ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Statistical Inference:

Part 1, Panel Data and Clustering

We now transition from talking about estimators to talking about how to perform statistical inference with these estimators. That is to say, we set aside the issue of consistency – we either assume our estimators are consistent or we assume that we are content with accepting the estimand for whatever it is – and instead focus on how to conduct statistical tests or construct confidence intervals for these estimators. We begin by discussing the issue of serial correlation (i.e., dependence between different observations in the same data set), particularly in the context of clustered data.

For many decades, applied micro papers were using conventional (or Eicker-White "robust") standard errors in data sets with a high degree of dependence between observations. Failing to appropriately account for this dependence can easily understate the standard errors by a factor of two or three, and it is now viewed as unacceptable to treat dependent observations as if they were independent when calculating standard errors. The paper that really brought this issue to the attention of applied researchers is Bertrand, Duflo, and Mullainathan (2004). We will review this paper and then discuss the appropriate techniques for adjusting standard errors depending on the number of independent groups (i.e., clusters) and the number of units inside each of these groups.

1 Bertrand, Duflo, and Mullainathan (2004)

1.1 Literature Review

Bertrand, Duflo, and Mullainathan (2004) (henceforth BDM) examine the performance of conventional standard errors in the context of diffs-in-diffs (DD) estimators that are popular

in many applied micro fields (labor, public, development, health, etc.). They begin by summarizing the state of the DD literature from 1990 to 2000. Using a survey of 92 DD papers drawn from six journals (AER, ILRR, JOLE, JPE, JPubE, and QJE), they find that 65 use more than two periods of data. Of these papers, the average number of periods used is 16.5, creating the potential for a large number of dependent observations within each cross-sectional unit. Only five of these papers make any correction for serial correlation across time within cross-sectional units. Four of the papers use parametric AR corrections (which turn out to be ineffective), and only one allows for arbitrary serial correlation within each cross-sectional unit (the recommended solution). To summarize, the state of the applied literature during this time period with respect to computing standard errors was nothing short of appalling.

1.2 Theory

Although it did not make it into the published version, the original BDM working paper contained a useful section on the bias of the OLS standard errors in the presence of AR(1) auto-correlation. Consider a simple bivariate regression of the form:

$$y_t = \alpha + \beta x_t + \varepsilon_t$$

Though we have only one cross-sectional unit in this case, we will in general assume that errors are independent across cross-sectional units (i.e., clusters) but dependent within cross-sectional units (i.e., over time within a given unit). Assume that x_t follows an AR(1) process with auto-correlation parameter λ and that ε_t follows an AR(1) process with auto-correlation parameter ρ. This process implies that the correlation between two observations of x_t that are t^* periods apart is λ^{t^*}. Likewise the correlation between two observations of ε_t that are t^* periods apart is ρ^{t^*}. It can be shown that

$$\text{Var}(\hat{\beta}) = \frac{\sigma_\varepsilon^2}{T\sigma_x^2}(1 + 2\rho\frac{\sum_{t=1}^{T-1} x_t x_{t+1}}{T\sigma_x^2} + 2\rho^2\frac{\sum_{t=1}^{T-2} x_t x_{t+2}}{T\sigma_x^2} + ... + 2\rho^{T-1}\frac{x_1 x_T}{T\sigma_x^2}).$$

The OLS standard errors in contrast are estimated as

$$\widehat{\text{Var}}(\hat{\beta}) = \frac{\hat{\sigma}_{\varepsilon}^2}{T\sigma_x^2}$$

Furthermore, as $T \to \infty$, the ratio of the estimated variance (i.e., $\widehat{\text{Var}}(\hat{\beta})$) to the true variance (i.e., $\text{Var}(\hat{\beta})$) equals $(1 - \rho\lambda)/(1 + \rho\lambda)$.

These formulas make clear several important points regarding serial correlation.[1] First, if *either* the residual, ε_t, or the regressor, x_t, is independent across observations, then there is no serial correlation bias in the standard errors. This is because if either $\rho = 0$ or $E[x_t x_s = 0]$ (for $t \neq s$), then $\text{Var}(\hat{\beta}) = \sigma_{\varepsilon}^2/T\sigma_x^2$ (asymptotically, $(1-\rho\lambda)/(1+\rho\lambda) = 1$). This fact explains why we are rarely concerned about serial correlation in randomized trials. For example, consider implementing a randomized intervention in the Bay Area in which some unemployed workers receive job training and others do not – we wish to compare outcomes of the treated workers to those of the untreated workers. These workers' outcomes are surely correlated with each other, since all of them are affected by similar macroeconomic shocks. Nevertheless, there is no serial correlation problem because the treatment is randomly assigned at the individual level, so by definition there can be no serial correlation in treatment assignments across different individuals.

Second, the presence of positive serial correlation (the most common type) in both ε_t and x_t will lead the estimated variance of $\hat{\beta}$ to be too small relative to the true variance of $\hat{\beta}$.[2] This should be obvious as $T \to \infty$ ($(1 - \rho\lambda)/(1 + \rho\lambda) < 1$), but it also holds for finite T – the ρ terms and the expectations of the $x_t x_s$ terms in $\text{Var}(\hat{\beta})$ will be positive. As ρ and λ increase (i.e., the serial correlation gets worse), the bias in the estimated variance gets worse. This fact should be somewhat intuitive – if there is positive dependence between observations on both the treatment and outcome sides, then you effectively have less data than it seems. In an extreme case, if you had perfect serial correlation in both the outcome

[1] Technically, these points are specific to the simple AR(1) model, but they will often hold in more general forms of serial correlation too.

[2] The same thing is true if both ε_t and x_t are negatively serially correlated, but negative serial correlation in both variables is rare in practice.

and the treatment, then you would just be repeating the same observation again and again as you added additional time periods, and no new information would actually enter the data set.

Third, the degree of bias is affected by T. Note that the ratio of $\text{Var}(\hat{\beta})/\widehat{\text{Var}}(\hat{\beta})$ is:

$$1 + 2\rho\frac{\sum_{t=1}^{T-1} x_t x_{t+1}}{T\sigma_x^2} + 2\rho^2\frac{\sum_{t=1}^{T-2} x_t x_{t+2}}{T\sigma_x^2} + ... + 2\rho^{T-1}\frac{x_t x_T}{T\sigma_x^2}$$

Although each denominator contains T, the numerator in each term should increase at rate $T/(T-1)$ or faster, while the denominators increase at only $(T+1)/T$. Furthermore, increasing T always adds more terms to the expression. Thus, the larger that T gets, the worse the bias gets, all other things being equal. This should be fairly intuitive – as T increases, the ratio of dependent observations to truly independent observations increases, and the downward bias in the standard errors becomes worse.

Finally, it is theoretically possible for the true variance of $\hat{\beta}$ to be less than the estimated variance of $\hat{\beta}$ (even ignoring sampling error). Specifically, if there were negative serial correlation in x_t and positive serial correlation in ε_t (or vice versa), then the estimated standard errors would overstate the true variance of $\hat{\beta}$. This might occur if the outcome has been first differenced (which tends to generate negative serial correlation due to mean reversion) or in a randomized experiment that implements a paired research design.

1.3 Simulations

BDM run a set of simulations using CPS data from 1979 to 1999 in order to ascertain how severe the bias in conventional standard errors is in practice. Specifically, they measure female wages in 50 states over 21 years (1,050 state-by-year cells) and then randomly generate laws that affect some states and not others. They randomly draw a year from the uniform distribution between 1985 and 1995 to determine when the simulated law takes effect. They then randomly draw 25 states that will be "treated," leaving the other 25 as controls. The treatment dummy is defined as unity for women living in a treated state during a treated

year, and zero otherwise. Note that this simulation procedure is very different from randomly assigning a treatment dummy in each state-by-year cell. Under pure random assignment, there would be no serial correlation in the treatment dummy, and the conventional standard errors should be correct (BDM confirm this fact in their simulations). But in the real world, laws don't randomly turn on and off from year to year in the same state – they turn on in a given state and then persist (i.e., there is serial correlation). Hence BDM design their simulations to replicate how laws are actually distributed in the real world.

By design these simulated laws, though serially correlated over time, are uncorrelated with any real outcome. Thus we know that on average the regression coefficient on the simulated treatment variable, $\hat{\beta}$, will equal zero.[3] The question of interest, however, is whether the standard errors are of the correct size, i.e. do we reject the null hypothesis of zero only 5% of the time at the $\alpha = 0.05$ level? The answer is no. Using the results from several hundred simulations, BDM find that they reject the null hypothesis of no effect an incredible 67% of the time when using micro level data. Aggregating the data to the state-by-year level improves matters a bit, but they still reject the null hypothesis 40-50% of the time.

BDM also experiment with changing the sample size along the two relevant dimensions (G and T). They find that reducing G while keeping T fixed at $T = 21$ has a minimal effect on the rejection rate (it still remains around 40% or higher). Reducing T to 5, however, lowers the rejection rate to 0.08 (G remains 50). Reducing T to 3 brings the rejection rate down to 0.05. So, as predicted by theory, the bias from serial correlation is less severe for relatively low values of T (i.e., small clusters).

1.4 Solutions

1.4.1 Parametric AR(1) Corrections

One possible correction entails assuming that the serial correlation takes the form of an AR(1) process (i.e., $\varepsilon_{t+1} = u_{t+1} + \rho\varepsilon_t$) and using a parametric correction based on this model (e.g.,

[3]The regression also include state and year fixed effects.

transform the data using the formula $\tilde{x}_t = x_t - \hat{\rho} x_{t-1}$). BDM find that parametric corrections are ineffective, presumably because the assumptions they impose on the exact form of the serial correlation are too restrictive. After implementing the parametric correction, BDM still find rejection rates from 18-24%. In summary, don't rely on parametric corrections that assume an AR(1) error process (or some similarly restrictive error process).

1.4.2 Collapse the Data

Another correction entails collapsing the data until the dependence issue disappears. Specifically, we solve the clustering problem by collapsing the clusters down until they only contain one or two observations each. Then the resulting data set has no dependence problem because the observations are independent of each other by virtue of the fact that the clusters are independent of each other. This method should almost always solve the dependence issue, albeit at the expense of lower precision (and possibly introducing some form of aggregation bias).

In the panel data/DD context, collapsing the data generally entails collapsing each cross-sectional unit into two time periods: pre-treatment and post-treatment. You can then estimate a regression of the outcome on a treatment indicator using the collapsed data; conventional standard errors should generate tests of the correct size (or at least close enough). This method is somewhat problematic, however, if the treatment activates at different times for different states – in that case, it's unclear what the counterfactual post-treatment period for the untreated states is. In this context, BDM suggest a variant of the following procedure:

1. Regress Y_{st} on state fixed effects, year dummies, and relevant covariates (if you have individual level data, Y_{st} corresponds to the state-by-year cell mean). Note that we are not including the treatment indicator in this regression. Collect the residuals from this regression – call them $\tilde{Y}_{st}$. Regress D_{st}, the treatment indicator, on state fixed effects, year dummies, and the same covariates that you use for Y_{st}. Collect the residuals from this regression – call them $\tilde{D}_{st}$.

2. *For the treatment states only*, divide the observations into two groups: observations from before the law and observations from after the law. Collapse the observations (which now consist of $\tilde{Y}_{st}$ and $\tilde{D}_{st}$) down to the state-by-treatment-status level (i.e., you will have two observations for each treated state: pre-treatment and post-treatment).

3. Using this collapsed data set with only treated states, regress $\tilde{Y}_{st}$ on $\tilde{D}_{st}$. The standard errors in this regression should be the correct size (or close enough).[4]

This is the Mike Anderson Approved$^{\text{TM}}$ method of collapsing the data when the policy change occurs at different periods for different cross-sectional units. It is slightly different than the methodology suggested in BDM – they suggest the same procedure except that they do not make any mention of residualizing the treatment indicator. Failing to residualize the treatment indicator results in an estimator that does not reproduce the standard collapsed DD estimator when the policy change occurs simultaneously for all treated states.[5]

When collapsing the data, BDM find that the rejection rate falls to 5-6%, using either the simple aggregation method or the residual aggregation method. When the number of clusters (states) falls to 10, they find that the simple aggregation method rejects 5% of the time while the residual aggregation method rejects 9% of the time. When the number of clusters falls to 6, the simple aggregation method rejects 7% of the time while the residual aggregation method rejects 10% of the time.

1.4.3 Arbitrary Variance-Covariance Matrix (Clustered Standard Errors)

A final correction (the generally recommended one) uses an empirical estimator that can accommodate an arbitrary variance-covariance matrix. This is the same clustered standard errors estimator that we covered in the panel data lectures. Assume that we have a panel

[4] You might think that the control states are adding nothing here since they have been discarded. However, they implicitly provided the counterfactual trajectories for the treated states when we estimated the year dummies in the first step.

[5] Anyone that can disprove me on this point gets an extra 20 percentage points added to their course grade.

data model of the form $y_{gt} = x_{gt}\beta + \varepsilon_{gt}$. Formally, the estimator is:

$$\text{Var}(\hat{\beta}) = \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \cdot \left(\sum_{g=1}^{G} X_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' X_g\right) \cdot \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1},$$

where X_g is a $T \times K$ matrix with the tth row equal to x_{gt}, and $\hat{\varepsilon}_g$ is the $T \times 1$ column vector with tth element equal to $\hat{\varepsilon}_{gt}$. We calculate $\hat{\varepsilon}_{gt}$ using the estimated regression coefficients (remember, serial correlation does not make the coefficient estimates inconsistent, it only affects the standard errors).[6]

This formula makes it clear why more clusters are better (from the perspective of computing standard errors). The middle term will provide a precise estimate of $E[X_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' X_g]$ only if G is of a reasonable size, i.e. we have a sufficient number of clusters. Otherwise, our estimate of $E[X_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' X_g]$ will be relatively unstable.

We explore the derivation of this estimator in the next section. For the moment, note that in practice the arbitrary variance-covariance estimator is implemented using Stata's ", cluster()" option. This option can be applied to a number of estimators, not just linear regression. The key underlying assumption is that although residuals may be correlated within a given cluster, they are independent across different clusters.

BDM explore the performance of the arbitrary variance-covariance matrix in their simulations. The rejection rate is 6% for 50 or 20 clusters, 8% for 10 clusters, and 12% for 6 clusters.

2 The Clustered Variance Estimator

2.1 Derivation

Clustered standard errors have replaced conventional standard errors in virtually all panel data applications, and in many other contexts as well (e.g., sampling groups of students

[6] In practice we also apply a correction for the number of clusters: $\frac{G}{G-1}\frac{N-1}{N-K} \approx \frac{G}{G-1}$, where N is the total sample size. Stata automatically applies this correction.

within classrooms, sampling groups of individuals within villages, etc.). Given their widespread adoption, you should have some understanding of the underlying algebra.

How do we derive the clustered variance estimator (i.e., the estimator that accommodates an arbitrary variance-covariance matrix)? First note that the conditional variance of $\hat{\beta}$ is:

$$E[(\hat{\beta}-\beta)(\hat{\beta}-\beta)'|X] = E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X]$$

Under homoskedasticity and independence between observations, $E[\varepsilon\varepsilon'|X] = \sigma^2 I$, and the formula above collapses to $\sigma^2(X'X)^{-1}$. Suppose that we relax these assumptions, however. Then we have:

$$E[(X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}|X] = (X'X)^{-1}X'E[\varepsilon\varepsilon'|X]X(X'X)^{-1}$$

Focus on the interior $K \times K$ matrix, $X'E[\varepsilon\varepsilon'|X]X$. The two outer $K \times K$ matrices are basically irrelevant since they are unaffected by assumptions about the residuals.

$$X'\varepsilon\varepsilon'X = X' \begin{pmatrix} \varepsilon_1^2 & \varepsilon_1\varepsilon_2 & \dots & \varepsilon_1\varepsilon_N \\ \varepsilon_2\varepsilon_1 & \varepsilon_2^2 & \dots & \varepsilon_2\varepsilon_N \\ \vdots & & \ddots & \\ \varepsilon_N\varepsilon_1 & \dots & & \varepsilon_N^2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$$

$$= \begin{pmatrix} x_1' & x_2' & \dots & x_N' \end{pmatrix} \begin{pmatrix} \sum_{j=1}^N x_j\varepsilon_j\varepsilon_1 \\ \sum_{j=1}^N x_j\varepsilon_j\varepsilon_2 \\ \vdots \\ \sum_{j=1}^N x_j\varepsilon_j\varepsilon_N \end{pmatrix} = \sum_{i=1}^N \sum_{j=1}^N x_i'x_j\varepsilon_j\varepsilon_i.$$

Reinserting the conditional expectation gives us:

$$\sum_{i=1}^N \sum_{j=1}^N x_i'x_j E[\varepsilon_j\varepsilon_i|X].$$

The independence assumption allows us to ignore all terms with $i \neq j$ in the sum $\sum_{i=1}^N \sum_{j=1}^N x_i'x_j E[\varepsilon_j\varepsilon_i|X]$, yielding $\sum_{i=1}^N x_i'x_i E[\varepsilon_i^2|X]$. This in turn gives us the Huber-Eicker-White heteroskedasticity robust standard errors: $(\sum_{i=1}^N x_i'x_i)^{-1} \sum_{i=1}^N x_i'x_i\hat{\varepsilon}_i^2 (\sum_{i=1}^N x_i'x_i)^{-1}$.

These standard errors go to zero as $N \to \infty$ because the two inverted matrices grow at a combined rate of N^2 while the interior matrix grows at a rate of only N.

Suppose, however, that we drop the independence assumption. The conditional variance of $\hat{\beta}$ is:

$$\text{Var}(\hat{\beta}) = (\sum_{i=1}^{N} x_i' x_i)^{-1} \sum_{i=1}^{N} \sum_{j=1}^{N} x_i' x_j E[\varepsilon_j \varepsilon_i | X] (\sum_{i=1}^{N} x_i' x_i)^{-1}$$

This expression presents two problems. First, this quantity need not converge to zero as $N \to \infty$ because the interior matrix can now grow at up to N^2, potentially matching the growth rate of the two inverted matrices.[7] Thus $\hat{\beta}$ can converge very slowly (or not at all, in cases of extreme dependence). Second, the empirical analog of this quantity is fatally flawed. Consider the empirical estimator for the interior matrix:

$$\sum_{i=1}^{N} \sum_{j=1}^{N} x_i' x_j \hat{\varepsilon}_j \hat{\varepsilon}_i = X' \hat{\varepsilon} \hat{\varepsilon}' X$$

The OLS residuals, $\hat{\varepsilon}$, are constructed to be orthogonal to the regressors, X. Thus $\hat{\varepsilon}' X = 0$ by construction, and an estimator of $\text{Var}(\hat{\beta})$ based on $\sum_{i=1}^{N} \sum_{j=1}^{N} x_i' x_j \hat{\varepsilon}_j \hat{\varepsilon}_i$ is guaranteed to equal zero.

Now consider a case in which we have G clusters. Within each cluster, we want to allow for dependence between observations of an arbitrary form, but we assume that observations in different clusters are independent. This assumption allows us to ignore all terms with $i \neq j$ in the sum $\sum_{i=1}^{N} \sum_{j=1}^{N} x_i' x_j E[\varepsilon_j \varepsilon_i | X]$, *as long as those terms are in different clusters.* This gives us the following estimator for the expectation of $\sum_{i=1}^{N} \sum_{j=1}^{N} x_i' x_j E[\varepsilon_j \varepsilon_i | X]$. In this estimator, we first sum up all of the cross terms within a given cluster and then sum up over all of the clusters:

$$\sum_{g=1}^{G} (\sum_{s=1}^{T} \sum_{t=1}^{T} x_s' x_t \hat{\varepsilon}_s \hat{\varepsilon}_t) = \sum_{g=1}^{G} X_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' X_g$$

[7] In other words, the estimated variance of $\sqrt{N}\hat{\beta}$ could go to infinity – we do not achieve root-N convergence rates.

We define X_g and $\hat{\varepsilon}_g$ as in Section 1.4.3. Because the regression coefficients are estimated for the entire sample, rather than for each cluster individually, we do not run into the problem that $\hat{\varepsilon}_g' X_g = 0$ by construction. The equality in the expression above should be clear when you consider that we already showed $X'\varepsilon\varepsilon' X = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i' x_j \varepsilon_j \varepsilon_i$ (replace N with T and i, j with s, t). Note also that we could easily accommodate clusters of varying sizes by indexing T as T_g.

The clustered variance estimator is thus:

$$\text{Var}(\hat{\beta}) = \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1} \cdot \left(\sum_{g=1}^{G} X_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' X_g\right) \cdot \left(\sum_{g=1}^{G} X_g' X_g\right)^{-1}$$

Note that this estimator (and the variance of $\hat{\beta}$) goes to zero as $G \to \infty$. We should be careful, however, if we have very few clusters, because we may not get a very precise estimate of $E[X_g' \hat{\varepsilon}_g \hat{\varepsilon}_g' X_g]$. We discuss this issue in the next section.

2.2 Rules for Clustering

Wooldridge sets out several rules of thumb for clustering as a function of the number of clusters (G) and the number of observations per cluster (T).

1. **Large G and Small T_g.** This is basically the case set out in BDM (G is greater than 20, and T_g is less than G). As demonstrated in BDM's simulations, the clustered standard errors perform well in this scenario.

2. **Large G and Large T_g.** You might think that problems would arise here because there are so many intra-cluster cross terms being estimated. However, for moderately large G (20? 50?), clustered standard errors appear to perform well even with large T_g.

3. **Small G and Large T_g.** If G is small (e.g., certainly if $G < 10$) and T_g is relatively large, then clustered standard errors are unlikely to perform well. Donald and Lang

(2007) suggest a method that is similar to the collapsing method discussed in Section 1.4.2 – simply collapse all the data down to the cluster level and run a regression on the G observations at the cluster level. In a DD scenario, it may be necessary to collapse the data to the cluster-by-treatment level. Note that t-statistics for the collapsed data will have only $G-K$ degrees of freedom, highlighting the inference problem with small numbers of clusters (e.g., consider the Card and Krueger (1994) paper – the data there would be collapsed to only 4 cells, essentially making inference impossible).

4. **Small G and Small T_g.** This is a less challenging version of the Small G/Large T_g scenario, so anything that works there should work here.

2.3 Additional Complications

Two factors exacerbate small numbers of clusters (small G). First, if the number of *treated* clusters is very small (e.g., one or two), the clustered standard errors can be biased even when G (total number of clusters) is relatively large. Second, if G is small and the number of observations per cluster (T_g) is very heterogeneous across clusters (i.e., some clusters have many observations, and others have only a few), then the clustered standard errors can be biased as well.

Sjoquist and Winters (2012) provide a nice example illustrating the first point above. They replicate a 2008 paper by Dynarski that estimates the effects of merit scholarship programs in Georgia and Arkansas on the stock of college-educated individuals in those states. Using a 1 percent Census sample, the Dynarski (2008) paper finds that the programs increase the stock of college-educated individuals by 0.03 percentage points with a standard error of 0.004 when clustering by state ($t \approx 7.5$). This appears highly significant. However, Sjoquist and Winters go back and estimate the same regression using a 5 percent Census sample. Since the sampling frames for the two datasets are identical, Sjoquist and Winters should be able to reproduce Dynarski's result up to sampling error. However, in the 5 percent sample they find a coefficient of 0.009 percentage points, with a clustered standard error of

0.003. The difference between these two coefficients (0.021 percentage points) is too large to be due to chance, at least according to the clustered standard errors.

The explanation of course is that the clustered standard errors are not conservative enough, so the regression estimates are not as precise as we are led to believe. Despite the fact that the total number of clusters is fairly large – I believe it includes all 50 states – there are only two treated states in the study (Georgia and Arkansas). Sjoquist and Winters explore whether two procedures that are supposed to be more robust to small numbers of clusters appear to fix the problem. The first procedure is based on a paper by Conley and Taber (2011). Their paper proposes using the distribution of residuals among control states to perform statistical inference about the program's effect in the treated state. This is very similar in spirit to the placebo tests that we implemented with the Abadie et al. synthetic control estimator back in the panel data lectures.

The exact Conley and Taber procedure for a case with a single treated state is as follows. We denote the outcome as Y, the treatment as D, and other time-varying covariates that we wish to control for as X. We denote the diffs-in-diffs estimate of the effect of D on Y as $\hat{\tau}$. Assume that state $s = 1$ is treated, and that states $s = 2, ..., S$ are controls.

1. Residualize Y_{st} and X_{st} with respect to the state fixed effects and time fixed effects. That is, regress Y_{st} on state and time fixed effects and collect the residuals from this regression, $\tilde{Y}_{st}$. Do the same with X_{st} to generate $\tilde{X}_{st}$.

2. Regress $\tilde{Y}_{st}$ on $\tilde{X}_{st}$. Save the residuals for control states from this regression. Call these residuals $\tilde{\eta}_{st}$.

3. Use the residuals $\hat{\eta}_{st}$ for control states to construct the empirical distribution of $\hat{\gamma}_s = \sum_{t=1}^{T}(D_{1t} - \bar{D}_1)\tilde{\eta}_{st} / \sum_{t=1}^{T}(D_{1t} - \bar{D}_1)^2$. D_{1t} are observations on the treatment indicator for the treated state, and $\bar{D}_1$ is the mean of the treatment indicator over time for the treated state. Note that you will have $S - 1$ observations on $\hat{\gamma}_s$, since $S - 1$ is the number of control states. Also note that $\hat{\gamma}_s$ bears some resemblance to a regression

coefficient from regressing Y on D (recall that $\tilde{\eta}_{st}$ are the residuals from a regression with Y as the dependent variable).

4. If the actual diffs-in-diffs estimate $\hat{\tau}$ is larger in magnitude than 95% of the $\hat{\gamma}_s$ observations, reject the null hypothesis of no treatment effect. Otherwise, do not.

The intuition here is that $\hat{\gamma}_s$ are similar to placebo estimates of the policy for all of the control states, since these states did not actually experience a policy change. If the actual treatment effect is much larger than (almost) all of the placebo estimates, then we can reject the null hypothesis of no treatment effect. With two or more treated states (as in Sjoquist and Winters), the procedure is only slightly more difficult; it involves calculating a different linear combination of the residuals in the third step. Software is available on Conley's or Taber's websites.

When Sjoquist and Winters apply the Conley and Taber procedure, they find that they can no longer reject the null hypothesis of no treatment effect in either the 1 percent or 5 percent Census samples. Thus their conundrum appears to be resolved. They reach a similar conclusion when using a bootstrap-based procedure by Cameron, Gelbach, and Miller. We discuss the Cameron, Gelbach, and Miller procedure in the bootstrapping lecture (not to be confused with the other Cameron, Gelbach, and Miller procedure below that addresses multi-way clustering).

2.4 Multi-way Clustering

We have so far assumed that there is only one unit of clustering – e.g., individuals (or time periods) are correlated within states but independent across states. But what if there are multiple levels of clustering? One possibility involves multiple levels of clustering that are nested within each other. For example, a panel data set might contain individuals living in cities nested within states with treatments that vary at the state level. In this scenario, the solution is to just cluster at the highest level (the state, in the example just given). Since the clustered variance estimator accommodates an arbitrary variance-covariance matrix within

each cluster, it is robust to the presence of sub-clusters within each cluster. Or, to put it another way, we only need independence across clusters, and we get that by clustering at the highest level.

If there are two levels of clustering that are not nested, however, then you may need to adjust for multi-way clustering. For example, consider a state-by-year panel data set. Our typical concern is that observations within a given state are correlated over time. However, it is also possible that there could be dependence between states within a given year; e.g., a hurricane might affect multiple southeastern states in one year. If the treatment also has a geographic correlation, then we will want to cluster by year as well as clustering by state. Constructing a cluster that contains all years and states, however, will result in having only one cluster in the data set – inference will be impossible.

Cameron, Gelbach, and Miller (2011) propose a simple way to accommodate multi-way clustering with non-nested clusters. Suppose that there are two dimensions of non-nested clustering, state (s) and year (t). The Cameron, et al. procedure requires the computation of three variance-covariance matrices for the estimator, $\hat{\beta}$. The first, $\text{Var}_s(\hat{\beta})$, is clustered at the state level. The second, $\text{Var}_t(\hat{\beta})$, is clustered at the year level. The third, $\text{Var}_{st}(\hat{\beta})$, is clustered at the state-by-year level, i.e. the intersection of the two one-way levels. The multi-way clustered standard errors are then calculated as the sum of the first two variance-covariance matrices minus the third variance-covariance matrix:

$$\text{Var}(\hat{\beta}) = \text{Var}_s(\hat{\beta}) + \text{Var}_t(\hat{\beta}) - \text{Var}_{st}(\hat{\beta})$$

The key assumption in multi-way clustering is that observations that differ in both s and t are independent of each other – this is the analog of the assumption in one-way clustering that observations with different g (i.e., different s in the state panel data case) are independent of each other.[8] Observations that share either s or t, however, may be arbitrarily correlated with each other.

[8]Whether this assumption will hold in practice depends on the context. It would appear to rule out, for example, common shocks to multiple states that persist over several years.

One theoretical issue that Cameron, et al. claim rarely occurs in practice is that the formula above could give negative estimates for one or more diagonal entries. If this occurs, they recommend simply choosing the maximum standard errors obtained from one-way clustering along each of the cluster dimensions. The procedure also generalizes to three-way clustering using an analogous formula – three one-way clustered matrices enter positively, three two-way clustered matrices enter negatively, and one three-way clustered matrix enters positively.[9]

Software to implement the multi-way clustered estimator in Stata is available on Doug Miller's website.

2.5 Bootstrap Based Improvements

Cameron, Gelbach, and Miller (2008) propose bootstrap based improvements to the clustered standard errors that should improve performance when using a small number of clusters (i.e., small G). We will discuss this technique when we cover bootstrapping.

3 Additional References

Donald, S. and K. Lang. "Inference with Difference-in-Differences and Other Panel Data." *Review of Economics and Statistics*, 2007, 89, 221-233.

Conley, T. and C. Taber. "Inference with 'Difference in Differences' with a Small Number of Policy Changes." *Review of Economics and Statistics*, 2011, 93, 113-125.

[9] You can easily figure out this formula by drawing a Venn diagram.