Lehigh University
Department of Economics

Muzhe Yang
Fall 2011

Treatment Effect Estimator with Matching[1]

If the treatment group and the control group are different in observables $\mathbf{x}$, then the difference in the outcome y between the two groups may not be attributable to the difference in the treatment status. One solution could be comparing only those individuals with the same or similar values of $\mathbf{x}$ from the two groups—*matching* on the observables. If $\mathbf{x}$ is high-dimensional, then it can be difficult to find exactly matched individuals. Nonetheless, there is one solution to the dimension problem—*propensity score matching*—which demands a *very strong* assumption—selection on observables.

1 Balancing Observables Using Propensity Scores

The *propensity score* $p(\mathbf{x})$ is the probability of receiving the treatment d conditional on observables $\mathbf{x}$, that is $\Pr(d = 1|\mathbf{x})$. The idea of using $p(\mathbf{x})$ to avoid the dimension problem actually has already been applied to inverse probability weighting (IPW) estimators. Using the (generalized version of the) law of iterated expectation, we have the following result:

$$\begin{aligned}\mathbb{E}\left(d|p(\mathbf{x})\right) &= \mathbb{E}\left[\mathbb{E}\left(d|\mathbf{x}\right)|p(\mathbf{x})\right] \text{ (noting that } p(\mathbf{x}) = \mathbb{E}\left(d|\mathbf{x}\right)\text{)} \\ &= p(\mathbf{x}).\end{aligned}$$

Using this result, we can obtain

$$\Pr\left(\mathbf{x} \leqslant t|d = 1, p(\mathbf{x})\right) = \Pr\left(\mathbf{x} \leqslant t|d = 0, p(\mathbf{x})\right), \text{ where } 0 < p(\mathbf{x}) < 1 \text{ for all } \mathbf{x}.$$

This means that *for the same $p(\mathbf{x})$, the distribution of $\mathbf{x}$ will be the same across the treatment group and the control group.*

Proof.

$$\begin{aligned}\Pr\left(d = 1, \mathbf{x} \leqslant t|p(\mathbf{x})\right) &= \Pr(d \cdot 1\{\mathbf{x} \leqslant t\} = 1|p(\mathbf{x})) \\ &= \mathbb{E}\left[d \cdot 1\{\mathbf{x} \leqslant t\}|p(\mathbf{x})\right] \\ &= \mathbb{E}\left[\mathbb{E}\left(d \cdot 1\{\mathbf{x} \leqslant t\}|\mathbf{x}\right)|p(\mathbf{x})\right] \\ &= \mathbb{E}\left[\mathbb{E}\left(d|\mathbf{x}\right) \cdot 1\{\mathbf{x} \leqslant t\}|p(\mathbf{x})\right] \\ &= \mathbb{E}\left[p(\mathbf{x}) \cdot 1\{\mathbf{x} \leqslant t\}|p(\mathbf{x})\right] \\ &= p(\mathbf{x}) \cdot \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right) \\ &= \mathbb{E}\left(d|p(\mathbf{x})\right) \cdot \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right) \\ &= \Pr\left(d = 1|p(\mathbf{x})\right) \cdot \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right).\end{aligned}$$

Thus, we have

$$\frac{\Pr\left(d = 1, \mathbf{x} \leqslant t|p(\mathbf{x})\right)}{\Pr\left(d = 1|p(\mathbf{x})\right)} = \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right),$$

which implies that

$$\Pr\left(\mathbf{x} \leqslant t|d = 1, p(\mathbf{x})\right) = \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right).$$

Furthermore, note that

$$\begin{aligned}\Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right) &= \Pr(d = 1|p(\mathbf{x})) \cdot \Pr(\mathbf{x} \leqslant t|d = 1, p(\mathbf{x})) + \Pr(d = 0|p(\mathbf{x})) \cdot \Pr(\mathbf{x} \leqslant t|d = 0, p(\mathbf{x})) \\ &= \Pr(d = 1|p(\mathbf{x})) \cdot \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right) + \Pr(d = 0|p(\mathbf{x})) \cdot \Pr(\mathbf{x} \leqslant t|d = 0, p(\mathbf{x})) \\ &\Rightarrow \Pr(d = 0|p(\mathbf{x})) \cdot \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right) = \Pr(d = 0|p(\mathbf{x})) \cdot \Pr(\mathbf{x} \leqslant t|d = 0, p(\mathbf{x})) \\ &\Rightarrow \Pr\left(\mathbf{x} \leqslant t|p(\mathbf{x})\right) = \Pr(\mathbf{x} \leqslant t|d = 0, p(\mathbf{x})).\end{aligned}$$

[1] This section is based on Lee (2005) and my notes from Kenneth Chay.

In summary we have

$$\Pr\left(\mathbf{x} \leqslant t | d=1, p(\mathbf{x})\right) = \Pr\left(\mathbf{x} \leqslant t | p(\mathbf{x})\right) = \Pr(\mathbf{x} \leqslant t | d=0, p(\mathbf{x})).$$

Given $p(\mathbf{x})$, the distribution of $\mathbf{x}$ is the same across the two groups. ■

Next we will discuss another way of removing bias due to selection on observables—an alternative to IPW.

2 Removing Bias due to Selection on Observables

Assume that d is independent of (y_0, y_1) conditional on $\mathbf{x}$, that is $d \amalg (y_0, y_1) | \mathbf{x}$. This assumption rules out selection bias due to unobservables. Also, this assumption is often referred to as one of the following—"selection-on-observables," "d is ignorable given $\mathbf{x}$," "randomization of d given $\mathbf{x}$," "conditional independence," "ignorable treatment assignment," "ignorability," or "unconfoundedness."

Rosenbaum and Rubin (1983) show that, if the conditional independence holds, then d is independent of (y_0, y_1) given just $p(\mathbf{x})$:

$$d \amalg (y_0, y_1) | \mathbf{x} \Rightarrow d \amalg (y_0, y_1) | p(\mathbf{x}).$$

However,

$$d \amalg (y_0, y_1) | p(\mathbf{x}) \not\Rightarrow d \amalg (y_0, y_1) | \mathbf{x}.$$

A sketch of the proof is the following:

$$\begin{aligned}
\mathbb{E}\left(d | y_0, y_1, p(\mathbf{x})\right) &= \mathbb{E}\left[\mathbb{E}\left(d | y_0, y_1, \mathbf{x}\right) | y_0, y_1, p(\mathbf{x})\right] \\
&= \mathbb{E}\left[\mathbb{E}\left(d | \mathbf{x}\right) | y_0, y_1, p(\mathbf{x})\right] \\
&= \mathbb{E}\left[p(\mathbf{x}) | y_0, y_1, p(\mathbf{x})\right] \\
&= p(\mathbf{x}) \\
&= \mathbb{E}\left[d | p(\mathbf{x})\right].
\end{aligned}$$

Note that d is binary, so the mean-independence is equivalent to independence:

$$\Pr(d=1 | y_0, y_1, p(\mathbf{x})) = \Pr(d=1 | p(\mathbf{x})) \text{ and } \Pr(d=0 | y_0, y_1, p(\mathbf{x})) = \Pr(d=0 | p(\mathbf{x})).$$

Under the assumption $d \amalg (y_0, y_1) | \mathbf{x}$*, which implies* $d \amalg (y_0, y_1) | p(\mathbf{x})$, we can identify the average treatment effect using $p(\mathbf{x})$:

$$\begin{aligned}
\text{ATE} &\equiv \mathbb{E}\left(y_1 - y_0\right) = \mathbb{E}\left[\mathbb{E}\left(y_1 - y_0 | p(\mathbf{x})\right)\right], \\
\text{where } y &= d y_1 + (1-d) y_0 \text{ and} \\
\mathbb{E}(y_1 - y_0 | p(\mathbf{x})) &= \mathbb{E}\left(y_1 | p(\mathbf{x})\right) - \mathbb{E}\left(y_0 | p(\mathbf{x})\right) \\
&= \mathbb{E}\left(y_1 | p(\mathbf{x}), d=1\right) - \mathbb{E}\left(y_0 | p(\mathbf{x}), d=0\right) \\
&= \mathbb{E}\left(y | p(\mathbf{x}), d=1\right) - \mathbb{E}\left(y | p(\mathbf{x}), d=0\right).
\end{aligned}$$

For propensity score matching, there must be overlap in the values of $p(\mathbf{x})$ between the treatment group and the control group.

Using $p(\mathbf{x})$ as opposed to $\mathbf{x}$ in the conditioning implies a considerable dimension reduction because $p(\mathbf{x})$ is a scalar. However, when matching is based on the estimated rather than the true propensity scores, we should correct for the error $\widehat{p}(\mathbf{x}) - p(\mathbf{x})$, which will affect the asymptotic variance of the treatment effect matching estimator. In practice, $p(\mathbf{x})$ is often estimated by logit or probit, and the error $\widehat{p}(\mathbf{x}) - p(\mathbf{x})$ is simply ignored. Alternatively, bootstrap can be used.

References

Lee, M.-j. (2005). *Micro-Econometrics for Policy, Program, and Treatment Effects*. Oxford; New York: Oxford University Press.

Rosenbaum, P. R. and D. B. Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70(1): 41-55.

Lecture: Selection on Observables

Evaluation/Selection Problem:

Ex. Linear additive model

$$y_i = \alpha + \theta \cdot T_i + X_i' \beta + \varepsilon_i$$

Focus on binary (0-1) treatment, homogeneous treatment effects

$$T_i = \begin{cases} 1, \text{ if treated} \\ 0, \text{ otherwise} \end{cases}$$

$$\theta_i = \theta \ \ \forall i$$

***i* has 2 potential outcomes**

y_{0i} **if** $T_i = 0$

y_{1i} **if** $T_i = 1$

Fundamental problem of causal inference ≡ unobserved counterfactual

Latent var.: $y_i^* = (y_{0i}, y_{1i})$

Observe: $y_i = (1 - T_i) y_{0i} + T_i y_{1i}$

Ex.

$$y_{0i} = \gamma_0 + g_0(X_i) + u_{0i}, \text{ if } T_i = 0$$

$$y_{1i} = \gamma_1 + g_1(X_i) + u_{1i}, \text{ if } T_i = 1$$

$$T_i = 1(T_i^* > 0), \ T_i^* = f(X_i, w_i) + v_i, \ \ w_i = \text{I.V.}$$

If $g_0(X_i) = X_i' \beta_0, \ g_1(X_i) = X_i' \beta_1, \ \beta_0 = \beta_1$

Then Average T.E.: ATE ≡ E($y_{1i} - y_{0i}$) = $\gamma_1 - \gamma_0 = \theta$ "causal effect"

For now, assume constant additive T.E. (fixed coeffs)

$y_{1i} = y_{0i} + \theta$ ➔ **strong homogeneous T.E. assumption**

"Gold standard" solution: Random Assignment of T_i

$T_i \amalg (y_{0i}, y_{1i})$

- **By definition of R.A., this identifies ATE**
- **Control group provides correct counterfactual as $N \to \infty$**

$$\bar{y}_1 - \bar{y}_0 \xrightarrow{p} \theta$$

- **Indirect test of R.A.: $\bar{X}_1 \approx \bar{X}_0 \ \ \forall x_{ik}$**

Linear model:

$$y_i = \theta \cdot T_i + X_i' \beta + u_i$$

$$T_i^* = X_i' \Pi_1 + w_i' \Pi_2 + v_i$$

$$T_i = 1(T_i^* \geq 0)$$

Usually assume E($u_i \cdot T_i$) = 0 ➔ E($u_i \cdot v_i$) = 0

Comparing mean of y_i by T_i

$$E(y_i|T_i=1) - E(y_i|T_i=0) = E(y_{1i} - y_{0i}|T_i=1) + [E(y_{0i}|T_i=1) - E(y_{0i}|T_i=0)]$$

= ATE if constant T.E. **= 0 if T_i R.A.**

$$y_i = \alpha + \theta \cdot T_i + \varepsilon_i$$

$$E(y_i|T_i=1) - E(y_i|T_i=0) = \theta + [E(\varepsilon_{1i}|T_i=1) - E(\varepsilon_{0i}|T_i=0)]$$

without R.A., observed correlation between (y_i, T_i) likely biased by omitted variables

<u>Random assignment conditional on observables</u> ≡ Selection on observables

$T_i \perp\!\!\!\perp (y_{0i}, y_{1i}) | X_i$

T_i <u>independent</u> of potential outcomes <u>conditional</u> on X_i

$$\Rightarrow E(y_{0i} | T_i = 1, X_i) - E(y_{0i} | T_i = 0, X_i) = 0$$

- **only source of bias due to X_i**
- **remove this bias**

<u>Ex.</u> Use as many X_i as possible and "kitchen sink" the regression
<u>Problem:</u> Data mining, can sometimes accentuate omitted variables bias (OVB)

<u>Approaches:</u> Multivariate matching, Propensity score, Regression discontinuity design

<u>Regression analogy:</u>

1) $y_i = \theta \cdot T_i + X_i' \beta + u_i, \qquad E\left(T_i \cdot u_i \middle| X_i' \beta \right) = 0$

- **low dimension problem, just control for $X_i' \beta$**

2) $y_i = \theta \cdot T_i + g(X_i) + u_i, \qquad E(T_i \cdot u_i | g(X_i)) = 0$

- **$g(X_i)$ may include polynomials and interactions ➔ high dimension**

<u>More generally</u>, $T_i \perp\!\!\!\perp (y_{0i}, y_{1i}) | X_i \Rightarrow E(T_i \cdot u_i | X_i) = 0$

- **Misspecify $g(X_i)$ ➔ O.V.B.**
- **Bias-efficiency trade-off**
- **"Problem" with linear regression: arbitrary specification of $g(X_i) = X_i' \beta$**

Multivariate Matching:

dim(X_i) = K

- **For each treatment observation, match control case with "identical" X_i ➔ Design problem if X_i, T_i collinear**
- **Fitting flexible functional form with K arguments**
  "Nonparametric" regression ≡ computational burden N^K
- **"Curse of dimensionality"**
  Reduce "partial" variation in T_i substantially

Ex. case-control method

i) Match each treatment to one control based on the "closeness" of X_i using some "distance metric"

ii) Using the matched pairs, run a regression controlling for "pair identifier" fixed effects.

How to reduce the dimension of problem and remove bias due to X_i?

Propensity Score Theorem:

If T_i R.A. conditional on X_i, then T_i R.A. conditional on the propensity score.

$p_i \equiv \Pr(T_i = 1|X_i) = E(T_i|X_i) \equiv p(x_i)$ = Prob. of treatment conditional on X_i

$$(y_{0i}, y_{1i}) \perp\!\!\!\perp T_i | X_i \Rightarrow (y_{0i}, y_{1i}) \perp\!\!\!\perp T_i | p(x_i)$$

Very strong assumption!

Idea: Since T_i binary, $E(T_i|X_i)$, $Var(T_i|X_i)$ determined by $p(x_i)$
➔ $p(x_i)$ sufficient statistic for T_i, X_i relation

$$T_i \perp\!\!\!\perp X_i | p(x_i)$$

Reduced dimension: Just control for flexible form of single index $p(x_i)$, instead of all X_i

2 steps:

1. Estimate pscore, $\hat{p}(x_i)$ such that it "balances" x_i

2. Estimate $\theta \equiv$ ATE controlling for $\hat{p}(x_i)$

Ex. match treated and control cases with similar $\hat{p}(x_i)$

---

1. Estimate $\hat{p}(x_i)$ by logit

$$Pr(T_i = 1 | X_i) = \frac{e^{h(X_i)}}{1 + e^{h(X_i)}}$$

- **$h(X_i)$ contains linear and maybe higher order terms**
- **include enough terms such that Treatments and Controls with similar $\hat{p}(x_i)$ have similar X_i (balanced)**
- **pscore reduces dimensionality**
  $p(x_i)$ single-index that balances X_i: "Match" $p_i \approx p_j$

$$y_i = \theta \cdot T_i + g(X_i) + \varepsilon_i$$

$$y_i = \theta \cdot T_i + g^0(p(x_i)) + \varepsilon_i^0$$

- **Adjusts for selection bias due to X_i in descriptive 2-step way**

Algorithm for estimating pscore:

1) parsimonious logit ➔ estimate $\hat{p}(x_i)$

2) stratify data into quintile blocks of $\hat{p}(x_i)$

3) Test $\overline{X}_1 = \overline{X}_0$ in T and C groups within each block

t-tests (F-tests) of significant difference in sample means of each x_k within each block

i) If X_i "balanced" in each block, stop

ii) If x_k not balanced in some blocks, divide block into 2 blocks and reevaluate

iii) If x_k not balanced in all blocks, add interaction and/or polynomial of x_k to logit and reevaluate

<u>Goal:</u> Balance of X_i in Treat and Control groups

Overlap in $\hat{p}(x_i)$ for $T_i = 1$ and $T_j = 0$ ➔ overlap in X_i

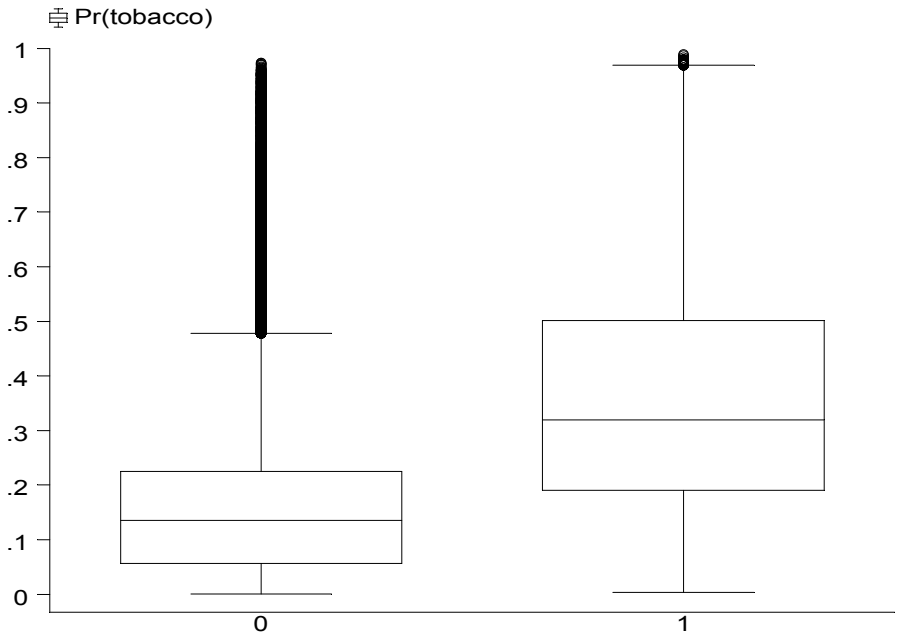
$$T_i \amalg X_i | p(x_i)$$

<u>Example of stopping rule:</u>
Stop when fail to reject $\bar{x}_{1k} = \bar{x}_{0k}$ for over 90% of t-tests within a block

---

Examining $\hat{p}(x_i)$ by T_i gives sense of "nonrandomness" of T_i assignment

Example: Box-Plot – 5%-tile, 25%-tile, 50%-tile, 75%-tile, 95%-tile of treatment and control distributions of predicted propensity scores.



Interpretation: Amount of "overlap" in plots ≈ similarity of X's in treatment and controls.

A lot of overlap ➔ very little selection on the X's (good research design)
Little overlap ➔ pure selection on X's (bad design) ➔ extrapolating across <u>non-comparable</u> populations.

Box-Plot if Random Assignment?

2. Estimate θ controlling for $\hat{p}(x_i)$

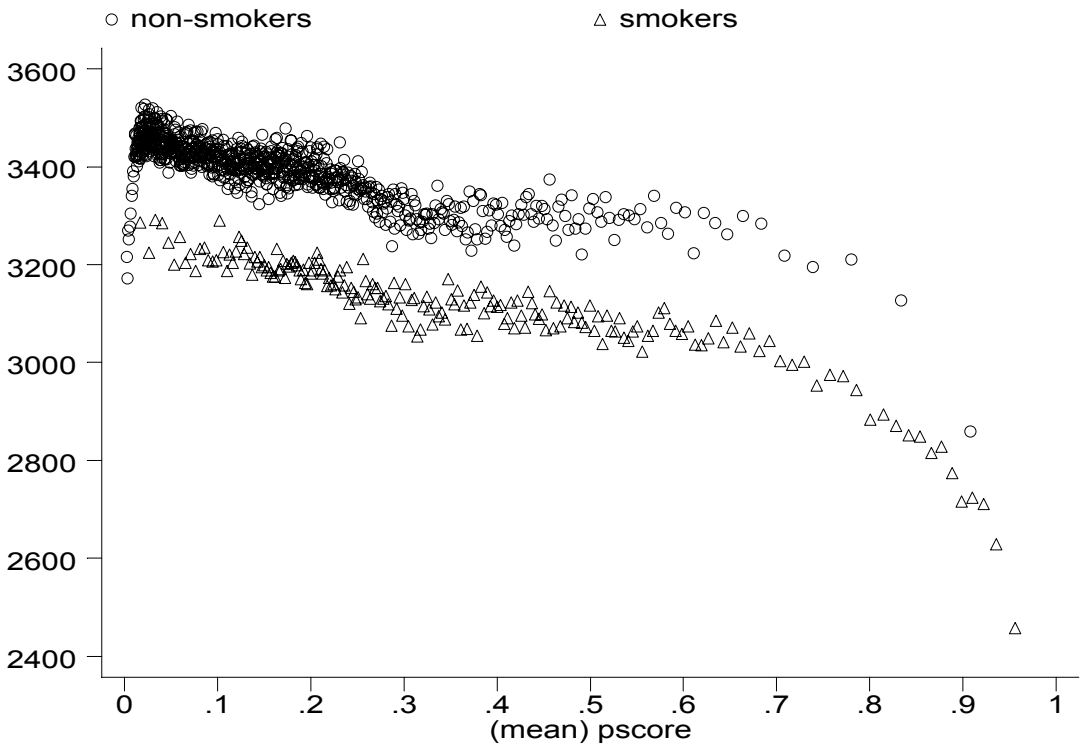
i) Most general (informative) use of $\hat{p}_i$ – Graphical analysis

Estimate $E\left(y_i \middle| \hat{p}_i, T_i = 0 \right) = f_0\left(\hat{p}_i \middle| T_i = 0 \right)$

$$E\left(y_i \middle| \hat{p}_i, T_i = 1 \right) = f_1\left(\hat{p}_i \middle| T_i = 1 \right)$$

- **Can estimate by bivariate nonparametric regression (e.g., kernel regression, local linear regression – "ksm" STATA command).**
- **More transparently, calculate means of outcome for 100+ blocks of $\hat{p}_i$, separately for treatment and controls, and plot against $\hat{p}_i$.**
- **Plot $\hat{f}_0$ and $\hat{f}_1$ against $\hat{p}_i$ – bias summarized by single index $\hat{p}_i$ ➔ no dimensionality problem.**
- **Useful if N_T, N_C are large.**

Example: Smoking during pregnancy and infant birth weight (grams)



- **Slope m_1 gives union selection on observables**
  unadjusted $\bar{y}_1 - \bar{y}_0$ biased since $m_1, m_0 \neq 0$
- **$(m_1 - m_0)$ gives differential selection into 2 sectors**
- **$\bar{y}_1^j - \bar{y}_0^j \big|$ fixed p_j = union wage gap adjusted for selection on X_i**
- **"Unrestricted" description of selection process and heterogeneity in "treatment effects" with the probability of observable selection – under <u>very strong</u> initial assumption.**

ii) Regression analog

$$y_i = \alpha + \theta \cdot T_i + \delta_1 \hat{p}_i + \delta_2 T_i \left(\hat{p}_i - \hat{\mu}_p \right) + u_i, \quad \hat{\mu}_p = \frac{1}{N} \sum \hat{p}_i$$

control for selection bias

- **$\delta_1 = m_0$, $\delta_2 = (m_1 - m_0)$**
- **restrictive linear specification – prone to misspecification**

  plim$\left(\hat{\theta} \right) = \theta$ if and only if $E\left(y_{1i} \middle| T_{1i} = 1, p(x_i) \right)$ linear in $p(x_i)$

  test by including polynomials of $\hat{p}(x_i)$
- **Simple summary (use bootstrap to calculate standard errors)**
- **Control for $X_i' \beta$ to gauge quality of balance through $\hat{p}(x_i)$**

Example – maternal smoking during pregnancy, birth weight, and the propensity score for smoking during pregnancy

```
. reg bweight tobacco

  Source |       SS       df       MS                  Number of obs =  514454
---------+------------------------------               F(  1,514452) =18890.45
   Model |  6.7965e+09     1  6.7965e+09               Prob > F      =  0.0000
Residual |  1.8509e+11514452  359787.703               R-squared     =  0.0354
---------+------------------------------               Adj R-squared =  0.0354
   Total |  1.9189e+11514453   372998.22               Root MSE      =  599.82

------------------------------------------------------------------------------
 bweight |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 tobacco |  -283.1877   2.060408   -137.443   0.000       -287.226   -279.1494
   _cons |   3394.513   .9396976   3612.346   0.000       3392.671    3396.355
------------------------------------------------------------------------------



. reg bweight tobacco pscore smkpscre

  Source |       SS       df       MS                  Number of obs =  514454
---------+------------------------------               F(  3,514450) = 8524.20
   Model |  9.0869e+09     3  3.0290e+09               Prob > F      =  0.0000
Residual |  1.8280e+11514450  355337.074               R-squared     =  0.0474
---------+------------------------------               Adj R-squared =  0.0473
   Total |  1.9189e+11514453   372998.22               Root MSE      =  596.10

------------------------------------------------------------------------------
 bweight |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 tobacco |  -194.4738   2.434303    -79.889   0.000      -199.2449   -189.7026
  pscore |  -373.1682   6.585591    -56.664   0.000      -386.0758   -360.2607
smkpscre |  -91.66314   10.49604     -8.733   0.000       -112.235   -71.09124
   _cons |   3456.679   1.440737   2399.244   0.000       3453.855    3459.503
------------------------------------------------------------------------------
```

(Should have added “robust” to the regression; also need to correct for sampling error in estimated/generated regressor = propensity score)