an effective number of clusters. In this data, $G_0^* = 15.2$ and $G_1^* = 15.4$ so there is relatively little effect from the heterogeneous sample sizes. Comparing this to 30 degrees of freedom, it is probably just on the edge of being an appropriate normal approximation.

A sophisticated GLM approach to this problem using over-dispersion to account for the correlation within the litters, leads to an estimated treatment effect of -0.961. Using the

"quasibinomial" family in `R`, the model summary reports a t statistics for the treatment effect of -1.778 (very close to our robust estimate.) A 95% confidence interval for the difference of the probabilities from this model is $[-0.358, 0.009]$. These results are broadly consistent with our simpler, more robust estimates.

A Proof of Results

A.1 Proof of Lemma 1

We just need to check a Lyapunov condition using

$$X_g = \frac{Y_{jg} - n_{jg}p_j}{n_j\sqrt{V_j}}$$

we have $\mathbb{E}X_g = 0$ and $\sum \operatorname{Var} X_g = 1$. Then a bound on

$$\sum_g \mathbb{E}X_g^4$$

implies our central limit theorem.

$$\mathbb{E}X_g^4 = \frac{\mathbb{E}\left(Y_{jg} - n_{jg}p_j\right)^4}{n^4V_j^2} \leq \frac{(1+\kappa)\left[\operatorname{Var}(Y_{jg})^2\right]}{n^4V_j^2}$$

by Condition 3. Then,

$$\sum_g \mathbb{E}X_g^4 \leq (1+\kappa)\frac{\sum_g \operatorname{Var}(Y_{jg})^2}{\left[\sum_g \operatorname{Var}(Y_{jg})\right]^2}$$

is bounded as in equation (15) by

$$\sum_g \mathbb{E}X_g^4 \leq \frac{1+\kappa}{G} + (1+\kappa)\sum_{g=1}^{G}\left(\frac{\operatorname{Var}(Y_g)}{n^2V} - \frac{1}{G}\right)^2$$

which goes to 0 by Condition 4.

A.2 Proof of Theorem 1

In order to get asymptotic normality of the testing procedure, we need to establish the consistency of the variance estimator. In particular,

$$\frac{\widehat{V}_1 + \widehat{V}_0}{V_1 + V_0} \xrightarrow{\mathbb{P}} 1 \implies \frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{V}_1 + \widehat{V}_0}} \rightsquigarrow \mathcal{N}(0,1),$$

where we have suppressed the fact that the estimators and the variance are functions of n.

The proof will decompose each term in the estimator

$$\widehat{V}_1 + \widehat{V}_0 = \frac{1}{n_0^2}\sum_{g=1}^{G_0}\left([Y_{0g} - n_{0g}p_0] + n_{0g}\left[p_0 - \widehat{p}_0\right]\right)^2 + \frac{1}{n_1^2}\sum_{g=1}^{G_1}\left([Y_{1g} - n_{1g}p_1] + n_{1g}\left[p_1 - \widehat{p}_1\right]\right)^2$$

Following Lemma A.1, it will be sufficient to show that

$$\widetilde{V}_A = \frac{1}{n_0^2}\sum_{g=1}^{G_0}\frac{(Y_{0g} - n_{0g}p_0)^2}{V_1 + V_0} + \frac{1}{n_1^2}\sum_{g=1}^{G_1}\frac{(Y_{1g} - n_{1g}p_1)^2}{V_1 + V_0} \xrightarrow{\mathbb{P}} 1,$$
$$\widetilde{V}_B = \frac{1}{n_0^2}\sum_{g=1}^{G_0}\frac{n_{0g}^2\left(p_0 - \widehat{p}_0\right)^2}{V_1 + V_0} + \frac{1}{n_1^2}\sum_{g=1}^{G_1}\frac{n_{1g}^2\left(p_1 - \widehat{p}_1\right)^2}{V_1 + V_0} \xrightarrow{\mathbb{P}} 0.$$

First, we note that

$$V_j = \mathrm{Var}(\widehat{p}_j) = \frac{1}{n_j^2}\sum_{g=1}^{G_j}\mathrm{Var}(Y_{jg})$$

Thus,

$$\begin{aligned}\mathbb{E}\left[\widetilde{V}_A\right] &= \frac{1}{(V_1+V_0)n_0^2}\sum_{g=1}^{G_0}\mathbb{E}(Y_{0g} - n_{0g}p_0)^2 + \frac{1}{(V_1+V_0)n_1^2}\sum_{g=1}^{G_1}\mathbb{E}(Y_{1g} - n_{1g}p_1)^2 \\ &= \frac{1}{(V_1+V_0)}\left[\frac{1}{n_0^2}\sum_{g=1}^{G_0}\mathrm{Var}(Y_{0g}) + \frac{1}{n_1^2}\sum_{g=1}^{G_1}\mathrm{Var}(Y_{1g})\right] = 1. \qquad (11)\end{aligned}$$

Therefore, $\widetilde{V}_A \xrightarrow{\mathbb{P}} 1$ if we can prove that $\mathrm{Var}(\widetilde{V}_A) \to 0$. This is the more difficult part of the calculation, and it is established in Lemma A.2 as a result of Conditions 3 and 4.

The second piece of the argument is to show that $\widetilde{V}_B \xrightarrow{\mathbb{P}} 0$

$$\widetilde{V}_B = \frac{(p_0 - \widehat{p}_0)^2}{V_0}\left(\left[\frac{V_0}{V_1+V_0}\right]\frac{1}{n_0^2}\sum_{g=1}^{G_0}n_{0g}^2\right) + \frac{(p_1 - \widehat{p}_1)^2}{V_1}\left(\left[\frac{V_1}{V_1+V_0}\right]\frac{1}{n_1^2}\sum_{g=1}^{G_1}n_{1g}^2\right)$$

The asymptotic normality of our estimator from Lemma 1 implies that

$$\frac{(p_j - \widehat{p}_j)^2}{V_j} = O_P(1).$$

Therefore, $\widetilde{V}_B$ converges to 0 if

$$\left[\frac{V_0}{V_1+V_0}\right]\frac{1}{n_0^2}\sum_{g=1}^{G_0} n_{0g}^2 \to 0,$$

and

$$\left[\frac{V_1}{V_1+V_0}\right]\frac{1}{n_1^2}\sum_{g=1}^{G_1} n_{1g}^2 \to 0.$$

We can see that this requires large numbers of somewhat homogeneous clusters

$$\sum_{g=1}^{G}\frac{n_g^2}{n^2} = \frac{1}{G} + \sum_{g=1}^{G}\frac{(n_g - n/G)^2}{n^2},$$

which goes to zero when Conditions 1 and 2 are true.

A.3 Statement and Proof of Additional Lemmas

We start with an instance of a continuous mapping theory.

Lemma A.1 *Assuming that X_{ni} and Y_{ni} are triangular arrays with $i = 1, \ldots, n$ and*

$$\sum X_{ni}^2 \xrightarrow{\mathbb{P}} c; \qquad \sum Y_{ni}^2 \xrightarrow{\mathbb{P}} 0.$$

Then

$$\sum_{i=1}^{n}(X_{ni}+Y_{ni})^2 \xrightarrow{\mathbb{P}} c. \tag{12}$$

Proof of Lemma A.1 The quadratic sum is

$$\sum_{i=1}^{n}(X_{ni}+Y_{ni})^2 = \sum_{i=1}^{n}X_{ni}^2 + \sum_{i=1}^{n}Y_{ni}^2 + 2\sum_{i=1}^{n}X_{ni}Y_{ni}. \tag{13}$$

The only technical issue is taking care of the interaction term. We can use Cauchy–Schwarz

$$\left|\sum_{i=1}^{n}X_{ni}Y_{ni}\right| \le \left[\sum_{i=1}^{n}X_{ni}^2\right]^{1/2}\left[\sum_{i=1}^{n}Y_{ni}^2\right]^{1/2} \xrightarrow{\mathbb{P}} c^{1/2}(0) = 0,$$

using Slutsky's Lemma and the continuous mapping theorem. Another application of Slustky's lemma to the sum in equation (12) completes the proof.

Finally, this technical lemma is needed to establish Theorem 1.

Lemma A.2 *Under Conditions 1, 3 and 4,*

$$\mathrm{Var}(\widetilde{V}_A) \to 0.$$

Proof of Lemma A.2 It is sufficient to bound the contribution to the variance by either the control or treatment group. Independence of the clusters implies

$$\mathrm{Var}(\widetilde{V}/V) = \frac{1}{n^4V^2}\sum_{g=1}^{G}\mathrm{Var}\left([Y_g - n_g p]^2\right), \tag{14}$$

where we have suppressed the index j which indicates control or treatment.

From Condition 3, $\mathrm{Var}\left([Y_g - n_g p]^2\right) \leq \kappa\ [\mathrm{Var}(Y_g)]^2$ we get a bound on (14)

$$\begin{aligned}
\mathrm{Var}(\widetilde{V}/V) &\leq \kappa\sum_{g=1}^{G}\frac{[\mathrm{Var}(Y_g)]^2}{n^4V^2} \\
&= \frac{\kappa}{G} + \kappa\sum_{g=1}^{G}\frac{\left(\mathrm{Var}(Y_g) - n^2V/G\right)^2}{n^4V^2} \\
&= \frac{\kappa}{G} + \kappa\sum_{g=1}^{G}\left(\frac{\mathrm{Var}(Y_g)}{n^2V} - \frac{1}{G}\right)^2
\end{aligned} \tag{15}$$

This will be small if the variance of the sums does not vary greatly from cluster to cluster. Our Conditions 1 and 4 will therefore imply that $\mathrm{Var}(\widetilde{V}_A) \to 0$.

B Markov Model Details

B.1 Variance

We first establish (6). The geometric decay property in (5) is our main tool for calculation the moments of Y.

$$\begin{aligned}
\mathrm{Var}(Y) = \mathbb{E}(Y - np)^2 &= \mathbb{E}\left[\sum_{j=1}^{n}(\mathbf{y}_j - p)\right]^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}(\mathbf{y}_j - p)(\mathbf{y}_i - p) \\
&= np(1-p) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\alpha^{j-i}p(1-p)
\end{aligned}$$

Then we can work out the value of this double sum,

$$\begin{aligned}
\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\alpha^{j-i} &= \sum_{i=1}^{n-1}\sum_{k=1}^{n-i}\alpha^{k} \\
&= \sum_{i=1}^{n-1}\frac{\alpha-\alpha^{n-i+1}}{1-\alpha} \\
&= \frac{\alpha}{1-\alpha}\sum_{i=1}^{n-1}\left(1-\alpha^{n-i}\right) \\
&= \frac{(n-1)\alpha}{1-\alpha}+\frac{\alpha^2}{(1-\alpha)^2}\left(1-\alpha^{n-1}\right) \\
&= \frac{n\alpha}{1-\alpha}+\frac{\alpha}{(1-\alpha)^2}\left(\alpha-\alpha^n-(1-\alpha)\right) \\
&= \frac{n\alpha}{1-\alpha}+\frac{\alpha}{(1-\alpha)^2}\left(2\alpha-1-\alpha^n\right).
\end{aligned}$$

Thus, the variance is

$$\begin{aligned}
\mathrm{Var}(Y_g) &= n_g p(1-p)\left[\frac{1+\alpha}{1-\alpha}\right]+\frac{2p(1-p)\alpha}{(1-\alpha)^2}\left(2\alpha-1-\alpha^{n_g}\right) \\
\bar{V} &= \frac{p(1-p)}{1-\alpha}\left[\frac{n}{G}(1+\alpha)+\frac{2\alpha(2\alpha-1)}{1-\alpha}-\frac{2\alpha}{G(1-\alpha)}\sum_h \alpha^{n_h}\right] \\
\mathrm{Var}(Y_g)-\bar{V} &= \frac{p(1-p)}{1-\alpha}\left[\left(n_g-\frac{n}{G}\right)(1+\alpha)+\frac{2\alpha}{G(1-\alpha)}\sum_h\left(\alpha^{n_h}-\alpha^{n_g}\right)\right] \\
\frac{\mathrm{Var}(Y_g)-\bar{V}}{\bar{V}} &= \frac{n_g-n/G}{n/G}\left(1-\frac{2\alpha(2\alpha-1)-2\alpha/G\sum_h\alpha^{n_h}}{\frac{n}{G}(1-\alpha^2)+2\alpha(2\alpha-1)-2\alpha/G\sum_h\alpha^{n_h}}\right)+\frac{2\alpha}{G(1-\alpha)}\sum_h\left(\alpha^{n_h}-\alpha^{n_g}\right)/\bar{V}
\end{aligned}$$

In particular, if the cluster sizes, n_g are reasonably large, then

$$\sum_g \frac{\left(\mathrm{Var}(Y_g)-\bar{V}\right)^2}{\bar{V}^2} \approx \sum_g \frac{(n_g-n/G)^2}{(n/G)^2}$$

This is negligible by Condition 2.

B.2 Kurtosis

The kurtosis of this distribution is very small, and for large n looks just like a normal distribution.

We want to calculate

$$\mathrm{Var}\left((Y-np)^2\right) = \mathbb{E}\left[\sum_j (\mathbf{y}_j - p)\right]^4 - n^2p^2(1-p)^2\left[\frac{1+\alpha}{1-\alpha}\right]^2 + O(n).$$

We need to extend our earlier calculation. First, note that $\mathbf{y}_j^2 = \mathbf{y}_j$ so that

$$\mathbb{E}(\mathbf{y}_j - p)^2(\mathbf{y}_i - p) = \mathbb{E}\left[\mathbf{y}_j - 2p\mathbf{y}_j + p^2\right](\mathbf{y}_i - p) = \\ \mathbb{E}\left[(\mathbf{y}_j - p)(1-2p) + p(1-p)\right](\mathbf{y}_i - p) = \alpha^{j-i}p(1-p)(1-2p)$$

If we then assume that $i \le j \le k \le \ell$,

$$\begin{aligned}\mathbb{E}(\mathbf{y}_i - p)(\mathbf{y}_j - p)(\mathbf{y}_k - p)(\mathbf{y}_\ell - p) &= \alpha^{\ell-k}\mathbb{E}(\mathbf{y}_i - p)(\mathbf{y}_j - p)(\mathbf{y}_k - p)^2 \\ &= \alpha^{\ell-k}\mathbb{E}(\mathbf{y}_i - p)(\mathbf{y}_j - p)\left[(\mathbf{y}_k - p)(1-2p) + p(1-p)\right] \\ &= \alpha^{\ell-j}(1-2p)\mathbb{E}(\mathbf{y}_i - p)(\mathbf{y}_j - p)^2 + \alpha^{\ell-k}p(1-p)\mathbb{E}(\mathbf{y}_i - p)(\mathbf{y}_j - p) \\ &= \alpha^{\ell-i}p(1-p)(1-2p)^2 + \alpha^{\ell-k+j-i}p^2(1-p)^2\end{aligned}$$

It is suggestive to look at the size of these terms when $\alpha = 0$. The first term is non-zero only when $j = i$, and so the sum over all (i, j, k, ℓ) will only be over n terms with $i = j = k = l$. Thus, the first term only contributes $O(n)$. Specifically,

$$\sum_{i,j,k,\ell} \alpha^{\max(\ell-j)} = n\left(24\left[\frac{\alpha}{1-\alpha}\right]^3 + 36\left[\frac{\alpha}{1-\alpha}\right]^2 + 14\left[\frac{\alpha}{1-\alpha}\right] + 1\right) + O(1).$$

However, the second term will be non-zero for $\alpha = 0$ whenever $i = j$ and $k = \ell$. This suggests that it will be of order n^2. Specifically, for large n,

$$\sum_{i,j,k,\ell} \alpha^{\ell-k+j-i} = 3n^2\left(\frac{1+\alpha}{1-\alpha}\right)^2 + O(n)$$

All combined, this means that

$$\begin{aligned}\mathrm{Var}\left((Y-np)^2\right) = 3n^2p^2(1-p)^2\left(\frac{1+\alpha}{1-\alpha}\right)^2 - n^2p^2(1-p)^2\left[\frac{1+\alpha}{1-\alpha}\right]^2 + O(n) \\ \le 2n^2p^2(1-p)^2\left(\frac{1+\alpha}{1-\alpha}\right)^2\end{aligned}$$

Therefore, we can use $\kappa = 2$ in our calculation.

B.2.1 Fourth order moment bounds

We want to bound the contribution of $\alpha^{\ell-i}$. The instance of indices being equal complicates the calculation so we consider 8 cases.

If $i < j < k < \ell$, then

$$\begin{aligned}\sum_{\ell>k>j}\sum_{i=1}^{j-1}\alpha^{\ell-i} &= \sum_{\ell>k>j}\frac{\alpha^{\ell-j+1}-\alpha^{\ell}}{1-\alpha}\\ &\leq \sum_{\ell>k}\sum_{j=2}^{k-1}\frac{\alpha^{\ell-j+1}}{1-\alpha}\\ &\leq \sum_{\ell}\sum_{k=3}^{\ell-1}\frac{\alpha^{\ell-k+2}}{(1-\alpha)^2}\\ &\leq \sum_{\ell=4}^{n}\frac{\alpha^3}{(1-\alpha)^3}\\ &\leq n\frac{\alpha^3}{(1-\alpha)^3}\end{aligned}$$

There are 24 orderings of these four indices.

If $i = j < k < \ell$, then

$$\begin{aligned}\sum_{\ell>k}\sum_{j=1}^{k-1}\alpha^{\ell-j} &\leq \sum_{\ell}\sum_{k=2}^{\ell-1}\frac{\alpha^{\ell-k+1}}{1-\alpha}\\ &\leq \sum_{\ell=3}^{n}\frac{\alpha^2}{(1-\alpha)^2}\\ &\leq n\frac{\alpha^2}{(1-\alpha)^2}\end{aligned}$$

There are 12 orderings of these four indices. We get the same result for $i < j = k < \ell$ and $i < j < k = \ell$.

If $i = j = k < \ell$, then

$$\begin{aligned}\sum_{\ell}\sum_{k=1}^{\ell-1}\alpha^{\ell-k} &\leq \sum_{\ell=2}^{n}\frac{\alpha}{1-\alpha}\\ &\leq n\frac{\alpha}{(1-\alpha)}\end{aligned}$$

There are four ways these can be ordered, and we get the same bound for $i < j = k = \ell$ and $i = j < k = \ell$. There are 6 ways to order the observations when $i = j < k = \ell$.

For $i = j = k = \ell$, the factor $\alpha^{\ell-j} = 1$ and so this contributes an n term.

These cases all together give us the bound,

$$\sum_{i,j,k,\ell} \alpha^{\ell-i} \leq \frac{n(1 + 11\alpha + 11\alpha^2 + 11\alpha^3)}{(1-\alpha)^3}.$$

The larger term The second term in the variance calculation is larger. Considering the same cases as above, if $i < j < k < \ell$

$$\begin{aligned}
\sum_{\ell>k>j} \sum_{i=1}^{j-1} \alpha^{\ell-k+j-i} &\leq \sum_{\ell} \sum_{k=3}^{\ell-1} \sum_{j=2}^{k-1} \frac{\alpha^{\ell-k+1}}{1-\alpha} \\
&= \sum_{\ell} \sum_{j=2}^{\ell-2} \sum_{k=j+1}^{\ell-1} \frac{\alpha^{\ell-k+1}}{1-\alpha} \\
&\leq \sum_{\ell=4}^{n} \sum_{j=2}^{\ell-2} \frac{\alpha^2}{(1-\alpha)^2} \\
&= \frac{\alpha^2}{(1-\alpha)^2} \sum_{\ell=4}^{n} \ell - 3 \\
&\leq \frac{n^2}{2} \frac{\alpha^2}{(1-\alpha)^2}
\end{aligned}$$

with 24 orderings.

If $i = j < k < \ell$, then

$$\begin{aligned}
\sum_{i=1}^{n-2} \sum_{k=i+1}^{n-1} \sum_{\ell=k+1}^{n} \alpha^{\ell-k} &\leq \sum_{i=1}^{n-2} \sum_{k=i+1}^{n-1} \frac{\alpha}{1-\alpha} \\
&= \frac{\alpha}{1-\alpha} \sum_{i=1}^{n-2} n - i - 1 \leq \frac{n^2}{2} \frac{\alpha}{1-\alpha}
\end{aligned}$$

with 12 orderings.

If $i < j < k = \ell$, then

$$\begin{aligned}
\sum_{ell=3}^{n} \sum_{j=2}^{\ell-1} \sum_{i=1}^{j-1} \alpha^{j-i} &\leq \sum_{ell=3}^{n} \sum_{j=2}^{\ell-1} \frac{\alpha}{1-\alpha} \\
&= \frac{\alpha}{1-\alpha} \sum_{ell=3}^{n} \ell - 2 \leq \frac{n^2}{2} \frac{\alpha}{1-\alpha}
\end{aligned}$$

with 12 orderings.

If $i = j < k = \ell$, then

$$\sum_{ell=2}^{n}\sum_{j=1}^{\ell-1} 1 = \sum_{ell=2}^{n} \ell - 1 = \frac{n(n-1)}{2} \leq \frac{n^2}{2}$$

with 6 orderings.

If $i \leq j = k \leq \ell$, then

$$\sum_{i=1}^{n}\sum_{k=i}^{n}\sum_{\ell=k}^{n} \alpha^{\ell-i} \leq \sum_{i=1}^{n}\sum_{k=i}^{n} \frac{\alpha^{k-i}}{1-\alpha} \leq \sum_{i=1}^{n} \frac{1}{1-\alpha} \leq \frac{n}{1-\alpha}$$

which is of order n and negligible. We don't need to work out all the orderings.

Putting together the interesting 4 cases,

$$\sum_{i,j,k,\ell} \alpha^{\ell-k+j-i} \leq n^2 \left(3 + 12\frac{\alpha}{1-\alpha} + 12\left(\frac{\alpha}{1-\alpha}\right)^2\right) = 3n^2\frac{(1+\alpha)^2}{(1-\alpha)^2}.$$

C Beta-Binomial Model Details

C.1 Fourth moment bound

The term in (9) involves the conditional variance, and we appeal to Lemma C.1 and Lemma C.2 for $Y_g \sim \text{Binomial}(n_g, p_g)$.

$$\begin{aligned}
&\text{Var}([Y_g - n_g p]^2) = \text{Var}([Y_g - n_g p_g]^2) + 4n_g(p_g - p)\mathbb{E}\left[Y_g - n_g p_g\right]^3 + 4n_g^2(p_g - p)^2 \,\text{Var}(Y_g - n_g p_g) \\
&= 2n_g(n_g - 3)p_g^2(1-p_g)^2 + n_g p_g(1-p_g) + 4n_g^2(p_g - p)p_g(1-p_g)(1-2p_g) + 4n_g^3(p_g - p)^2 p_g(1-p_g) \\
&= n_g p_g(1-p_g)\left[4n_g^2(p_g - p) + 2n_g\left(p_g(1-p_g) + 2(p_g - p)(1-2p_g)\right) + 1 - 6p_g(1-p_g)\right]
\end{aligned} \tag{16}$$

The next step would seem to be to use the moments of the distribution of p_g to calculate the expectation of this conditional variance. The leading term is

$$\begin{aligned}
n_g^3\mathbb{E}p_g(1-p_g)(p_g - p) &= n_g^3\mathbb{E}\left(p_g(p_g - p) - p_g^2(p_g - p)\right) \\
&= n_g^3(1-2p)\,\text{Var}(p_g) - n_g^3\mathbb{E}(p_g - p)^3 = n_g^3\left(\sigma^2 p(1-p)(1-2p) - \mathbb{E}(p_g - p)^3\right)
\end{aligned} \tag{17}$$

However, seeing as $p_g \in [0,1]$, we can rely on rough bounds like $0 \leq p_g(1-p_g) \leq 1/4$ and $p_g(1-p_g)|1-6p_g(1-p_g)| \leq 1/8$ to get

$$\mathbb{E}\operatorname{Var}([Y_g - n_g p]^2 \mid p_g) \leq n_g^3 + n_g^2 + n_g.$$

This bound is conveniently agnostic to the distribution on p_g.

The term in (10) is the variance of

$$\mathbb{E}\left([Y_g - n_g p]^2 \mid p_g\right) = n_g p_g(1-p_g) + n_g^2(p-p_g)^2$$

The variance of this expression is (I have omitted the working for the moment)

$$\begin{aligned}\operatorname{Var}\left(n_g p_g(1-p_g) + n_g^2(p-p_g)^2\right) = n_g^4\left[\operatorname{Var}((p_g-p)^2) - 4p\mathbb{E}(p_g-p)^3 + 4p^2\operatorname{Var}(p_g)\right] + \\ + n_g^3\left[-2\operatorname{Var}((p_g-p)^2) + 4\mathbb{E}(p_g-p)^3 - 8p(1-p)\operatorname{Var}(p_g)\right] + \\ + n_g^2\left[\operatorname{Var}((p_g-p)^2) - 4(1-p)\mathbb{E}(p_g-p)^3 + 4(1-p)^2\operatorname{Var}(p_g)\right] \quad (18)\end{aligned}$$

Though really, we could bound this by something like $n_g^4 + n_g^2$.

C.2 Binomial Moment Lemmas

A bound on the variance of quadratic polynomials of random variables.

Lemma C.1 *Assume that $\mathbb{E}X = 0$ and $\mathbb{E}X^4 < \infty$ then*

$$\operatorname{Var}((X+a)^2) = \operatorname{Var}(X^2) + 4a\mathbb{E}X^3 + 4a^2\operatorname{Var}(X)$$

This is simply an algebraic result

$$\begin{aligned}\operatorname{Var}((X+a)^2) &= \mathbb{E}(X+a)^4 - \left[\mathbb{E}(X+a)^2\right]^2 \\ &= \mathbb{E}X^4 + 4a\mathbb{E}X^3 + 6a^2\mathbb{E}X^2 + 4a^3\mathbb{E}X + a^4 - \left[\mathbb{E}X^2 + a^2\right]^2 \\ &= \mathbb{E}X^4 - \left[\mathbb{E}X^2\right]^2 + 4a\mathbb{E}X^3 + 4a^2\mathbb{E}X^2 \\ &= \operatorname{Var}(X^2) + 4a\mathbb{E}X^3 + 4a^2\operatorname{Var}(X).\end{aligned}$$

More specifically,

Lemma C.2 *For X a random variable from a binomial distribution on n trials with p chance of success,*

$$\operatorname{Var}((X-np)^2) = 2n(n-3)p^2(1-p)^2 + np(1-p)$$

This is a reformulation of a standard calculation of the excess kurtosis of a binomial distribution.

References

Chaganty, N. R. and Joe, H. (2004). Efficiency of generalized estimating equations for binary responses. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **66** 851–860.

Djogbenou, A., MacKinnon, J. and Nielsen, M. (2018). Asymptotic theory and wild bootstrap inference with clustered errors. *manuscript* **Queens University**.

Freedman, D. A. (2006). On the so-called "Huber sandwich estimator" and "robust standard errors". *The American Statistician* **60** 299–302.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96** 1387–1396.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association* **89** 330–335.

Ochi, Y. and Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika* **71** 531–543.