Problem Set 7[1]
Due 11/2

This exercise examines the following research question: What is the effect of maternal smoking during pregnancy on infant birth weight and death? The required reading for this problem set is Almond, Chay, and Lee (2005). The data extract, "lbw.dta", is from the 1989 Linked National Natality-Mortality Details Files, which are an annual census of births in the U.S., derived from Certificates of Live Birth. Information on subsequent infant death within a year of birth is derived from Death Certificates. This extract consists of all births in Pennsylvania in 1989. The observational unit of the data is the mother-infant outcome match. There are 139,149 observations and 32 variables. For this problem set, observations with missing values for any of the variables below were dropped from the original sample (about 17%).

The key variables are:

dbirwt	birth weight of the infant (in grams)
death	indicator equal to one if the infant died within one year of birth and zero, otherwise
tobacco	indicator equal to one if the mother smoked during pregnancy and zero, otherwise

The relevant control variables are:

- Mother's attributes:

  dmage (mother's age), dmeduc (mother's educational attainment), mblack (indicator=1 if mother is Black), motherr (=1 if neither Black nor White), mhispan (=1 if Hispanic), dmar (=1 if mother is unmarried), foreignb (=1 if mother is foreign born)

- Father's attributes:

  dfage (father's age), dfeduc (father's educational attainment), fblack (indicator=1 if father is Black), fotherr (=1 if neither Black nor White), fhispan (=1 if Hispanic)

- Other risky behavior:

  alcohol (indicator=1 if mother drank alcohol during pregnancy), drink (number of drinks per week)

- Medical care:

  tripre1, tripre2, tripre3 (indicators=1 if 1st prenatal care visit in 1st, 2nd, or 3rd trimester, respectively), tripre0 (=1 if no prenatal care visits), nprevist (total number of prenatal care visits)

- Pregnancy history and maternal health:

  first (=1 if first-born), dlivord (birth order), deadkids (number previous births where newborn died), dislb (months since last birth), preterm (=1 if previous birth premature or small for gestational age), pre4000 (=1 if previously had $>$ 4000 gram newborn), plural (=1 if twins or greater birth), phyper (=1 if mother had pregnancy associated hypertension), diabete (=1 if mother diabetic), anemia (=1 if mother anemic)

- 410 observations were dropped from the data (using Stata's "duplicates drop" command) due to duplicates in terms of all variables.

- Notation:

  y_{1i} : infant-mother-pair i's potential outcome in the treatment (T) state
  (infant's birth weight if mother smokes during pregnancy)

  y_{0i} : infant-mother-pair i's potential outcome in the control (C) state
  (infant's birth weight if mother doesn't smoke during pregnancy)

  y_i : infant-mother-pair i's observed outcome (infant's birth weight)

  d_i : infant-mother-pair i's treatment status $= \begin{cases} 1 & \text{(infant's mother smokes during pregnancy)} \\ 0 & \text{(infant's mother doesn't smoke during pregnancy)} \end{cases}$

[1] This problem set requires that you download a user-written command "pscore" and its associated files from this website http://www.stata-journal.com/software/sj5-3 under the folder name "st0026_2." Equivalently, you can type in "net search pscore" in *Stata*, and then click the link "st0026_2 from http://www.stata-journal.com/software/sj5-3" shown on the *Stata* result window.

- The observables ($\mathbf{x}$), which are determined prior to pregnancy, are: mother's and father's age, education, and race, marital status, immigrant status; indicators for alcohol use, and indicators for medical risk factors; prenatal care usage, number of previous live births and terminations, months since last birth, indicators for previous births over 4000 grams.
- Under selection-on-observables, we have the following results:

$$\begin{aligned} f(y_j|d) &\neq f(y_j) \text{ but } f(y_j|d,\mathbf{x}) = f(y_j|\mathbf{x}), \text{ where } j = 0,1 \\ \text{that is, } (y_0, y_1) \amalg d|\mathbf{x} &\Rightarrow \text{ ATE identifiable (and ATT, ATC as well)} \\ f(y_0|d) &\neq f(y_0) \text{ but } f(y_0|d,\mathbf{x}) = f(y_0|\mathbf{x}) \\ \text{that is } y_0 \amalg d|\mathbf{x} &\Rightarrow \text{ only ATT identifiable} \\ f(y_1|d) &\neq f(y_1) \text{ but } f(y_1|d,\mathbf{x}) = f(y_1|\mathbf{x}) \\ \text{that is } y_1 \amalg d|\mathbf{x} &\Rightarrow \text{ only ATC identifiable} \end{aligned}$$

a. Under what conditions can one identify the average treatment effect of maternal smoking by comparing the unadjusted difference in mean birth weight of infants of smoking and non-smoking mothers? Under the assumption that maternal smoking is randomly assigned, estimate its impact on birth weight. Provide some evidence for or against the hypothesis that maternal smoking is randomly assigned.

- Random assignment of maternal smoking during pregnancy, **unconditional** on observables, is the "gold standard" solution to identifying ATE by comparing the unadjusted difference in mean birth weight of infants of smoking and non-smoking mothers.

$$\begin{aligned} f(y_j|d) &= f(y_j), \text{ where } j = 0,1 \\ &\Rightarrow \text{ ATE} \equiv \theta = \mathbb{E}(y_1 - y_0) = \mathbb{E}(y_1 - y_0|d) = \mathbb{E}(y|d=1) - \mathbb{E}(y|d=0) \\ &\Rightarrow \widehat{\theta}_{\text{ATE}} = \frac{1}{N_1}\sum_{i=1}^{N_1} y_{i\in T} - \frac{1}{N_0}\sum_{j=1}^{N_0} y_{j\in C} = \overline{y}_{i\in T} - \overline{y}_{j\in C} = \overline{y}_1 - \overline{y}_0 \end{aligned}$$

- Assuming $d \amalg (y_1, y_0)$, we have $\widehat{\theta}_{\text{ATE}} = \overline{y}_1 - \overline{y}_0 = 3168.036 - 3425.658 = -257.622$. Smoking mothers have babies that weigh 258 grams less, on average, than the babies of non-smoking mothers.

Table 1: Unadjusted Mean Comparison

Variable	**Coefficient**	**(Std. Err.)**
tobacco	-257.622**	(3.993)
Intercept	3425.658**	(1.700)

- *Stata* code: reg dbirwt tobacco, vce(robust)
- It is self-selection into treatment (smoking) that easily invalidates the random assignment of maternal smoking during pregnancy. Some women might quit smoking during pregnancy to avoid infant's low birth weight (LBW). The following mean comparison t-tests of the observables between the treated and the control group provide evidence against the claimed random assignment of maternal smoking during pregnancy.

Rejection rate of equal means	$\alpha = 0.05$	$\alpha = 0.01$
Mother's characteristics	100%	100%
Father's characteristics	100%	100%
Other risky behaviors	100%	100%
Medical care	100%	100%
Pregnancy history and maternal health	90%	100%

b. Suppose that maternal smoking is randomly assigned conditional on the other observable determinants of infant birth weight.

(1) What does this imply about the relationship between maternal smoking and unobservable determinants of birth weight conditional on the observables?

- This implies "selection (only) on the observables": maternal smoking is independent of (or at least uncorrelated with) unobservable determinants of birth weight once conditional on the observables.

(2) Use a basic linear regression model to estimate the impact of smoking and report your estimates.

- *Stata* code: reg dbirwt tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids disllb preterm pre4000 plural phyper diabete anemia, vce(robust)
- $\widehat{\theta}_{\text{ATE}} = -218.166$. Note that after adjusting for overt bias, the effect (in magnitude) becomes smaller. See Table 2 for details.

Table 2: Linear Regression Model

Variable	**Coefficient**	**(Std. Err.)**
tobacco	-218.166**	(3.977)
dmage	-2.822**	(0.487)
dmeduc	5.604**	(0.929)
mblack	-155.308**	(13.691)
motherr	-79.949**	(19.818)
mhispan	-60.458**	(14.119)
dmar	-38.519**	(4.557)
foreignb	-12.482	(8.676)
dfage	-0.237	(0.365)
dfeduc	4.030**	(0.846)
fblack	-55.625**	(13.440)
fotherr	-107.689**	(19.379)
fhispan	-72.711**	(13.219)
alcohol	-24.524†	(13.808)
drink	-7.896**	(2.761)
tripre1	38.785†	(22.049)
tripre2	113.886**	(21.642)
tripre3	230.072**	(22.145)
tripre0	0.000	(0.000)
nprevist	31.992**	(0.647)
first	-85.384**	(5.053)
dlivord	33.591**	(1.966)
deadkids	-12.167**	(2.243)
disllb	-0.183**	(0.064)
preterm	-522.642**	(15.136)
pre4000	468.431**	(13.569)
plural	-953.617**	(13.536)
phyper	-171.246**	(11.336)
diabete	45.301**	(13.012)
anemia	-37.578*	(15.434)
Intercept	3012.185**	(24.651)

(3) Under what conditions is the average treatment effect identified?

- The **potential** outcomes in the treated and in the control state are **independent** of the treatment conditional on **observables**, that is, $f(y_j|d) \neq f(y_j)$ but $f(y_j|d, \mathbf{x}) = f(y_j|\mathbf{x})$, where $j = 0, 1$, or written as $(y_0, y_1) \amalg d|\mathbf{x}$. (To be strict, we also need: 1) ATE is additive; 2) ATE is homogeneous (constant) or heterogeneous with the heterogeneity independent of observables; and 3) the conditional expectation of y given $\mathbf{x}$ is linear in $\mathbf{x}$, that is, no specification errors.)

c. Under the assumption of random assignment conditional on the observables:

(1) What are the sources of misspecification bias in the estimates generated by the linear model estimated in part (b)?

- Consider the following:

$$\begin{aligned} y &= dy_1 + (1-d)y_0 \\ &= d[\mathbb{E}(y_1|\mathbf{x}) + u_1] + (1-d)[\mathbb{E}(y_0|\mathbf{x}) + u_0] \\ &= \mathbb{E}(y_0|\mathbf{x}) + d[\mathbb{E}(y_1|\mathbf{x}) - \mathbb{E}(y_0|\mathbf{x})] + u \end{aligned}$$

Let $u \equiv du_1 + (1-d)u_0$, $\mathbb{E}(y_0|\mathbf{x}) = \gamma_0 + g(\mathbf{x})$, and $\mathbb{E}(y_1|\mathbf{x}) = \gamma_1 + g(\mathbf{x})$. We have the following:

$$y = \gamma_0 + g(\mathbf{x}) + d(\gamma_1 - \gamma_0) + u.$$

And ATE (θ) is identified by:

$$\begin{aligned} \theta &= \mathbb{E}(y_1 - y_0) = \mathbb{E}[\mathbb{E}(y_1 - y_0|\mathbf{x})] \\ &= \mathbb{E}[\mathbb{E}(y_1|\mathbf{x}) - \mathbb{E}(y_0|\mathbf{x})] = \gamma_1 - \gamma_0. \end{aligned}$$

Therefore, we have

$$y = \gamma_0 + \underbrace{g(\mathbf{x})}_{\text{overt bias}} + \underset{\text{exogenous}}{d} \cdot \underbrace{\theta}_{\text{ATE}} + u$$

If we know $g(\cdot)$, then θ can be easily identified with a correctly specified regression. We require $\mathbb{E}(du|g(\mathbf{x})) = 0$ so that d is exogenous in the regression. However, if we only use a linear regression, then we are imposing a strong assumption (belief) on the functional form of $g(\mathbf{x})$, which is assumed to be $\mathbf{x}'\beta$. Then for the following model,

$$y = \gamma_0 + \mathbf{x}'\beta + d\theta + u,$$

we must have $\mathbb{E}(du|\mathbf{x}'\beta) = 0$ so that d is exogenous in the regression. Because

$$\mathbb{E}(du|g(\mathbf{x})) = 0 \not\Rightarrow \mathbb{E}(du|\mathbf{x}'\beta) = 0 \text{ if } g(\mathbf{x}) \neq \mathbf{x}'\beta,$$

we will have bias in estimating θ due to misspecifying $g(\mathbf{x})$ in the linear regression. For example, excluded nonlinear terms in $\mathbf{x}$ may generate OVB.

(2) Use an approach in the spirit of multivariate matching, that is, estimate the smoking effects using a flexible functional form for the control variables (e.g., higher order terms and interactions).

- *Stata* code: reg dbirwt tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids disllb preterm pre4000 plural phyper diabete anemia dmage2 dmage3 dmeduc2 dfage2 dfage3 dfeduc2 dmage_dmeduc dmage_mblack dmage_motherr dmage_mhispan dmage_dmar dfage_dfeduc dfage_fblack dfage_fotherr dfage_fhispan dmage_alcohol dmage_drink, vce(robust)
- $\widehat{\theta}_{\text{ATE}} = -216.043$. See Table 3 for details.
- Since we do not know the functional form of $g(\cdot)$, we simply kitchen-sink the regression by including as many terms as we can, wishing for cleaning up specification errors and OVB.

(3) What are the benefits and drawbacks to this approach?

- Without knowing the true functional form of $g(\cdot)$, the benefits come from the *wish* that this approach could reduce specification errors (by including many nonlinear terms of observables) and avoid OVB (since the likelihood of missing any important factors might be lowered when more and more observables are controlled for). The drawbacks of this approach are multifold: 1) it could **not** really solve the OVB problem if the omitted variable is never observable or hardly measurable (e.g., ability, effort, etc); 2) it may induce multicollinearity problems (reducing variation in the treatment, maternal smoking, and making the estimate less precise); and 3) it could exacerbate the attenuation bias due to measurement errors in any of these control variables.

Table 3: Multivariate Matching

Variable	Coefficient	(Std. Err.)
tobacco	-216.043**	(4.016)
dmage	13.338	(14.891)
dmeduc	7.138	(6.285)
mblack	-110.215**	(33.824)
motherr	-56.785	(70.494)
mhispan	-92.723†	(52.617)
dmar	47.101*	(22.109)
foreignb	-15.390†	(8.766)
dfage	8.334	(6.577)
dfeduc	4.318	(5.817)
fblack	-57.876†	(29.626)
fotherr	-85.118	(70.869)
fhispan	-75.718†	(45.238)
alcohol	-78.776	(74.588)
drink	17.738	(14.503)
tripre1	37.174†	(22.088)
tripre2	113.653**	(21.678)
tripre3	228.148**	(22.180)
tripre0	0.000	(0.000)
nprevist	32.066**	(0.648)
first	-84.353**	(5.339)
dlivord	32.486**	(2.041)
deadkids	-12.014**	(2.244)
disllb	-0.135*	(0.066)
preterm	-522.231**	(15.129)
pre4000	468.237**	(13.555)
plural	-953.915**	(13.526)
phyper	-170.683**	(11.339)
diabete	45.504**	(13.016)
anemia	-36.343*	(15.442)
dmage2	-0.053	(0.541)
dmage3	-0.001	(0.006)
dmeduc2	0.689**	(0.230)
dfage2	-0.211	(0.192)
dfage3	0.002	(0.002)
dfeduc2	0.152	(0.194)
dmage_dmeduc	-0.705**	(0.173)
dmage_mblack	-1.874	(1.280)
dmage_motherr	-0.898	(2.486)
dmage_mhispan	1.420	(2.110)
dmage_dmar	-3.250**	(0.890)
dfage_dfeduc	-0.129	(0.122)
dfage_fblack	0.138	(0.970)
dfage_fotherr	-0.748	(2.225)
dfage_fhispan	0.201	(1.593)
dmage_alcohol	2.020	(2.621)
dmage_drink	-0.961†	(0.532)
Intercept	2657.069**	(136.299)

d. Describe the propensity score approach to the problem of estimating the average treatment effect of smoking when the treatment is randomly assigned conditional on the observables. How does it reduce the dimensionality problem of multivariate matching?

- Read the lecture notes.
- The key points are here. We may identify ATE in three ways: (1) regression, where we must correctly specify the functional form for $g(\cdot)$; (2) matching, where we need to compare individuals in the treated and in the untreated groups with similar $\mathbf{x}$; or (3) weighting, where we must estimate the propensity score $p(\mathbf{x})$ correctly, which is almost equivalent to getting $g(\cdot)$ right in the regression approach. For approach (2), we can use $p(\mathbf{x})$ to make matching only on one dimension instead of multi-dimensions. The identification of ATE comes from:

$$\begin{aligned}&\text{(1) selection-on-observables: } \mathbb{E}(y_j|d,\mathbf{x}) = \mathbb{E}(y_j|\mathbf{x}) \Rightarrow \mathbb{E}(y_j|d,p(\mathbf{x})) = \mathbb{E}(y_j|p(\mathbf{x}));\\ &\text{(2) } \theta = \mathbb{E}(y_1 - y_0) = \mathbb{E}[\mathbb{E}(y_1 - y_0|p(\mathbf{x}))] = \mathbb{E}[\underbrace{\mathbb{E}(y|d=1,p(\mathbf{x})) - \mathbb{E}(y|d=0,p(\mathbf{x}))}_{\text{matching on } p(\mathbf{x}) \text{ (a scalar) instead of } \mathbf{x} \text{ (a vector)}}] \quad .\end{aligned}$$

e. Implement the propensity score approach (hints below) to the evaluation problem using two methods:

- Below is my Stata code for estimating the propensity score without specifying the number of blocks at the beginning. Stata generated 33 blocks in the end. For balancing results, please see the attached log file.

  pscore tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids dislb preterm pre4000 plural phyper diabete anemia dmage2 dmeduc2 dfage2 dfeduc2 dmage_dmeduc dmage_mblack dmage_motherr dmage_mhispan dmage_dmar dfage_dfeduc dfage_fblack dfage_fotherr dfage_fhispan dmage_alcohol dmage_drink, pscore(pscore) blockid(block) logit level(0.01)

- Below is my Stata code for estimating the propensity score with 100 blocks.

  pscore tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids dislb preterm pre4000 plural phyper diabete anemia dmage2 dmeduc2 dfage2 dfeduc2 dmage_dmeduc dmage_mblack dmage_motherr dmage_mhispan dmage_dmar dfage_dfeduc dfage_fblack dfage_fotherr dfage_fhispan dmage_alcohol dmage_drink, pscore(pscore100) blockid(block100) logit level(0.001) numblo(101)

- Below is my Stata code for estimating the propensity score with 200 blocks.

  pscore tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids dislb preterm pre4000 plural phyper diabete anemia dmage2 dmeduc2 dfage2 dfeduc2 dmage_dmeduc dmage_mblack dmage_motherr dmage_mhispan dmage_dmar dfage_dfeduc dfage_fblack dfage_fotherr dfage_fhispan dmage_alcohol dmage_drink, pscore(pscore200) blockid(block200) logit level(0.005) numblo(201)

(1) Method 1: control directly for the estimated propensity scores in a regression model.

- If we use the following regression model,

$$y = \beta_0 + d\theta + \beta_1 \widehat{p}(\mathbf{x}) + u,$$

  we will ignore the heterogeneity in the effect due to the interaction between d and $\mathbf{x}$. Here we get $\widehat{\theta}_{\text{ATE}} = -212.571$. See Table 4 for details.

Table 4: Control directly for the propensity scores I

Variable	**Coefficient**	**(Std. Err.)**
tobacco	-212.571**	(4.312)
pscore	-316.255**	(12.137)
Intercept	3476.398**	(2.510)

- If we take the heterogeneity in the effect due to the interaction between d and $\mathbf{x}$ into account, then the following regression model should be used:

$$y = \beta_0 + d\theta + \beta_1 \widehat{p}(\mathbf{x}) + \beta_2 d(\widehat{p}(\mathbf{x}) - \overline{\widehat{p}(\mathbf{x})}) + u.$$

Here we get $\widehat{\theta}_{\text{ATE}} = -207.944$. See Table 5 for details.

Table 5: Control directly for the estimated propensity score II

Variable	Coefficient	(Std. Err.)
tobacco	-207.944**	(4.576)
pscore	-294.852**	(14.714)
Tpscore	-66.289*	(26.000)
Intercept	3472.964**	(2.812)

(2) Method 2: use the estimated propensity score in a classification scheme to stratify the sample.

- (1) Stratify the sample by $\widehat{p}(\mathbf{x})$ into 33 blocks ($j = 1, \cdots, J = 33$). Within each block (j), we assume selection-on-observables only for y_0:

$$\mathbb{E}(y_{0j}|d_j, p_j(\mathbf{x})) = \mathbb{E}(y_{0j}|p_j(\mathbf{x})).$$

Hence, a simple group mean comparison in each block (j) between the treated and the control group, conditional on the propensity score, will identify ATT.

$$\begin{aligned}
&\mathbb{E}(y_j|d_j = 1, p_j(\mathbf{x})) - \mathbb{E}(y_j|d_j = 0, p_j(\mathbf{x})) \\
&= \mathbb{E}(y_{1j}|d_j = 1, p_j(\mathbf{x})) - \mathbb{E}(y_{0j}|d_j = 0, p_j(\mathbf{x})) \\
&= \mathbb{E}(y_{1j}|d_j = 1, p_j(\mathbf{x})) - \mathbb{E}(y_{0j}|d_j = 1, p_j(\mathbf{x})) \\
&= \mathbb{E}(y_{1j} - y_{0j}|d_j = 1, p_j(\mathbf{x})) \\
&= \theta_{j\text{ATT}} \text{ (average effect of treatment on the treated for block } j\text{)}.
\end{aligned}$$

What is doing here is essentially requiring us to find an individual in the control group to match (based on the similar propensity score) a chosen one in the treated group, that is, to look for counterfactuals for the treated from the control group.

(2) If we look for counterfactuals for each treated subject (in the treatment group) from the control group, we will identify ATT instead of ATE. Using the law of iterated expectation, we can identify ATT by a classification scheme based on the stratified sample.

$$\begin{aligned}
\theta_{\text{ATT}} &= \mathbb{E}(y_1 - y_0|d = 1) = \mathbb{E}[\mathbb{E}(y_{1j} - y_{0j}|d = 1, p_j(\mathbf{x}))] \\
&= \mathbb{E}[\mathbb{E}(\theta_{j\text{ATT}}|d = 1, p_j(\mathbf{x}))].
\end{aligned}$$

(3) The estimation procedure is the following:

$$\begin{aligned}
\widehat{\theta}_{j\text{ATT}} &= \overline{y}_{1j} - \overline{y}_{0j} \\
&= \frac{\sum_{i\in j} d_i y_i}{\sum_{i\in j} d_i} - \frac{\sum_{i\in j}(1-d_i)y_i}{\sum_{i\in j}(1-d_i)} \text{ (for each block } j\text{)} \\
\widehat{\theta}_{\text{ATT}} &= \sum_{j=1}^{J} \left(\frac{\sum_{i\in j} d_i}{N_1} \widehat{\theta}_{j\text{ATT}} \right) \\
&= \sum_{j=1}^{J} \left(\frac{\sum_{i\in j} d_i}{N_1} \right) \left[\frac{\sum_{i\in j} d_i y_i}{\sum_{i\in j} d_i} - \frac{\sum_{i\in j}(1-d_i)y_i}{\sum_{i\in j}(1-d_i)} \right] \\
&= \overline{y}_1 - \sum_{j=1}^{J} \left(\frac{\sum_{i\in j} d_i}{N_1} \right) \overline{y}_{0j} \\
&= -211.285.
\end{aligned}$$

A simple Stata command for this estimator based on stratified sample using the estimated propensity score is "atts." To use this command, you need to obtain the estimated propensity scores using "pscore" first.

- *Stata* code: atts dbirwt tobacco, pscore(pscore) blockid(block)

Table 6: Use propensity scores in classification scheme

Variable	Coefficient	(Analy. Std. Err.)
tobacco	-211.285**	(4.825)

(3) Provide empirical evidence that your implementation is reasonable and evidence on the overlap of the observables of smokers and non-smokers. Present your findings and interpret the results. (This is an open-ended question, so show me what you know and be creative and thoughtful.)

- The following "box-and-whisker" plots (or simply called box plots) show some evidence on the overlap of the observables of smokers and non-smokers through estimated propensity scores. The box ends at the quartiles (25%-tile and 75%-tile), extending the "whiskers" to the farthest points that are not outliers (5%-tile and 95%-tile). The statistical median is a horizontal line in the box. The overlap in plots implies similarities of the observables in the treated and the control groups. Three cases are investigated: data are stratified into 33, 100 and 200 blocks by estimated propensity scores. Results are consistent. See Figure 1, Figure 2, and Figure 3.

- To show propensity scores are reasonably estimated, we compare the fraction of the treated within each block with estimated propensity scores. Because of selection-on-observables, the probability of being treated is given by the propensity score, which stratify data into blocks and balances the observables between the treated and the untreated in each block. Within each block, the fraction of the treated should approximate the associated propensity score (resulting from the binomial distribution). Accordingly, scatter plots should resemble a 45-degree line from the origin. The better the propensity score is estimated, the less deviation of the scatter plots is from the 45-degree line. Again, three cases are investigated: data are stratified into 33, 100 and 200 blocks by estimated propensity scores. It is noticeable that propensity score is less well estimated at the upper tail of the distribution! It should not be surprising because it should be more difficult to find a non-smoker in the control group to match a smoker in the treated group when both have very high propensity to smoke under the selection-on-observables assumption. See Figure 4, Figure 5 and Figure 6 for illustration.

f. Use the estimated propensity scores to reweigh the outcomes of non-smokers and estimate the average treatment effect.

- When using propensity score to reweigh the outcomes of non-smokers, we are identifying ATT. In this case, we do not need to reweigh the outcomes of smokers:

$$\mathbb{E}(y_1 - y_0|d=1) = \underbrace{\mathbb{E}(y|d=1)}_{\text{factual}} - \underbrace{\mathbb{E}(y_0|d=1)}_{\text{counterfactual}} = \frac{1}{\Pr(d=1)}\mathbb{E}\left(\frac{d-p(x)}{1-p(x)}y\right).$$

The estimator for ATT and the estimate based on the estimated propensity score (with 33 blocks) are:

$$\begin{aligned}\widehat{\theta}_{\text{ATT}} &= \frac{1}{N_1}\sum_{i=1}^{N}\frac{d_i - \widehat{p}(\mathbf{x}_i)}{1-\widehat{p}(\mathbf{x}_i)}y_i = \overline{y}_1 - \frac{1}{N_1}\sum_{i=1}^{N}\frac{\widehat{p}(\mathbf{x}_i)}{1-\widehat{p}(\mathbf{x}_i)}(1-d_i)y_i \\ &= -237.506.\end{aligned}$$

(1) Compare the estimates to those in part (e) and interpret your findings.

- See the comparison below.

Controlling for $p(\mathbf{x})$	Treatment Effect $(\widehat{\theta})$
(i) directly as a regressor alone	$\widehat{\theta}_{\text{ATE}} = -212.571$
(ii) (i) + interacted with treatment	$\widehat{\theta}_{\text{ATE}} = -207.944$
(iii) used in a classification scheme	$\widehat{\theta}_{\text{ATT}} = -211.285$
(iv) used as adjustment weights	$\widehat{\theta}_{\text{ATT}} = -237.506$

- Assuming no unobservable factors accounting for smoking choice, that is, no selection-on-unobservables, we find that maternal smoking during pregnancy has significant negative impacts on lowering infant birth weight, on average, by approximately 210 to 240 grams. Such an effect is robust to several controlling schemes ((i), (ii), (iii) and (iv)) of the observables.

(2) What are the benefits and drawbacks of approaches that use the estimated propensity scores as weights?

- See the summary below.

Using $p(\mathbf{x})$	"Gains" besides reducing dimensionality	"Loss" besides violating $d \amalg (y_1, y_0) \vert p(\mathbf{x})$
(i) directly as a regressor alone	get ATE right when it is constant	bias due to misspecifying $p(\mathbf{x})$
(ii) (i) + interacted with treatment	get ATE right when treatment interacted with observables	bias due to misspecifying $p(\mathbf{x})$
(iii) used in a classification scheme	get ATT right when it is heterogeneous	sensitive to balancing procedure and the number of blocks
(iv) used as adjustment weights	might be more efficient than (iii) and get ATT right when it is heterogeneous	weighting is sensitive to $p(\mathbf{x})$, too sensitive at high $p(\mathbf{x})$ which is often poorly estimated

g. A more informative way to describe the birth weight effects of smoking is to estimate the nonparametric conditional mean of birth weight as a function of the estimated propensity score, separately for smokers and non-smokers.

(1) To do this, simply stratify the smokers into 100 equal-sized cells based on their propensity scores and calculate the mean birth weight and propensity score in each cell. Do the same for the non-smokers. Plot these two conditional mean functions, with the mean scores on the x-axis and mean birth weight on the y-axis.

- Use the estimated propensity scores with 100 blocks.

  [pscore tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids disllb preterm pre4000 plural phyper diabete anemia dmage2 dmeduc2 dfage2 dfeduc2 dmage_dmeduc dmage_mblack dmage_motherr dmage_mhispan dmage_dmar dfage_dfeduc dfage_fblack dfage_fotherr dfage_fhispan dmage_alcohol dmage_drink, pscore(pscore100) blockid(block100) logit level(0.001) numblo(101)]

- See Figure 7 generated by the following Stata codes:

  (1) [egen mpscore100 = mean(pscore100), by(block100)]; (2) [bysort block100: egen mdbirwt_t = mean(dbirwt) if tobacco==1]; (3) [bysort block100: egen mdbirwt_c = mean(dbirwt) if tobacco==0]; (4) [label variable mdbirwt_t "smokers"]; (5) [label variable mdbirwt_c "non-smokers"]; (6) [twoway (scatter mdbirwt_t mpscore100, sort msymbol(triangle_hollow)) (scatter mdbirwt_c mpscore100, sort msymbol(circle_hollow)), ytitle(, orientation(horizontal)) ylabel(#7, angle(horizontal)) xtitle(mean propensity scores in 100 equal-sized cells) xlabel(0(0.1)1) scheme(sj)].

(2) Interpret your findings and relate them to the results in part (e) and part (f).

- This graphical analysis explicitly reveals how the bias due to selection on the observables (overt bias) is corrected by the estimated propensity scores. The distance, on y-axis, between the gravity centers of "smokers" and "non-smokers" gives:

$$\widehat{\theta}_{\text{ATE}} = 3168.036 - 3425.658 = -257.622.$$

  This unadjusted mean difference between the treated and the control group (without conditioning on those observables) over-estimates the ATE because it does not take into account the correction on overt bias through propensity scores (along the x-axis). Given the assumption of selection-on-observables, ATE (or ATT) should be identified through the "blockwise" averaging scheme or weighting. For the estimated propensity scores (with 100 blocks), we use the classification scheme illustrated in part (e) to obtain this adjusted estimate:

$$\widehat{\theta}_{\text{ATT}}|_{100\_\text{blocks}} = -211.227.$$

  Therefore, the gain from using the propensity score, which corrects selection bias on observables (overt bias), is:

$$\text{overt bias } = -257.622 - (-211.227) = -46.395.$$

This amount is the averaged discrepancies between two unadjusted means across all blocks.

- It is clear in Figure 7 that propensity score is not good at correcting the selection bias (due to observables) at its upper tail. This indicates that the efficiency of estimating ATT in part (f), can be vulnerable to high propensity scores, let alone its sensitivity to misspecified propensity scores.
- It is clear in Figure 7. that propensity score is not good at correcting the selection bias (due to observables) at its upper tail. This indicates that the efficiency of estimating ATT in part (f), is very vulnerable to high propensity scores, let alone its sensitivity to misspecified propensity scores.

(3) Redo the above using 200 equal-sized cells for smokers and non-smokers.

- Use the estimated propensity scores with 200 blocks.

  [pscore tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids dislllb preterm pre4000 plural phyper diabete anemia dmage2 dmeduc2 dfage2 dfeduc2 dmage_dmeduc dmage_mblack dmage_motherr dmage_mhispan dmage_dmar dfage_dfeduc dfage_fblack dfage_fotherr dfage_fhispan dmage_alcohol dmage_drink, pscore(pscore200) blockid(block200) logit level(0.005) numblo(201)]

- Using this 200 equal-sized blocks as another stratification scheme, we obtain a new adjusted estimate:

$$\widehat{\theta}_{\text{ATT}}|_{200\_\text{blocks}} = -211.701.$$

Therefore, the gain from using the propensity score, which corrects selection bias on observables (overt bias), is:

$$\text{overt bias } = -257.622 - (-211.701) = -45.921.$$

This amount is the averaged discrepancies between two unadjusted means across all blocks.

- See Figure 8 generated by the following Stata codes:

  (1) [egen mpscore200 = mean(pscore200), by(block200)]; (2) [bysort block200: egen mdbirwt_t = mean(dbirwt) if tobacco==1]; (3) [bysort block200: egen mdbirwt_c = mean(dbirwt) if tobacco==0]; (4) [label variable mdbirwt_t "smokers"]; (5) [label variable mdbirwt_c "non-smokers"]; (6) [twoway (scatter mdbirwt_t mpscore200, sort msymbol(triangle_hollow)) (scatter mdbirwt_c mpscore200, sort msymbol(circle_hollow)), ytitle(, orientation(horizontal)) ylabel(#7, angle(horizontal)) xtitle(mean propensity scores in 200 equal-sized cells) xlabel(0(0.1)1) scheme(sj)].

h. Low birth weights (less than 2500 grams) are considered particularly undesirable since they comprise a large share of infant deaths).

(1) Redo part (g) using an indicator for low birth weight as the outcome of interest.

- Since we are interested in the marginal and average effects, we can just use the linear probability model.
- See Figure 9 and Figure 10. [Stata codes are similar to part (g).]

(2) Interpret your findings.

- Using estimated propensity scores in a similar classification scheme to part (g), we obtain an adjusted estimate:

$$\begin{aligned}\widehat{\theta}_{\text{LBW}}|_{100\_\text{blocks}} &= 0.040;\\ \widehat{\theta}_{\text{LBW}}|_{200\_\text{blocks}} &= 0.040.\end{aligned}$$

Therefore, the gain from using the propensity score, which corrects selection bias on observables (overt bias), is:

$$\text{overt bias } = 0.052 - 0.040 = 0.012.$$

Assuming no unobservable factors accounting for smoking choice, we find that maternal smoking during pregnancy has negative impact on birth weight. It, on average, increases the probability of LBW by approximately 4 percentage points.

i. Estimate the impact of maternal smoking on infant death using the methods in parts (a), (b), and (g), using the 100 equal-sized cells for smokers and non-smokers).

- Since we are interested in the marginal and average effects, we can just use the linear probability model.
- Under the assumption that maternal smoking is randomly assigned, estimate its impact on infant death: $\widehat{\theta}_{\text{death}} = 0.003$.

  [reg death tobacco, vce(robust)]

  See Table 7 for details.

Table 7: Unadjusted Mean Comparison

Variable	**Coefficient**	**(Std. Err.)**
tobacco	0.003**	(0.001)
Intercept	0.006**	(0.000)

- Under the assumption that maternal smoking is randomly assigned conditional on the observables: $\widehat{\theta}_{\text{death}} \doteq 0$.

  [reg death tobacco dmage dmeduc mblack motherr mhispan dmar foreignb dfage dfeduc fblack fotherr fhispan alcohol drink tripre1 tripre2 tripre3 tripre0 nprevist first dlivord deadkids disllb preterm pre4000 plural phyper diabete anemia, vce(robust)]

  See Table 8 for details.

- Graphical analysis: see Figure 11 and Figure 12.

  Stata codes for Figure 11:

  (1) [egen mpscore100 = mean(pscore100), by(block100)]; (2) [bysort block100: egen mdeath_t = mean(death) if tobacco==1]; (3) [bysort block100: egen mdeath_c = mean(death) if tobacco==0]; (4) [label variable mdeath_t "smokers"]; (5) [label variable mdeath_c "non-smokers"]; (6) [twoway (scatter mdeath_t mpscore100, sort msymbol(triangle_hollow)) (scatter mdeath_c mpscore100, sort msymbol(circle_hollow)), ytitle(infant mortality) ylabel(0(0.1)1) xtitle(mean propensity scores in 100 equal-sized cells) xlabel(0(0.1)1) scheme(sj)].

  Stata codes for Figure 12:

  (1) [egen mpscore200 = mean(pscore200), by(block200)]; (2) [bysort block200: egen mdeath_t = mean(death) if tobacco==1]; (3) [bysort block200: egen mdeath_c = mean(death) if tobacco==0]; (4) [label variable mdeath_t "smokers"]; (5) [label variable mdeath_c "non-smokers";]; (6) [twoway (scatter mdeath_t mpscore200, sort msymbol(triangle_hollow)) (scatter mdeath_c mpscore200, sort msymbol(circle_hollow)), ytitle(infant mortality) ylabel(0(0.1)1) xtitle(mean propensity scores in 200 equal-sized cells) xlabel(0(0.1)1) scheme(sj)].

(1) Interpret your findings.

- Using estimated propensity scores in a similar classification scheme to part (g), we obtain an insignificant estimate adjusted for overt bias:

$$\begin{aligned} \widehat{\theta}_{\text{death}}|_{100\_\text{blocks}} &= 0.000; \\ \widehat{\theta}_{\text{death}}|_{200\_\text{blocks}} &= 0.000. \end{aligned}$$

  Therefore, the gain from using the propensity score, which corrects selection bias on observables (overt bias), is:

$$\text{overt bias } = 0.003 - 0.000 = 0.003.$$

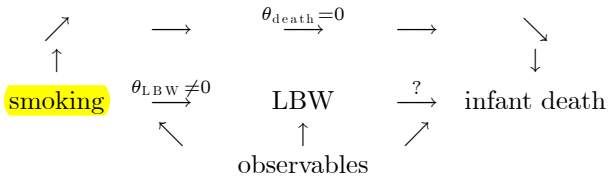
  Assuming no unobservable factors accounting for smoking choice, we find that maternal smoking during pregnancy, on average, has no significant impact on infant mortality. The simple mean comparison of infant mortality between smokers and non-smokers overstates a significant impact by neglecting the selection bias due to observables.

(2) From your results, what might you conclude about the relationship between birth weight and infant death?

Table 8: Linear Regression Model

Variable	**Coefficient**	**(Std. Err.)**
tobacco	0.000	(0.001)
dmage	0.000	(0.000)
dmeduc	0.000	(0.000)
mblack	-0.003	(0.002)
motherr	0.000	(0.003)
mhispan	-0.002	(0.002)
dmar	0.002**	(0.001)
foreignb	-0.004**	(0.001)
dfage	0.000	(0.000)
dfeduc	0.000	(0.000)
fblack	0.004†	(0.002)
fotherr	0.005	(0.003)
fhispan	0.003	(0.002)
alcohol	-0.002	(0.002)
drink	-0.001†	(0.000)
tripre1	-0.024**	(0.006)
tripre2	-0.029**	(0.006)
tripre3	-0.037**	(0.006)
tripre0	0.000	(0.000)
nprevist	-0.002**	(0.000)
first	0.002†	(0.001)
dlivord	0.000	(0.000)
deadkids	0.002**	(0.000)
disllb	0.000	(0.000)
preterm	0.019**	(0.003)
pre4000	-0.002	(0.002)
plural	0.026**	(0.004)
phyper	0.002	(0.002)
diabete	0.004*	(0.002)
anemia	0.003	(0.003)
Intercept	0.052**	(0.006)

- Consider the following causal graph:



  We have found that maternal smoking during pregnancy has a significant negative impact on the birth weight of infants, but we do not find significant impacts of smoking on infant mortality. Therefore, conditional on the observables available to a researcher, the relationship between birth weight and infant death seems to be very weak; otherwise, $\theta_{\text{death}} = 0$ should be statistically significant from 0.

j. Concisely and coherently summarize your results above. Describe the estimated effects of maternal smoking on infant-birth weight and infant mortality, and whether you think your "best" estimate of the effects of smoking is credibly identified. State why or why not.

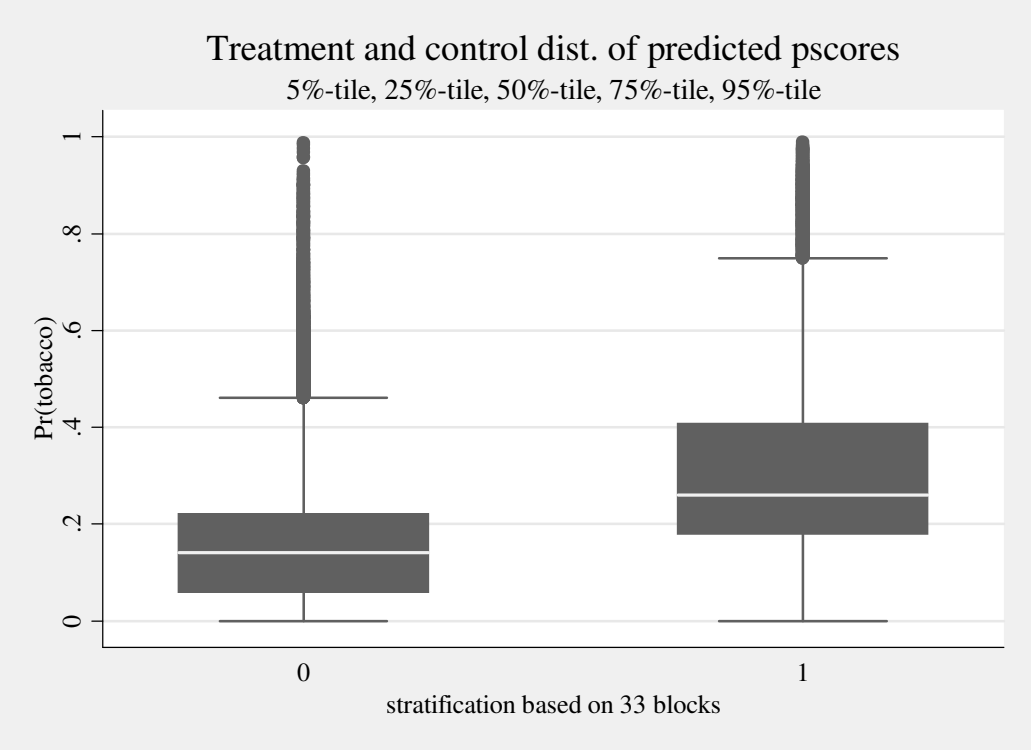
- Comparing birth weights of infants whose mothers smoked during pregnancy with the ones whose mothers did not, we find a mean difference about 260 grams less in the case of maternal smoking. This indicates a negative effect of maternal smoking on infant birth weight. Controlling for variables (observed) which are determined prior to pregnancy and which affect both the treatment (maternal smoking) and the outcome (infant birth weight), we attempt to identify the average effect of maternal smoking on infant birth weight. Assuming the treatment is determined only based upon the observables in the data, we adjust this estimate to roughly 210 grams, which implies an over-estimate (of about 50 grams) of the previous simple mean comparison due to selection of treatment on the observables. Using the same approach and assuming selection only on the observables, we also find maternal smoking could increase the probability of low birth weight (less than 2,500 grams) by approximately 4 percentage point. As to the case of infant death, we cannot reject the null hypothesis that on average there is no impact of maternal smoking on infant mortality.

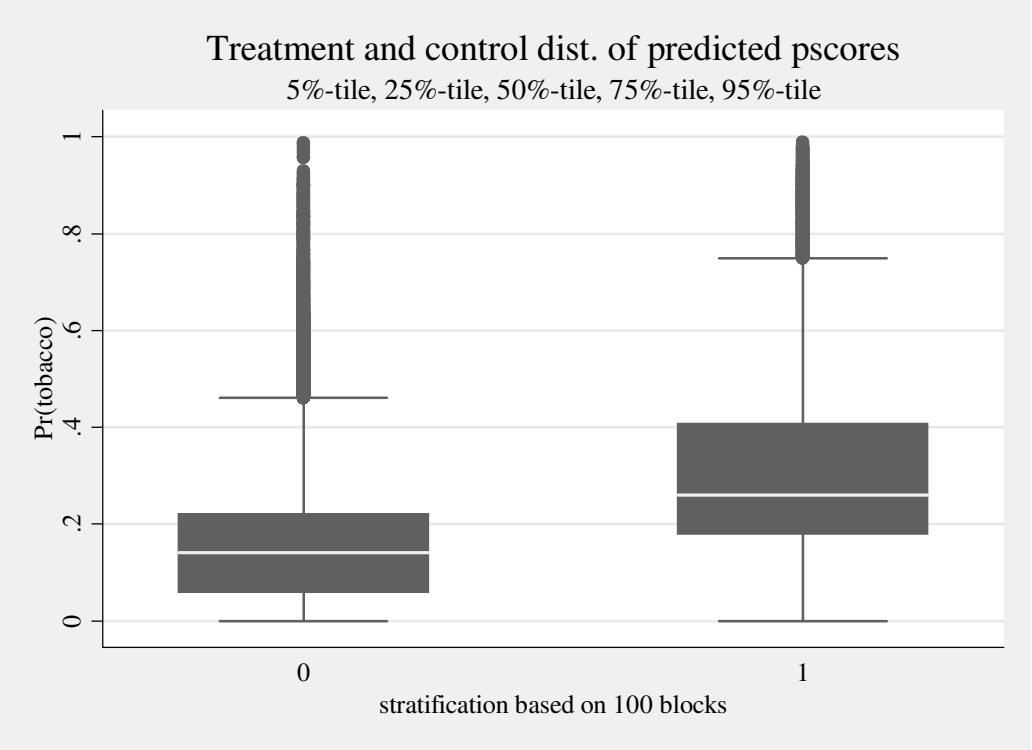
- Whether such treatment effects are identified depends on the validity of the assumption of selection (only) on the observables. If this is true, then the treatment status is purely exogenously after controlling for those observables. Therefore, and obviously, the regression analysis of infant's health outcome is *assumed* free from the endogeneity concern of the smoking choice. The main problem that we need to deal with is the specification of those control variables and possible heterogeneities in treatment effects for people with different treatment propensity. Using propensity score is a reasonable way to overcome the dimensionality problem of matching the treated and the untreated on observables. In addition, either the stratification or weighting scheme suggests some ways to account for the heterogeneities in the treatment effects, which is a more flexible way than doing least squares where a fixed weighting scheme is used rather than a flexible weight based on the propensity score.

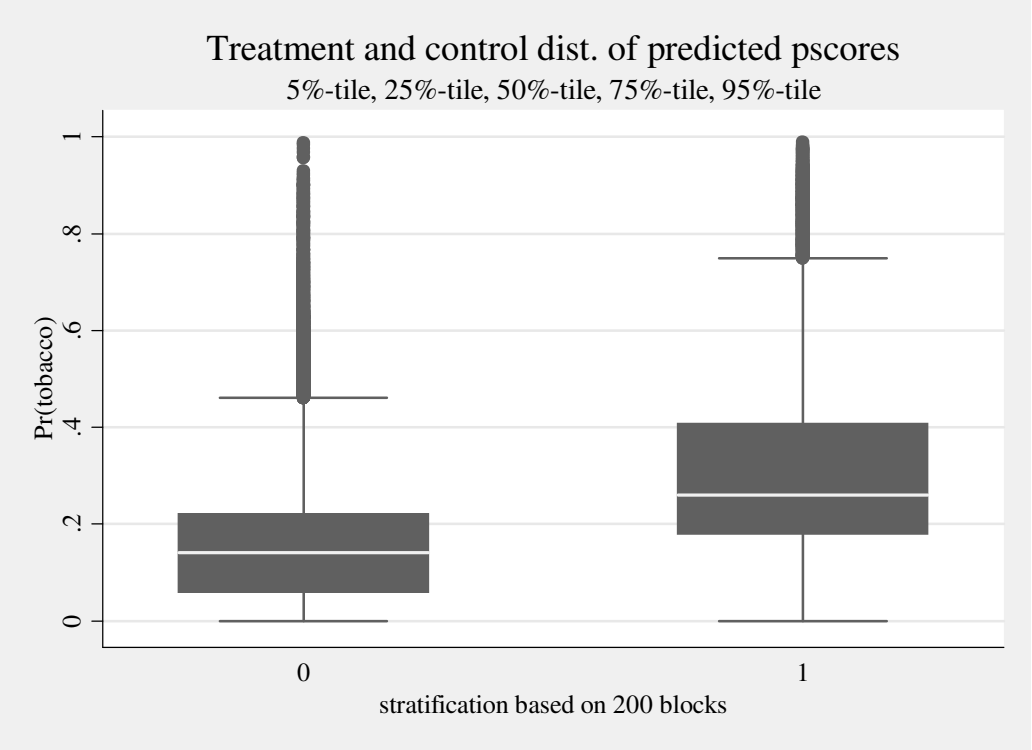
- Whether treatment effects can be identified in this maternal smoking case or in a more general context using the propensity score approach hinges upon the *exogeneity* of the treatment selection. Assuming selection only on the observables, we deny possible impacts on both the outcome and the treatment selection from unobserved factors, which also induce self-selection into the treatment based on potential gains. Making a bold (and heroic) assumption like selection-on-observables will allow us to answer a question of how to get to the destination if we start from the wrong direction.

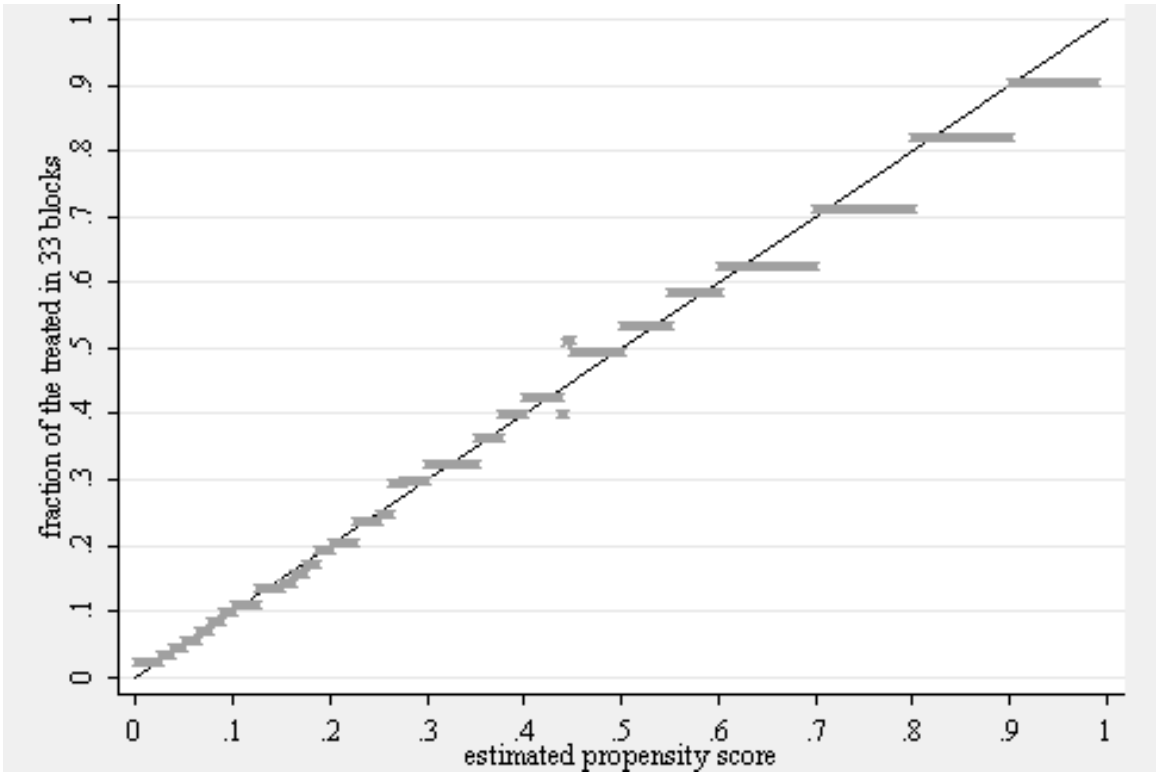
References

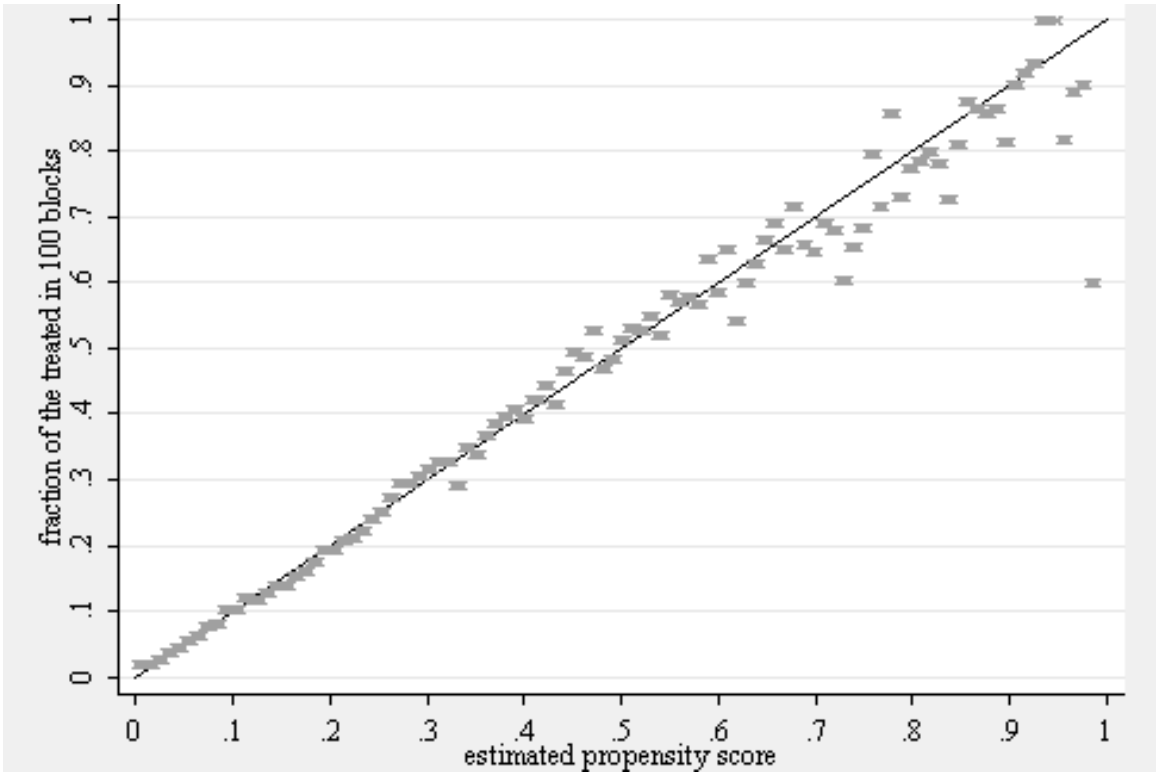
Almond, D., K. Y. Chay, and D. Lee (2005). "The Costs of Low Birth Weight." *Quarterly Journal of Economics* 120(3): 1031-1083.

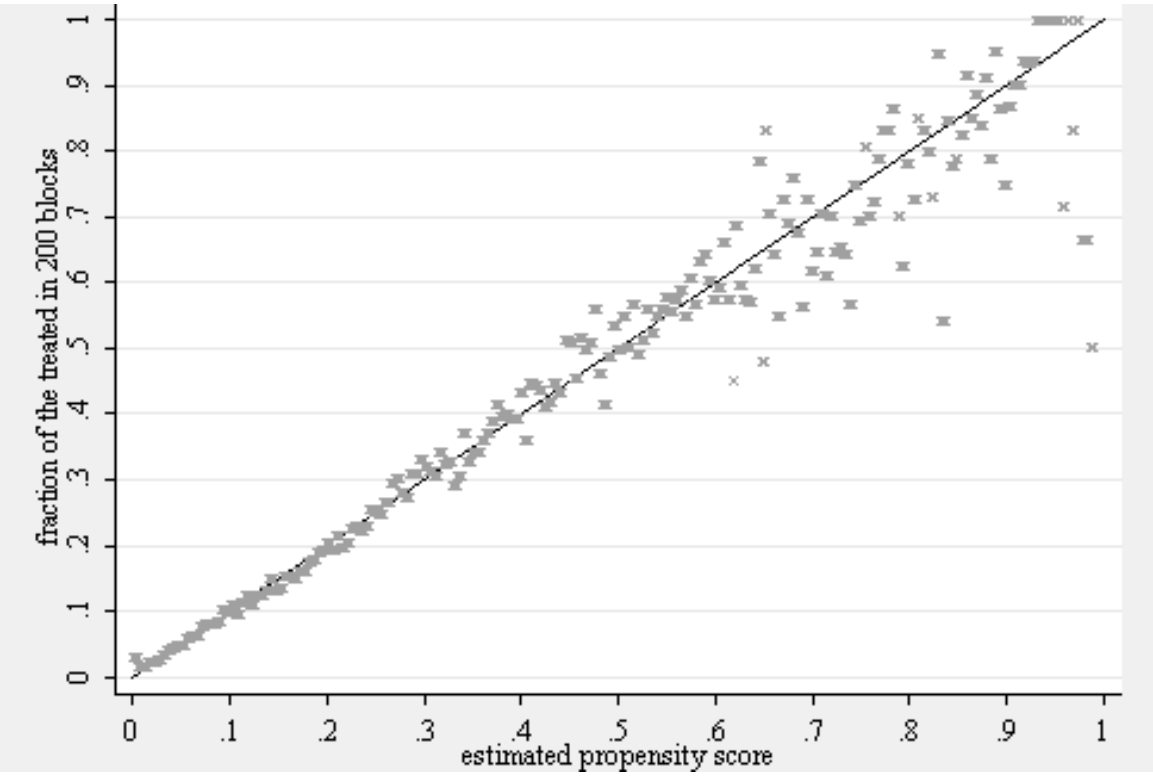

Treatment and control dist. of predicted pscores
5%-tile, 25%-tile, 50%-tile, 75%-tile, 95%-tile
Pr(tobacco)
0
.2
.4
.6
.8
1
0
1
stratification based on 33 blocks

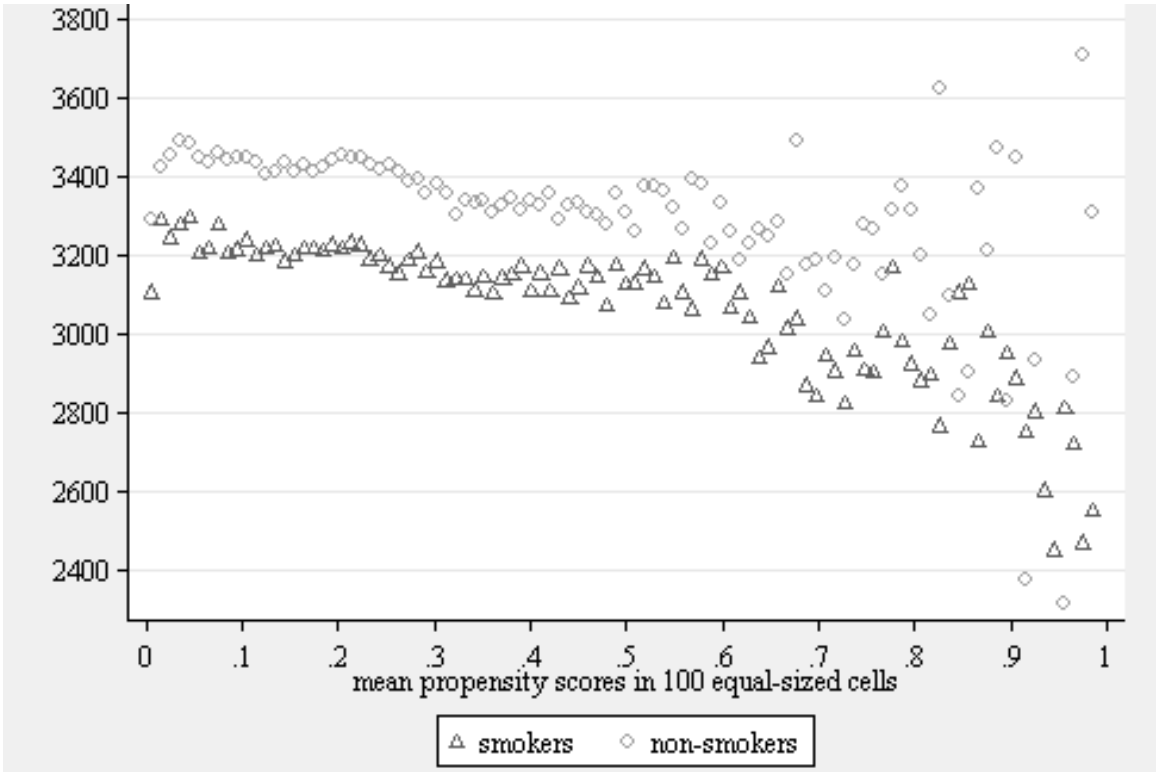

Treatment and control dist. of predicted pscores
5%-tile, 25%-tile, 50%-tile, 75%-tile, 95%-tile
Pr(tobacco)
0
.2
.4
.6
.8
1
0
1
stratification based on 100 blocks

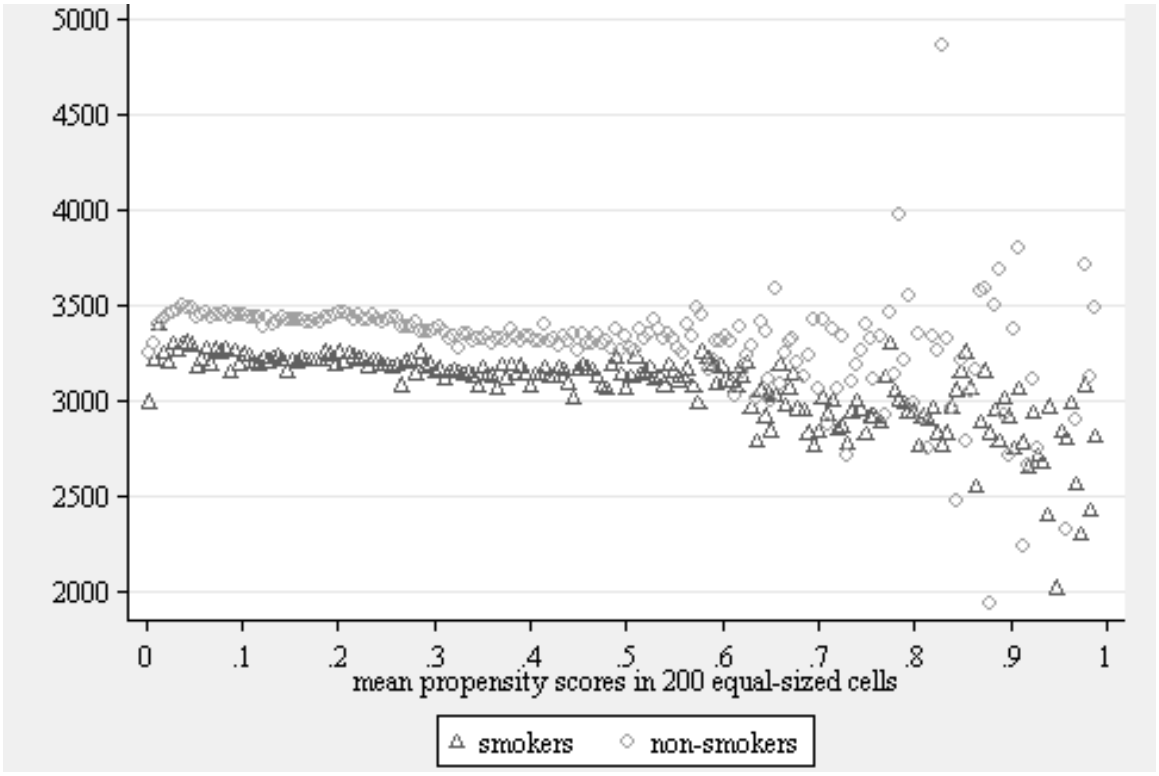

Treatment and control dist. of predicted pscores
5%-tile, 25%-tile, 50%-tile, 75%-tile, 95%-tile
Pr(tobacco)
0
.2
.4
.6
.8
1
0
1
stratification based on 200 blocks

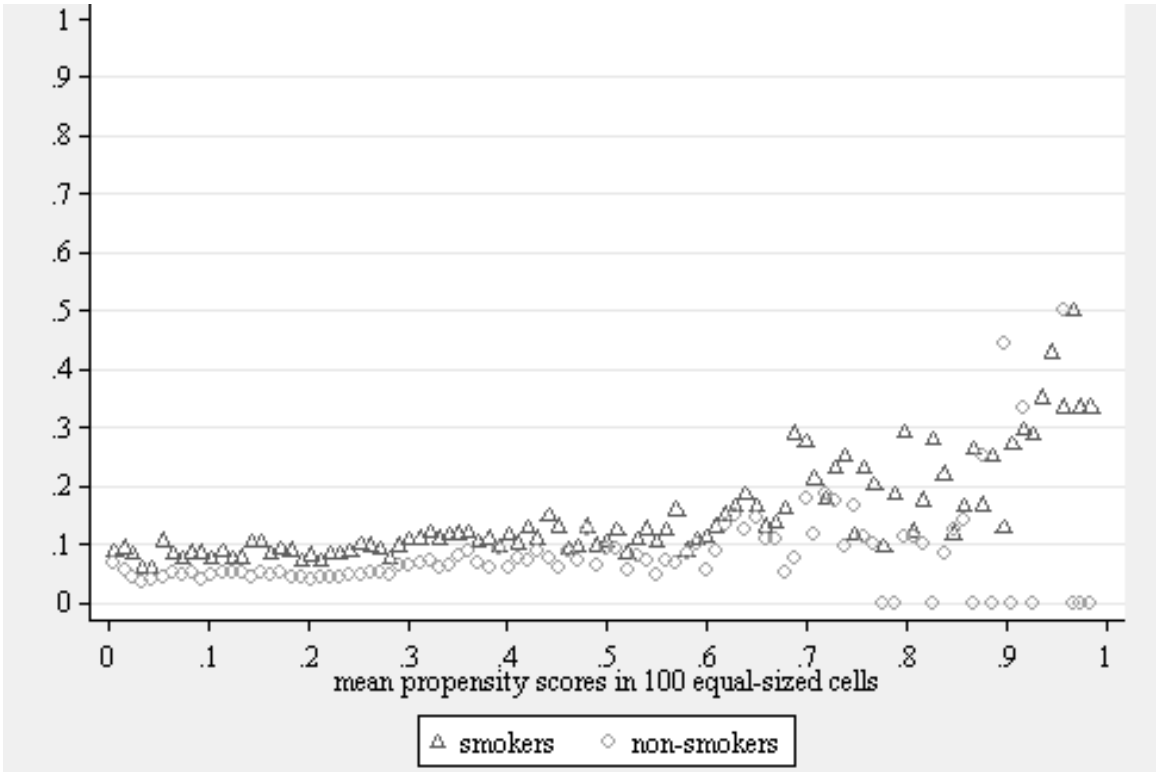

fraction of the treated in 33 blocks
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
estimated propensity score

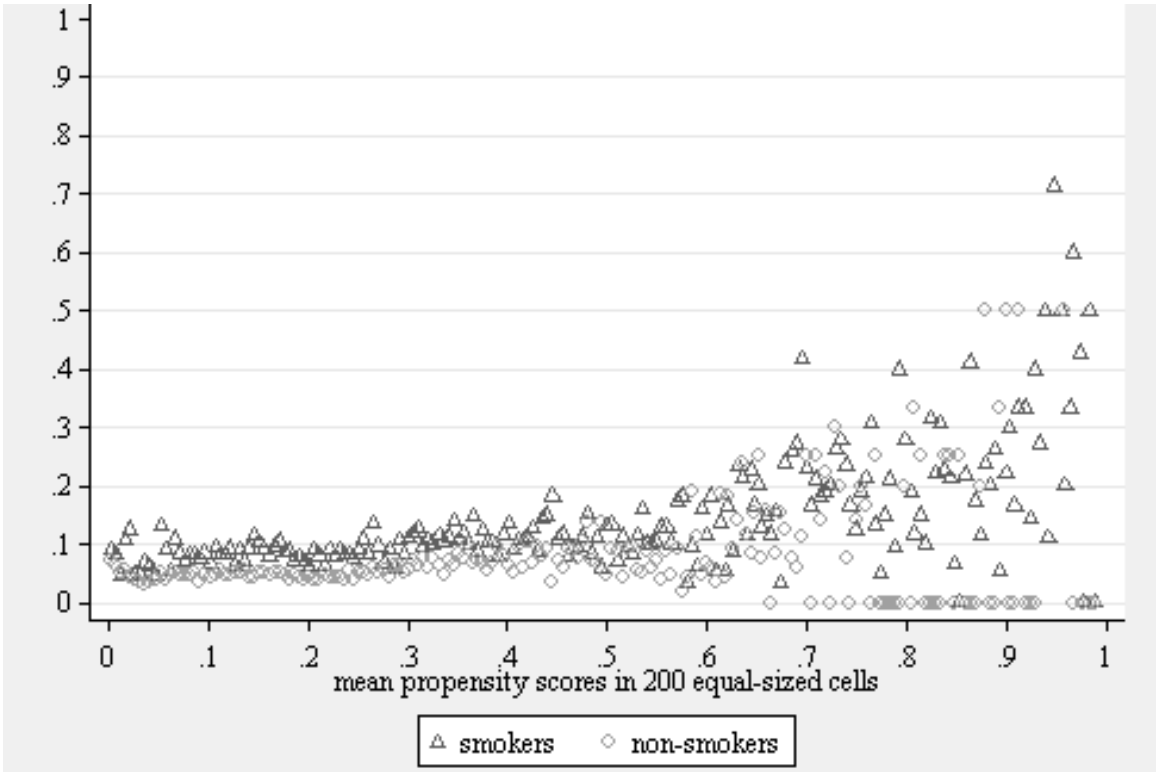

fraction of the treated in 100 blocks
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
estimated propensity score

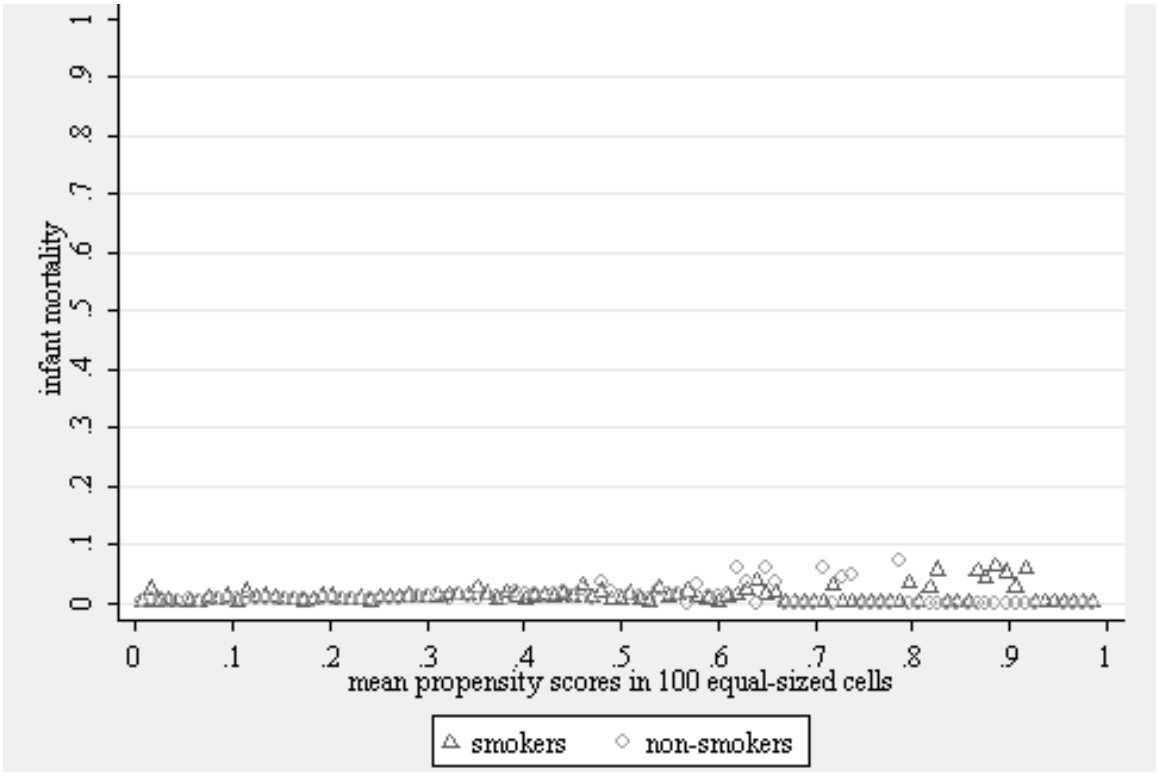

fraction of the treated in 200 blocks
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
estimated propensity score

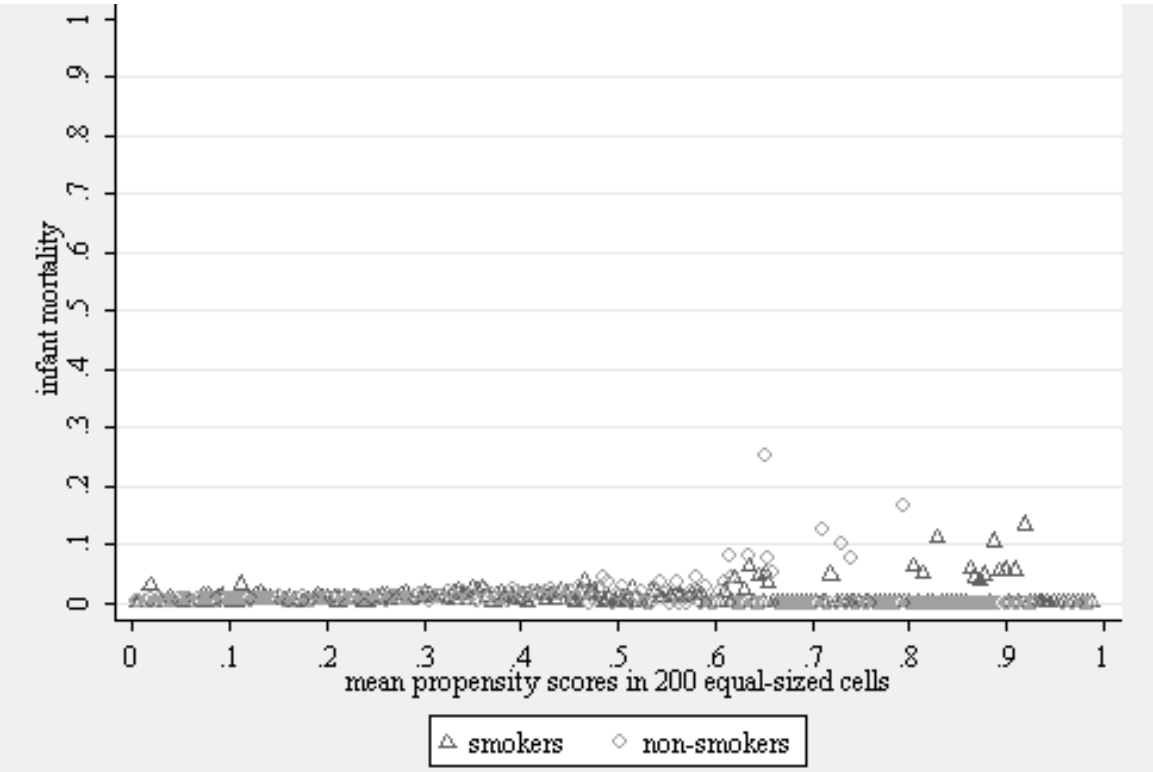

3800
3600
3400
3200
3000
2800
2600
2400
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
mean propensity scores in 100 equal-sized cells
smokers
non-smokers


5000
4500
4000
3500
3000
2500
2000
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
mean propensity scores in 200 equal-sized cells
smokers
non-smokers


1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
mean propensity scores in 100 equal-sized cells
smokers
non-smokers


1
.9
.8
.7
.6
.5
.4
.3
.2
.1
0
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
mean propensity scores in 200 equal-sized cells
smokers
non-smokers


infant mortality
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
mean propensity scores in 100 equal-sized cells
smokers
non-smokers


infant mortality
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
mean propensity scores in 200 equal-sized cells
smokers
non-smokers