Appendix for "Ecological Regression with Partial Identification," forthcoming, *Political Analysis*

Wenxin Jiang* Gary King† Allen Schmaltz‡ Martin A. Tanner§

January 28, 2019

Appendix A Derivation of Confidence Interval in Proposition 4

Note that from (13), $\beta_i^b = b_i(w_1, \theta) + e_i^b$, where the residual e_i^b has mean 0.

For the district level parameter, the residuals can be averaged out over many precincts due to the central limit theorem and we can get a potentially useful conservative confidence interval, without modeling the variance of the residuals:

$$B = \sum_i N_i X_i \beta_i^b / \sum_i N_i X_i = \sum_i N_i X_i [b_i(w_1, \theta) + e_i^b] / \sum_i N_i X_i$$

$$B = \frac{\sum_i N_i X_i e_i^b}{\sum_i N_i X_i} + \frac{\sum_i N_i X_i b_i(0, \theta)}{\sum_i N_i X_i} - w_1 \frac{\sum_i N_i X_i (1 - X_i)}{\sum_i N_i X_i}.$$

The unidentified parameter $w_1 \in [wl, wu]$.

Therefore $B \in [BL(\theta), BU(\theta)]$, where

$$BL(\theta) \equiv \frac{\sum_i N_i X_i e_i^b}{\sum_i N_i X_i} + \frac{\sum_i N_i X_i b_i(0, \theta)}{\sum_i N_i X_i} - wu(\theta) \frac{\sum_i N_i X_i (1 - X_i)}{\sum_i N_i X_i};$$

$$BU(\theta) \equiv \frac{\sum_i N_i X_i e_i^b}{\sum_i N_i X_i} + \frac{\sum_i N_i X_i b_i(0, \theta)}{\sum_i N_i X_i} - wl(\theta) \frac{\sum_i N_i X_i (1 - X_i)}{\sum_i N_i X_i}.$$

*wjiang@northwestern.edu, Institute of Finance (Adjunct), Shandong University, and Department of Statistics, Northwestern University

†king@harvard.edu, Institute for Quantitative Social Science, Harvard University

‡schmaltz@fas.harvard.edu, Institute for Quantitative Social Science, Harvard University

§mat132@northwestern.edu, Department of Statistics, Northwestern University

Here wl, wu depend linearly on $\theta \equiv (w_0, c_1, d_1)$. The $b_i(0,\theta) \equiv b_i^0 + (b_i^1)^T\theta \equiv 0 + (1,1,X_i)(w_0,c_1,d_1)^T$ also depends linearly on $\theta \equiv (w_0, c_1, d_1)$, which is estimated by quadratic regression (20) as $\hat{\theta} \equiv (\hat{w}_0, \hat{c}_1, \hat{d}_1)$, with robust asymptotic variance matrix $V = \hat{a}var(\hat{\theta})$ based on a sandwich formula.[1]

Denote the first term in $BL(\theta)$ or $BU(\theta)$ as

$$TERM_1 = \frac{\sum_i N_i X_i e_i^b}{\sum_i N_i X_i}.$$

Then $E(TERM_1) = 0$. Assuming independent precincts, then the first term $TERM_1$ has asymptotic variance $Var(TERM_1) = \sum_i \left(\frac{N_i X_i}{\sum_i N_i X_i}\right)^2 var(e_i^b | N_i, X_i)$. Note that $var(e_i^b|N_i, X_i) = var(\beta_i^b|N_i, X_i)$, where β_i^b is a proportion valued in $[0,1]$. The variance of a bounded random variable in $[a,b]$ is at most $[(b-a)/2]^2$. Therefore, $var(\beta_i^b|N_i,X_i) \le (1/2)^2$ and $Var(TERM_1) \le \sum_i \left(\frac{N_i X_i (1/2)}{\sum_i N_i X_i}\right)^2$. Therefore, we know that the asymptotic standard error of the first term is bounded above by

$$sd(TERM_1) \le S_1 = (1/2)\sqrt{\sum_{i=1}^{p}\left(\frac{N_i X_i}{\sum_{i=1}^{p} N_i X_i}\right)^2}.$$

Now $wl = wl(\theta)$ is of the form $\max_{j=1}^J\{gl_j^0 + gl_j^T\theta\}$ for some constant vectors gl_j; $wu = wu(\theta)$ is of the form $\min_{j=1}^J\{gu_j^0 + gu_j^T\theta\}$, for some constant vectors gu_j. Denote $r \equiv \frac{\sum_i N_i X_i(1-X_i)}{\sum_i N_i X_i}$, $h_0 \equiv \frac{\sum_i N_i X_i b_i^0}{\sum_i N_i X_i}$, $h \equiv \frac{\sum_i N_i X_i b_i^1}{\sum_i N_i X_i}$.

Then

$$BL(\theta) = TERM_1 + \frac{\sum_i N_i X_i(b_i^0 + (b_i^1)^T\theta)}{\sum_i N_i X_i} - \min_{j=1}^{J}\{gu_j^0 + gu_j^T\theta\}r$$

$$= TERM_1 + h_0 + h^T\theta - \min_{j=1}^{J}\{gu_j^0 + gu_j^T\theta\}r.$$

We can write $BL(\theta) = \max_{j=1}^J\{BL_j\}$ where $BL_j = TERM_1 + h_0 - rgu_j^0 + (h - rgu_j)^T\theta$. Similarly, we can write $BU(\theta) = \min_{j=1}^J\{BU_j\}$ where $BU_j = TERM_1 + h_0 - rgl_j^0 + (h - rgl_j)^T\theta$.

Now define

$$\hat{B}L = \max_{j=1}^{J}\{\hat{B}L_j\}, \tag{1}$$

[1]See, e.g., https://www.stata.com/manuals/p_robust.pdf

where $\hat{B}L_j = h_0 - rgu_j^0 + (h - rgu_j)^T\hat{\theta}$;

$$\hat{B}U = \min_{j=1}^{J}\{\hat{B}U_j\}, \tag{2}$$

where $\hat{B}U_j = h_0 - rgl_j^0 + (h - rgl_j)^T\hat{\theta}$.

[It can be verified that in the previous notation of (21), we have $\hat{B}L = B(wl(\hat{\theta}), \hat{\theta})$ and $\hat{B}U = B(wu(\hat{\theta}), \hat{\theta})$.]

Note that

$$\hat{B}L_j - BL_j = -TERM_1 + (h - rgu_j)^T(\hat{\theta} - \theta);$$

$$\hat{B}U_j - BU_j = -TERM_1 + (h - rgl_j)^T(\hat{\theta} - \theta).$$

By an asymptotic normality argument, $\hat{B}L_j \approx N(BL_j, sl_j^2)$ where $sl_j \leq SL_j \equiv S_1 + \sqrt{(h - rgu_j)^T V(h - rgu_j)}$; $\hat{B}U_j \approx N(BU_j, su_j^2)$ where $su_j \leq SU_j \equiv S_1 + \sqrt{(h - rgl_j)^T V(h - rgl_j)}$, for all $j = 1, ..., J$. Assuming V is of order $O_{\mathrm{p}}(1/p)$, then all SU_j and SL_j's are also of order $O_{\mathrm{p}}(1/\sqrt{p})$. The sample variations $\hat{B}U_j - BU_j$ and $\hat{B}L_j - BL_j$ are also of order $O_{\mathrm{p}}(1/\sqrt{p})$.

Now consider various cases of the bound $B \in [BL(\theta), BL(\theta)]$. Assume that $N_i X_i(1 - X_i)$ is not almost surely 0, then the large sample limit of the sensitivity parameter $\frac{BU(\theta)-BL(\theta)}{wu(\theta)-wl(\theta)} = r = \frac{\sum_i N_i X_i(1-X_i)}{\sum_i N_i X_i}$ is a positive number due to the law of large numbers. Assume that $wu(\theta) \neq wl(\theta)$ (and therefore $BU(\theta) \neq BL(\theta)$). Then w_1 can be $close$ (within $O_{\mathrm{p}}(1/\sqrt{p})$) to only one of the end points of $[wl(\theta), wu(\theta)]$, and consequently B can be close to only one end point of $[BL(\theta), BU(\theta)]$. Without loss of generality we assume that B is close to $BL(\theta)$. (The other possibility would be similar.) Assume that the minimizing entry of $wu = \min_{j=1}^{J}\{gu_j^0 + gu_j^T\theta\}$ is unique and not tied with the other entries. Then the maximizing entry $BL(\theta) = max_{j=1}^{J}\{BL_j\}$ is unique and has an order-1 gap from the other entries, that is greater than $O_{\mathrm{p}}(1/\sqrt{p})$, which is the order of all $(\hat{B}U_j - BU_j)$'s and $(\hat{B}L_j - BL_j)$'s. Therefore $\max_{j=1}^{J}\{\hat{B}L_j\}$ $(= \hat{B}L)$ and $\max_{j=1}^{J}\{BL_j\}$ $(= BL(\theta))$ are achieved at a same j, with probability tending to 1 as $p \to \infty$. We will call this same j as $\hat{j}$. Then

$$\hat{B}L = \hat{B}L_{\hat{j}} = BL_{\hat{j}} + (\hat{B}L_{\hat{j}} - BL_{\hat{j}})$$

$$= BL(\theta) + (\hat{B}L_{\hat{j}} - BL_{\hat{j}})$$

where the last term is asymptotically normal and of order $O_{\mathrm{p}}(1/\sqrt{p})$.

There is a similar equation relating $\hat{B}U$ to $BU(\theta)$. These imply that the $\hat{B}U$ is close to $BU(\theta)$.

Since we have assumed that B is only close to one end point $BL(\theta)$, and not close to $BU(\theta)$, then B must not be close to $\hat{B}U$ or $\hat{B}U + u'$ either, for any u' of order $O_{\mathrm{p}}(1/\sqrt{p})$. Then we have $P(B > \hat{B}U + u')$ converges to 0. Then $P(B \notin [\hat{B}L - l', \hat{B}U + u']) \approx P(B < \hat{B}L - l') \approx P(B < BL(\theta) - l' + (\hat{B}L_{\hat{j}} - BL_{\hat{j}})) \leq P(l' < (\hat{B}L_{\hat{j}} - BL_{\hat{j}}))$ since $B > BL(\theta)$. Then $P(l' < (\hat{B}L_{\hat{j}} - BL_{\hat{j}})) \approx \Phi(-l'/sl_{\hat{j}}) \leq \Phi(-l'/SL_{\hat{j}}) = \Phi(-x)$ if we take $l' = xSL$. Setting $u' = xSU$ of order $O_{\mathrm{p}}(1/\sqrt{p})$ leads to an approximate upper bound of $P(B \notin [\hat{B}L - l', \hat{B}U + u'])$ being $1 - \Phi(-x) = \Phi(x)$ (for large p). Q.E.D.

Remark 1. *The coverage probability of CI$_x$ can be improved to 1, if we have w_1 lying in the interior of the bound (wl, wu). This would allow any $x > 0$ to be used in finding a confidence interval. However, the condition on w_1 cannot be checked due to its non-identifiability. The tie-breaking conditions that we assumed about the identified θ, however, can be checked by data. Then we can, e.g., use $x = 1$ and achieve coverage probability at least $\Phi(x) \approx 84\%$, or use $x = 1.282$ and achieve at least 90% coverage probability.*

Appendix B Non-emptiness of CI$_0$

In this Appendix B, we prove in the large p limit that when model assumptions hold, CI$_0$ should be nonempty.

By tracing the definition of CI$_0$ and applying the Law of Large Numbers, we find that in the large sample limit

$$CI_0 = [\inf_{v_1 \in [wl_j, wu_j]} B(v_1, \theta), \sup_{v_1 \in [wl_j, wu_j]} B(v_1, \theta)] \cap DD,$$

where DD denotes the larges sample limit of the DD bound,

$$B(v_1, \theta) = E\{N_i X_i[(w_0 + c_1 + d_1 X_i) + v_1(X_i - 1)]\}/E\{N_i X_i\}$$

as summarized in (16) and (17) before, and $[wl_j, wu_j]$, $j \in \{1, 2, 3\}$ indicate the bound of w_1 to be used according to the jth Proposition.

Proposition 1. *(Non-emptiness of CI$_0$.) For $j \in \{1,2,3\}$, assume linear contextual effects $E(\beta_i^w|X_i, N_i) = w_0 + w_1 X_i$ and $E(\beta_i^w|X_i, N_i) = b_0 + b_1 X_i$, and let $[wl_j, wu_j]$ indicate the bound of w_1 to be used according to the jth Proposition. Define in the large sample limit*

$$CI_0 = [\inf_{v_1 \in [wl_j, wu_j]} B(v_1, \theta), \sup_{v_1 \in [wl_j, wu_j]} B(v_1, \theta)] \cap DD,$$

where

$$DD = [E[N_i \max\{0, T_i - (1 - X_i)\}]/E(N_i X_i), E[N_i \min\{T_i, X_i\}]/E(N_i X_i)],$$

and

$$B(v_1, \theta) = E\{N_i X_i[(w_0 + c_1 + d_1 X_i) + v_1(X_i - 1)]\}/E\{N_i X_i\},$$

where c_1, d_1 follow (7).

Then CI$_0$ is nonempty.

Proof:

$$B(v_1, \theta) = E\{N_i X_i[(w_0 + c_1 + d_1 X_i) + v_1(X_i - 1)]\}/E\{N_i X_i\}$$

$$= E\{N_i X_i[(w_0 + c_1 X_i + d_1 X_i^2 - (w_0 + v_1 X_i)(1 - X_i)]/X_i\}/E\{N_i X_i\}$$

$$= E\{N_i X_i[(T_i - (w_0 + v_1 X_i)(1 - X_i)]/X_i\}/E\{N_i X_i\}$$

$$= E\{N_i X_i[\beta_i^b X_i + \beta_i^w(1 - X_i) - (w_0 + v_1 X_i)(1 - X_i)]/X_i\}/E\{N_i X_i\}$$

$$= E\{N_i X_i[\beta_i^b X_i + (w_0 + w_1 X_i)(1 - X_i) - (w_0 + v_1 X_i)(1 - X_i)]/X_i\}/E\{N_i X_i\}$$

$$= E\{N_i X_i[\beta_i^b + (w_1 - v_1)(1 - X_i)\}/E\{N_i X_i\}$$

$$\equiv B + (w_1 - v_1)r.$$

where we denote $r = E\{N_i X_i(1 - X_i)\}/E\{N_i X_i\}$ and $B = E\{N_i X_i \beta_i^b\}/E\{N_i X_i\}$. Then

$$CI_0 = [B + (w_1 - wu_j)r, B + (w_1 - wl_j)r] \cap DD.$$

On the other hand, the jth Proposition implies that

$$w_1 \in [wl_j, wu_j].$$

Then

$$B \in [B + (w_1 - wu_j)r, B + (w_1 - wl_j)r],$$

since $r \geq 0$. Now we also have

$$B \in DD,$$

since

$$X_i\beta_i^b = T_i - (1 - X_i)\beta_i^w \in [\max\{0, T_i - (1 - X_i)\}, \min\{T_i, X_i\}]$$

due to Duncan and Davis (1953). Then we have

$$B \in [B + (w_1 - wu_j)r, B + (w_1 - wl_j)r] \cap DD = CI_0$$

in the large sample limit. Therefore CI_0 is non-empty.

Q.E.D.

Remark 2. *In practice, one can apply the converse of this Proposition to rule out data sets with empty* CI_0*, which likely suggests either some assumptions are violated or the size of the data is not large enough for the method to work reliably. The logic is that the interval should not be empty if the assumptions all hold and the sample size is large enough.*

Appendix C Covariate Contextual Model

In this Appendix we extend the simple linear context model to include a covariate Z_i in addition to the basic regressor X_i. For example, this Z can be a function of the population N_i of the ith precinct, such as $Z_i = \log N_i$. This Z_i can be easily extended to be a vector with several covariates.

Assumption 1* *(Covariate linear contextual effects.) Assume that* $(\beta_i^b, \beta_i^w, X_i, Z_i)$*, for* $i = 1..., p$*, are iid random vectors satisfying*

$$E(\beta_i^w | X_i, Z_i) = w_0(Z_i) + \tilde{w}_1 X_i, \tag{3}$$

$$E(\beta_i^b | X_i, Z_i) = b_0(Z_i) + \tilde{b}_1 X_i, \tag{4}$$

where

$$w_0(Z_i) = \tilde{w}_0 + \tilde{w}_2 Z_i, \tag{5}$$

$$b_0(Z_i) = \tilde{b}_0 + \tilde{b}_2 Z_i, \tag{6}$$

and $\tilde{w}_0, \tilde{w}_1, \tilde{w}_2, \tilde{b}_0, \tilde{b}_1, \tilde{b}_2$ *are six non-random real parameters.*

Under this assumption, for the observed response

$$T_i = \beta_i^w (1 - X_i) + \beta_i^b X_i, \tag{7}$$

we have

$$E(T_i|X_i, Z_i) = \tilde{w}_0 + (\tilde{w}_1 + \tilde{b}_0 - \tilde{w}_0)X_i + \tilde{w}_2 Z_i + (\tilde{b}_1 - \tilde{w}_1)X_i^2 + (\tilde{b}_2 - \tilde{w}_2)X_i Z_i. \tag{8}$$

So if we do a five-parameter regression based on a model

$$E(T_i|X_i, Z_i) = \tilde{w}_0 + \tilde{c}_1 X_i + \tilde{w}_2 Z_i + \tilde{d}_1 X_i^2 + \tilde{d}_2 X_i Z_i, \tag{9}$$

we will be able to identify five parameters

$$\tilde{w}_0,\ \tilde{c}_1 = \tilde{w}_1 + \tilde{b}_0 - \tilde{w}_0,\ \tilde{w}_2,\ \tilde{d}_1 = \tilde{b}_1 - \tilde{w}_1,\ \tilde{d}_2 = \tilde{b}_2 - \tilde{w}_2. \tag{10}$$

Again there is one unidentified parameter, which we can choose as $\tilde{w}_1$.

Under this assumption, using a method similar to that of the main paper, we have the following results for bounding the unidentified $\tilde{w}_1$:

Proposition 1**(Tightest bound with if and only if.) Let* $[L_i, U_i]$ *be the DD bound for* β_i^w. *Assume* $E(\beta_i^w|X_i, Z_i) = w_0(Z_i) + \tilde{w}_1 X_i$ *for all* $(X_i, Z_i) \in A \subset (0, 1) \times \Re$*, Then*

$$E(L_i|X_i, Z_i) \leq E(\beta_i^w|X_i, Z_i) \leq E(U_i|X_i, Z_i),$$

for all $(X_i, Z_i) \in A$*, if and only if the nonidentifiable* $\tilde{w}_1$ *satisfies*

$$\sup_{(X_i, Z_i) \in A} X_i^{-1}[E(L_i|X_i, Z_i) - w_0(Z_i)] \leq \tilde{w}_1 \leq \inf_{(X_i, Z_i) \in A} X_i^{-1}[E(U_i|X_i, Z_i) - w_0(Z_i)]. \tag{11}$$

Now, similar to the proof of Proposition 2, we can use the expression of the DD bound in terms of T_i, X_i and the Jenson's inequality to express $E(L_i|X_i, Z_i)$ and $E(U_i|X_i, Z_i)$ in terms of $E(T_i|X_i, Z_i)$, which can be obtained by the five-parameter regression as before.

Proposition 2*_(Bound using five-parameter regression and general A.) Assume_ $E(\beta_i^w|X_i, Z_i) = w_0(Z_i) + \tilde{w}_1 X_i$ _for all_ $(X_i, Z_i) \in A \subset (0,1) \times \Re$_, then we have that the nonidentifiable_ $\tilde{w}_1$ _satisfies_

$$\sup_{(X_i,Z_i)\in A} X_i^{-1}[\max\{0, (E(T|X_i, Z_i) - X_i)/(1 - X_i)\} - w_0(Z_i)] \leq$$

$$\tilde{w}_1 \leq \inf_{(X_i,Z_i)\in A} X_i^{-1}[\min\{1, E(T|X_i, Z_i)/(1 - X_i)\} - w_0(Z_i)]. \tag{12}$$

When assuming Assumption 1*, we have $w_0(Z_i) = \tilde{w}_0 + \tilde{w}_2 Z_i$ _and the five parameter regression for_ $E(T_i|X_i, Z_i)$_. Then the bound can be expressed in the form of the five identified parameters_ $\psi = (\tilde{w}_0, \tilde{c}_1, \tilde{d}_1, \tilde{w}_2, \tilde{d}_2)^T$.

Regarding the choice of $A \subset (0,1) \times \Re$, as a set of (X_i, Z_i) values where we believe in Assumption 1*, one particularly convenient possibility is to reduce consider a set of (X_i, Z_i) values to a line segment that encompasses the center of data, i.e., $(X_i, Z_i) = (X_i, a_0 + a_1 X_i)$ for some $X_i \in [l, u]$, where a_0, a_1 may be obtained by regressing Z_i on X_i. This way, the formula in Proposition 2* can be made very similar to that of Proposition 2 since the set is now also for $X_i \in [l, u]$.

Proposition 3*_(Bound related to Proposition 2 in the main text using a 1-dimensional A.) Assume Assumption 1* for all_ $(X_i, Z_i) \in A = \{(x, a_0 + a_1 x) : \ x \in [l, u] \subset (0, 1)\}$, _then we have that the nonidentifiable_ $\tilde{w}_1$ _satisfies_

$$\tilde{w}_1 = w_1 - a_1 \tilde{w}_2, \tag{13}$$

where w_1 _satisfies the bound in Proposition 2 of the main paper, and the parameter_ $\theta = (w_0, c_1, d_1)^T$ _in that bound satisfies the following linear relation to the five-parameter_ $\psi = (\tilde{w}_0, \tilde{c}_1, \tilde{d}_1, \tilde{w}_2, \tilde{d}_2)^T$_:_

$$w_0 = \tilde{w}_0 + a_0 \tilde{w}_2, \ c_1 = \tilde{c}_1 + a_1 \tilde{w}_2 + a_0 \tilde{d}_2, \ d_1 = \tilde{d}_1 + a_1 \tilde{d}_2. \tag{14}$$

Remark 1* *(How to transform and apply the bound in Proposition 2 of the main text.) According to the procedure suggested by this proposition, we can first do the five parameter regression and determine ψ, and then use the relations here to transform it obtain θ, and then apply the bound of w_1 in Proposition 2, and then use $\tilde{w}_1 = w_1 - a_1\tilde{w}_2$ to obtain the bound for the unidentified $\tilde{w}_1$ in the current covariate linear contextual model.*

Proof of Proposition 3:*

This is derived straightforwardly from Proposition 2* and we omit the details. We only pointing out how the relation of the un-tilded parameters are related to the tilded parameters due to the particular line segment A that we chose here. In particular for relation $\tilde{w}_1 = w_1 - a_1\tilde{w}_2$, we obtain it from from Assumption 1* which states that $E(\beta_i^w|X_i, Z_i) = \tilde{w}_0 + \tilde{w}_1 X_i + \tilde{w}_2 Z_i$. Then in the set A we plug in (*) $Z_i = a_0 + a_1 X_i$ and obtain $E(\beta_i^w|X_i, Z_i = a_0 + a_1 X_i) = w_0 + w_1 X_i$ where we have $w_0 = \tilde{w}_0 + a_0\tilde{w}_2$ and $w_1 = \tilde{w}_1 + a_1\tilde{w}_2$. The second result implies $\tilde{w}_1 = w_1 - a_1\tilde{w}_2$. Likewise, plug (*) into the five parameter-regression equation, we obtain $E(T_i|X_i, Z_i) = \tilde{w}_0 + \tilde{c}_1 X_i + \tilde{w}_2 Z_i + \tilde{d}_1 X_i^2 + \tilde{d}_2 X_i Z_i = w_0 + c_1 X_i + d_1 X_i^2$, with $w_0 = \tilde{w}_0 + a_0\tilde{w}_2$, $c_1 = \tilde{c}_1 + a_1\tilde{w}_2 + a_0\tilde{d}_2$, and $d_1 = \tilde{d}_1 + a_1\tilde{d}_2$. Q.E.D.

Remark 2* *(Estimation of the district parameter B.) Note that*

$$B = \sum_i N_i X_i \beta_i^b / \sum_i N_i X_i = \sum_i N_i X_i (\tilde{b}_0 + \tilde{b}_1 X_i + \tilde{b}_2 Z_i + e_i^b) / \sum_i N_i X_i, \tag{15}$$

where $e_i^b = \beta_i^b - E(\beta_i^b|X_i, Z_i)$. These parameters parameters can be related to the five-parameter regression (9) and (10) by

$$\tilde{b}_0 = \tilde{w}_0 + \tilde{c}_1 - \tilde{w}_1,\ \tilde{b}_1 = \tilde{w}_1 + \tilde{d}_1,\ \tilde{b}_2 = \tilde{w}_2 + \tilde{d}_2. \tag{16}$$

Then B can be expressed as

$$B = B(\tilde{w}_1, \theta) + \frac{\sum_i N_i X_i e_i^b}{\sum_i N_i X_i}, \tag{17}$$

where the last term is now an average of iid mean zero random variables if Z_i includes information of N_i, since $e_i^b = \beta_i^b - E(\beta_i^b|X_i, Z_i)$. The first term is then an unbiased for

B, and is expressed as

$$B(\tilde{w}_1, \theta) = \tilde{w}_0 + \tilde{c}_1 + \frac{\sum_i N_i X_i^2}{\sum_i N_i X_i}\tilde{d}_1 + \frac{\sum_i N_i X_i Z_i}{\sum_i N_i X_i}(\tilde{w}_2 + \tilde{d}_2) - \frac{\sum_i N_i X_i(1 - X_i)}{\sum_i N_i X_i}\tilde{w}_1. \quad (18)$$

This is linear in the five parameters that can be determined from the regression equation (9), as well as in the unidentified $\tilde{w}_1$. The unidentified $\tilde{w}_1$ can be bounded by Proposition 3.*

Remark 3* *(Confidence interval for covariate linear contextual model.) Now we briefly comment on how to obtain the confidence interval when applying the bounds of Proposition 3*. This proposition provides bounds of $\tilde{w}_1$ that are similar to the bounds of w_1 in Proposition 2 in the main text. The upper bound (or respectively the lower-bound) of w_1 is a minimum (or respectively maximum) over four linear combinations of $(1, \theta^T)$. Similarly, The upper bound (or respectively the lowerbound) of $\tilde{w}_1$ is a minimum (or respectively maximum) over four linear combinations of $(1, \psi^T)$. The conservative confidence intervals for the district parameter B, related to these bounds, can be derived similarly. In particular, for the formulas in Proposition 4 of the main text,*

- *DD, x, r, S_1, J remain unchanged,*
- *$\hat{\theta}$ should be replaced by $\hat{\psi}$ from the five-parameter regression,*
- *V should be replaced by the asymptotic variance estimate of $\hat{\psi}$,*
- *$h_0, h, gl_0, gl_j's, gu_0, gu_j's$ should be replaced respectively by $\tilde{h}_0, \tilde{h}, \tilde{g}l_0, \tilde{g}l_j's, \tilde{g}u_0, \tilde{g}u_j's$ to be defined below,*
- *$\tilde{h}_0 = h_0$, $\tilde{g}l_0 = gl_0$, $\tilde{g}u_0 = gu_0$,*
- *$h^T = \frac{\sum_i N_i X_i (1,1,X_i,Z_i,Z_i)}{\sum_i N_i X_i}$,*
- *$\tilde{g}l_j^T = gl_j^T D - a_1 1_{\tilde{w}_2}^T$, $\tilde{g}u_j^T = gu_j^T D - a_1 1_{\tilde{w}_2}^T$, where a_1 is the tuning parameter in the set A in Proposition 3*, $1_{\tilde{w}_2}^T = (0, 0, 0, 1, 0)$, D is the 3×5 matrix such that (14) can be expressed as $\theta = D\psi$.*

Remark 4* *(Specification of a random set of A where we believe the linear contextual assumptions hold.) Our confidence intervals are derived for a set A (where we* *believe that the linear contextual assumptions hold, whether with or without covariates) that is pre-determined and non-stochastic. When they are data related, for example when the parameters l, u or a_0, a_1 need to be estimated from data somehow to form a random estimated version of A, deriving confidence interval for B may still be possible by finding the joint asymptotic distribution of the estimates of all the parameters, including those for A (such as l, u) and those for the regression model of T (i.e., θ or ψ). This will be more complicated and may not be necessary. In practice, we expect our proposed confidence intervals to remain valid even when ignoring the randomness in A, if it is chosen conservatively, i.e., if we believe that the linear contextual assumptions actually hold in a larger set that contains A with overwhelming probability (with probability tending to 1 as the number of precincts increases to infinity).*