Solutions to Advanced Econometrics Homework 4
Maternal smoking during pregnancy

Ashwin Iyenggar (1521001)
ashwin.iyenggar15@iimb.ernet.in

December 7, 2016

Question (a)

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

Conditions for comparing unadjusted means

Let us consider a situation where maternal smoking were assigned randomly not conditional upon any other observable variables (like physical characteristics of the mother or known habits). Under this condition, we would expect the treatment and control group to be similar on all respects other than on maternal smoking habits. It may thus be reasonable to compare the unadjusted mean of the difference in birth weight of infants between smoking and non-smoking mothers to determine the causal effects of maternal smoking on infant birth weight. Therefore, the answer to the first part of the question is that we would need random assignment of maternal smoking during pregnancy that is unconditional on any other observable variables.

Estimation of impact

I demonstrate the impact of maternal smoking on birthweight under assumption of random assignment of maternal smoking in two ways. Table 1 shows the result of a t-test on birthweight between smoking and non-smoking mothers. The results indicate that there is a statistically significant difference of 257.6 grams in birthweight between children born to smoking mothers and those born to non-smoking mothers.

Table 2 is the output from regressing birthweight on mother smoking status, and this suggests the same result as the t-test in Table 1 i.e., indicating an average drop in birthweight of 257.6 grams for smoking mothers as compared to the non-smoking mothers in the sample.

Evidence against random assignment

If maternal smoking were randomly assigned, then we would expect that there would be no significant differences between treatment and control on most (if not all) observable parameters of interest (in this case for the birth of the child). Intuitively, it is clear that maternal smoking

Table 1: T-test for birth weight by maternal smoking status

	(1)
	tobacco
dbirwt	257.6***
	(65.38)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Regression Results

Maternal Smoking Effects on Birthweight	
	(1)
VARIABLES	Random Assignment
tobacco	-257.5724***
	(3.9895)
Constant	3,425.5559***
	(1.6958)
Observations	139,149
R^2	0.0298

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

during pregnancy may not be independent of the state of pregnancy (some mothers may decide to give up smoking so as to not hurt the foetus), or of socio-economic and racial characteristics of the parents. Empirically though, this may be demonstrated by comparing the mean values of various observable characteristics between the treatment and control group. Table 3 provides the results of t-tests on several observable variables[1] between smoking and non-smoking mothers. While the list in Table 3 is not exhaustive on the observable characteristics, it is clear that smoking and non-smoking mothers vary distinctly on most observable characteristics. This is therefore sufficient to demonstrate that maternal smoking is not randomly assigned.

The Stata code for this question is as below

```
use smoking_labels, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/

estpost ttest dbirwt, by(tobacco)
esttab using ‘imagepath’a1.tex, title("T-test for birth weight by
   maternal smoking status\label{a1}") mtitle("tobacco") replace

reg dbirwt tobacco, vce(robust)
outreg2 using ‘imagepath’a2.tex, title("Maternal Smoking Effects on
   Birthweight\label{a2}") ctitle("Random Assignment") tex(pretty frag)
   dec(4) replace


estpost ttest dmage dmeduc dmar dlivord mblack fblack alcohol tripre0
   tripre1 tripre2 tripre3, by(tobacco)
esttab using ‘imagepath’a3.tex, title("T-tests by Maternal Smoking Status
   \label{a3}") mtitle("Mean Difference") replace
```

[1] I am grateful to Prof. Shailender Swaminathan to have pointed out this method in class, correcting my previous misconception of looking at variable correlations as the appropriate method to determine this

Table 3: T-tests by Maternal Smoking Status

	(1) Mean Difference
dmage	1.865***
	(49.85)
dmeduc	1.253***
	(85.42)
dmar	-0.234***
	(-83.52)
dlivord	-0.167***
	(-20.70)
mblack	-0.0295***
	(-13.43)
fblack	-0.0379***
	(-16.80)
alcohol	-0.0431***
	(-49.49)
tripre0	-0.0171***
	(-23.97)
tripre1	0.119***
	(45.06)
tripre2	-0.0819***
	(-34.43)
tripre3	-0.0196***
	(-16.61)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Question (b)

Suppose maternal smoking is randomly assigned conditional on the observables determinants of infant birth weight. What does this imply about the relationship between maternal smoking and unobservable determinants of birth weight conditional on the observables? Use a linear regression model to estimate the impact of maternal smoking on birth weight, and report your estimates. Compare the estimate to the one from part (a).

Implications for relationship with unobservable variables

By definition, the statement above implies that maternal smoking is uncorrelated (ideally, independent) with unobservable determinants of infant birth weight. This may be concluded, because we have been given that maternal smoking is randomly assigned conditional on the observable determinants - it automatically follows that maternal smoking is not determined by any other factors, including unobservable determinants of birth weight.

Table 4: tobacco Randomly Assigned Conditional on Observables

	(1) Birth Weight
tobacco	-218.2***
	(-56.07)
dmage	-2.847***
	(-5.98)
dmeduc	5.443***
	(6.00)
dmar	-35.41***
	(-7.96)
dlivord	33.20***
	(17.13)
nprevist	29.63***
	(49.05)
disllb	-0.175**
	(-2.79)
dfage	-0.151
	(-0.43)
dfeduc	3.833***
	(4.65)
anemia	-33.77*

	(-2.29)
diabete	49.93***
	(3.87)
phyper	-168.3***
	(-15.14)
pre4000	466.3***
	(34.96)
preterm	-498.7***
	(-35.08)
alcohol	-26.40
	(-1.93)
drink	-8.562**
	(-3.04)
foreignb	-17.15*
	(-2.02)
plural	-921.8***
	(-73.90)
deadkids	-9.282***
	(-4.28)
mblack	-159.7***
	(-11.95)
motherr	-80.27***
	(-4.12)
mhispan	-63.62***
	(-4.57)
fblack	-50.20***
	(-3.82)
fotherr	-101.0***
	(-5.27)
fhispan	-68.91***
	(-5.32)
tripre1	-175.6***
	(-19.12)

tripre2	-106.3***
	(-11.77)
tripre3	0
	(.)
tripre0	-183.9***
	(-9.19)
first	-83.27***
	(-16.80)
death	-1259.9***
	(-35.16)
_cons	3262.1***
	(218.59)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Estimation of impact and comparison with random assignment

While the question asks us to use a linear model, this is clearly inappropriate given that the random assignment of maternal smoking is conditional on observable variables. A better estimation approach would have used bins to place observations with similar characteristics. However, as asked by the question we regress our dependent variable, birth weight on all available variables in the data set. The regression results are as in Table 4 where the effect of tobacco has come down to 218.2 grams from being 257.6 grams in question (a). This is so because our regression here has accounted for all observable variables (though not recognised them into appropriate bins)

The Stata code for this question is as below

```
use smoking_labels, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
set more off

local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
reg dbirwt tobacco dmage dmeduc dmar dlivord nprevist disllb dfage dfeduc
    anemia diabete phyper pre4000 preterm alcohol drink foreignb plural
   deadkids mblack motherr mhispan fblack fotherr fhispan tripre1 tripre2
    tripre3 tripre0 first death, vce(robust)
esttab using ‘imagepath’b1.tex, title("tobacco Randomly Assigned
   Conditional on Observables\label{b1}") mtitle("Birth Weight")
   longtable replace
```

Question (c)

Under the assumption of random assignment conditional on the observables, what are the sources of misspecification bias in the estimates generated by the linear regression model estimated in part (b)? Now use an approach in the spirit of multivariate matching ? i.e., estimate the smoking effects using a flexible functional form for the control variables (e.g., higher order terms and interactions). What are the benefits and drawbacks to this approach?

Sources of misspecification

The primary source of misspecification in using a linear model arise from:
1. Functional form being required to be linear in the observable variables, and
2. From potential omitted variable bias due to the exclusion of non-linear terms in the linear model specification including interaction terms. This leads us to recognize that we really do need to modify our specification based matching on some characteristics of the data.

Estimation using flexible functional form

Table 5: Flexible functional form regression

	(1) Birth Weight
tobacco	-217.0***
	(-55.42)
dmage	-2.515***
	(-4.90)
dmeduc	1.075
	(0.20)
dmar	-15.11
	(-0.81)
dlivord	33.21***
	(17.09)
nprevist	29.68***
	(49.06)
disllb	-0.156*
	(-2.48)
dfage	-0.234
	(-0.61)
dfeduc	0.103

	(0.02)
anemia	-33.99*
	(-2.30)
diabete	49.95***
	(3.88)
phyper	-168.0***
	(-15.11)
pre4000	466.0***
	(34.93)
preterm	-498.5***
	(-35.07)
alcohol	-61.70
	(-0.82)
drink	15.86
	(1.08)
foreignb	-17.49*
	(-2.06)
plural	-921.9***
	(-73.92)
deadkids	-9.095***
	(-4.20)
mblack	-125.3***
	(-3.87)
motherr	-81.61***
	(-4.19)
mhispan	-63.59***
	(-4.56)
fblack	-62.66*
	(-2.26)
fotherr	-101.6***
	(-5.29)
fhispan	-69.74***
	(-5.37)

tripre1	-174.6***
	(-18.99)
tripre2	-105.6***
	(-11.69)
tripre3	0
	(.)
tripre0	-183.2***
	(-9.15)
first	-82.81***
	(-16.69)
death	-1259.7***
	(-35.16)
dmeduc2	0.171
	(0.83)
dfeduc2	0.141
	(0.78)
dmage_mblack	-1.428
	(-1.16)
dmage_dmar	-0.872
	(-1.13)
dfage_fblack	0.485
	(0.54)
dmage_alcohol	1.354
	(0.52)
dmage_drink	-0.914
	(-1.70)
_cons	3303.8***
	(85.91)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Since it is not clear which higher order terms and which interaction terms to use, we may need to try out several (if not all) of them. By doing so, we hope to eliminate omitted variable bias, but there can be no guarantee that we are indeed doing so. The results of the estimation using a few higher order and interaction terms is presented in Table 5

Benefits and drawbacks of flexible functional form

As I described in the section above, the primary benefit in trying to include as many imaginable higher order and interaction terms is to hope to reduce/eliminate omitted variable bias. The drawbacks though are manifold. The main class of drawbacks arise from the fact that introducing all these higher order and interaction variables may bring with them associated problems. Prominent among them could be multicollinearity between the introduced terms, and potential exacerbation of any measurement errors present in those variables. The other set of problems arising out this approach may even be termed philosophical in that we now hope to have reduced the omitted variable and functional form misspecification but we have no way of how well we are doing on this approach.

The Stata code for this question is as below

```
use smoking_labels, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/

gen dmeduc2 = dmeduc^2
gen dfeduc2 = dfeduc^2
gen dmage_mblack = dmage*mblack
gen dmage_dmar =dmage*dmar
gen dfage_fblack = dfage*fblack
gen dmage_alcohol = dmage*alcohol
gen dmage_drink = dmage*drink

save smoking_labels, replace
reg dbirwt tobacco dmage dmeduc dmar dlivord nprevist disllb dfage dfeduc
    anemia diabete phyper pre4000 preterm alcohol drink foreignb plural
   deadkids mblack motherr mhispan fblack fotherr fhispan tripre1 tripre2
    tripre3 tripre0 first death dmeduc2 dfeduc2 dmage_mblack dmage_dmar
   dfage_fblack dmage_alcohol dmage_drink, vce(robust)
esttab using ‘imagepath’c1.tex, title("Flexible functional form
   regression\label{c1}") mtitle("Birth Weight") longtable replace
```

Question (d)

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

The primary problem we face in estimating the effect of maternal smoking on birth weight is that we lack data on the counterfactual. One way we may get around this problem is by estimating the counterfactual. When we are able to make a strong assumption about random selection conditional on observables, we may then construct a propensity score (which is like a probability of sameness score) so as to then match the treatment and control to those with the most similar propensity scores. By doing so, we overcome the problem of omitted variable bias that I highlighted in the previous section.

Question (e)

Implement the propensity score approach to the evaluation problem using two methods: 1) control directly for the estimated propensity scores in a regression model; 2) use the estimated propensity score in a subclassification scheme. In doing so, use your own stopping rule ? e.g., use the 1% significance level when assessing the balance of the covariates within each block (t-test), and stop the algorithm when fail to reject the equality of mean covariates for over 80% of t-tests within a block. Provide empirical evidence on the overlap of the observables of smokers and non-smokers. Estimate the average treatment affect and the average treatment effect on the treated. Interpret the results.

For this question, I use the `pscore` command in Stata without specifying the number of blocks. Stata generates 31 blocks and highlights that the distribution is unbalanced. However, for this question I carry on with the propensity scores generated by Stata in the variable `ps31`

Regression with controls for propensity scores

Table 6 reports the results of regressing birth weight on `tobacco` and `ps31`.

Table 6: Control for Propensity Score (ps31)

	(1) Birth Weight
tobacco	-213.3***
	(-49.63)
ps31	-318.7***
	(-25.94)
_cons	3476.8***
	(1375.97)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Propensity scores in subclassification scheme

I create a subclassification index as a function of the propensity score `ps31`, the mean propensity score, and `tobacco`. I get 100 subclasses from -19 to 80. Table 7 reports the results of regressing birth weight on `tobacco`, `ps31` and `subclass`.

Evidence of overlap of observables

The box and whisker plot in Figure 1 suggests that there is significant overlap of predicted propensity scores between treatment and control with 31 blocks.

Average Treatment Effect and Average Treatment Effect on the Treated

Since we assume that treatment is selected on observables, we may safely use the propensity score as the probability of being treated. Within each stratum, we would expect that the fraction of treated equals the propensity score. As we note in Figure **??**e4), the scatter plots are clustered

Table 7: Regression with Subclassification on Propensity Score (subclass)

	(1) Birth Weight
tobacco	-208.3***
	(-45.72)
ps31	-295.4***
	(-19.88)
subclass	-0.730**
	(-2.77)
_cons	3473.0***
	(1226.87)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

around the 45 degree line, suggesting that propensity score are reasonably estimated. However, they may be less so at the higher end of the distribution (for values closer to 1).
The Stata code for this question is as below

```
set more off
use smoking_labels, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
pscore tobacco dmage dmeduc dmar dlivord nprevist dislb dfage dfeduc
   anemia diabete phyper pre4000 preterm alcohol drink foreignb plural
   deadkids mblack motherr mhispan fblack fotherr fhispan tripre1 tripre2
    tripre3 tripre0 first death dmeduc2 dfeduc2 dmage_mblack dmage_dmar
   dfage_fblack dmage_alcohol dmage_drink, pscore(ps31) blockid(blo31)
   logit level(0.01)
save smoking.ps31.dta, replace

use smoking.ps31.dta, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
reg dbirwt tobacco ps31, vce(robust)
esttab using `imagepath'e1.tex, title("Control for Propensity Score (ps31
   )\label{e1}") mtitle("Birth Weight") replace

egen mean = mean(ps31)
gen subclass = round(100 * tobacco * (ps31-mean))
drop mean
reg dbirwt tobacco ps31 subclass, vce(robust)
esttab using `imagepath'e2.tex, title("Regression with Subclassification
   on Propensity Score (subclass)\label{e2}") mtitle("Birth Weight")
   replace

graph box ps31, over(tobacco)
```

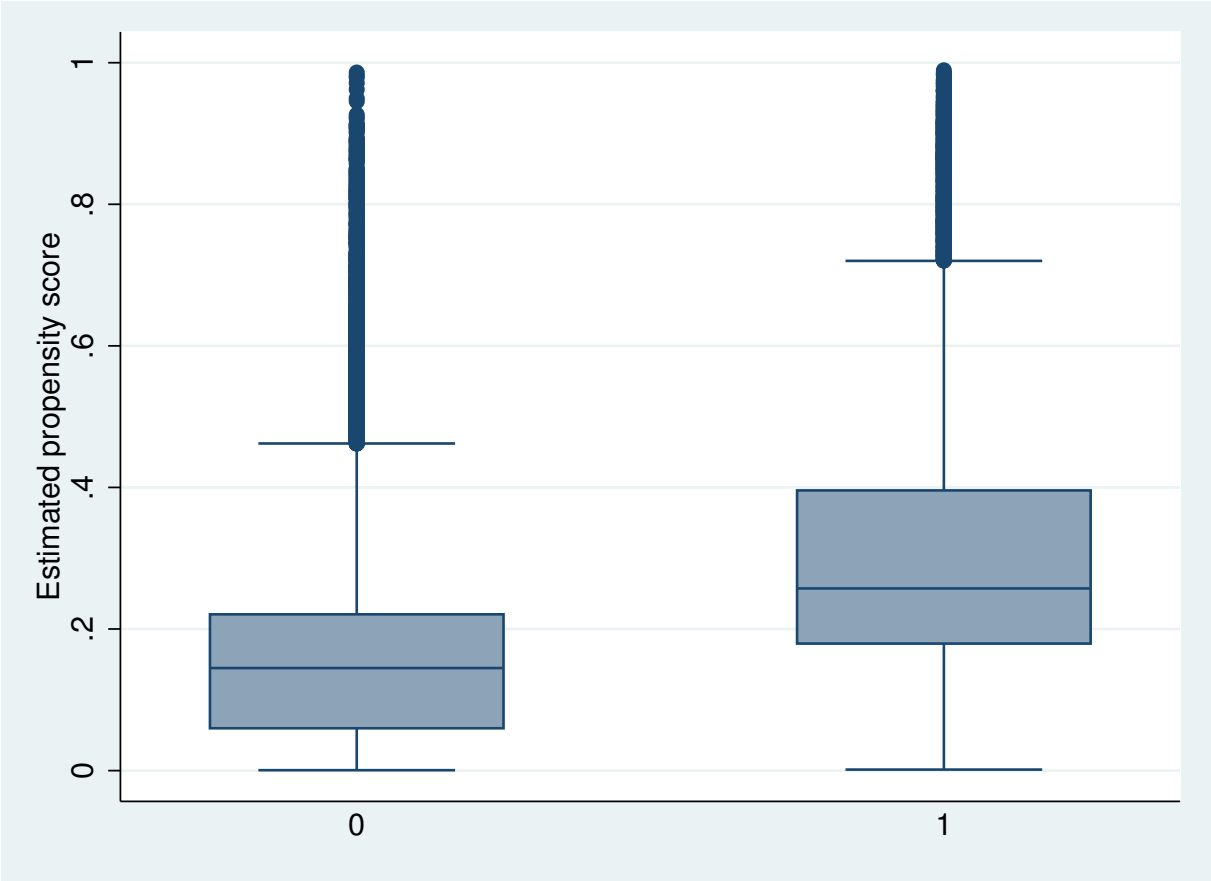

Figure 1: Distribution of Propensity Scores by Treatment

```
graph2tex, epsfile(‘imagepath’e3) ht(5) caption(Distribution of Predicted
    Propensity Scores)

sort blo31
by blo31, sort: egen smoke=count(tobacco) if tobacco==1
by blo31, sort: egen total=count(tobacco)
gen fraction=smoke/total
drop if smoke==.
duplicates drop blo31, force
graph twoway (scatter fraction ps31) (function y=x, range(0 1)), ytitle("
   Predicted Propensity Score") xtitle("Fraction Treated") ///
       title("Distribution of Propensity Scores Across Bins") legend(
          label(1 Propensity Score) label(2 45-Line))
graph2tex, epsfile(‘imagepath’e4) ht(5) caption(Fraction treated by bin)
```

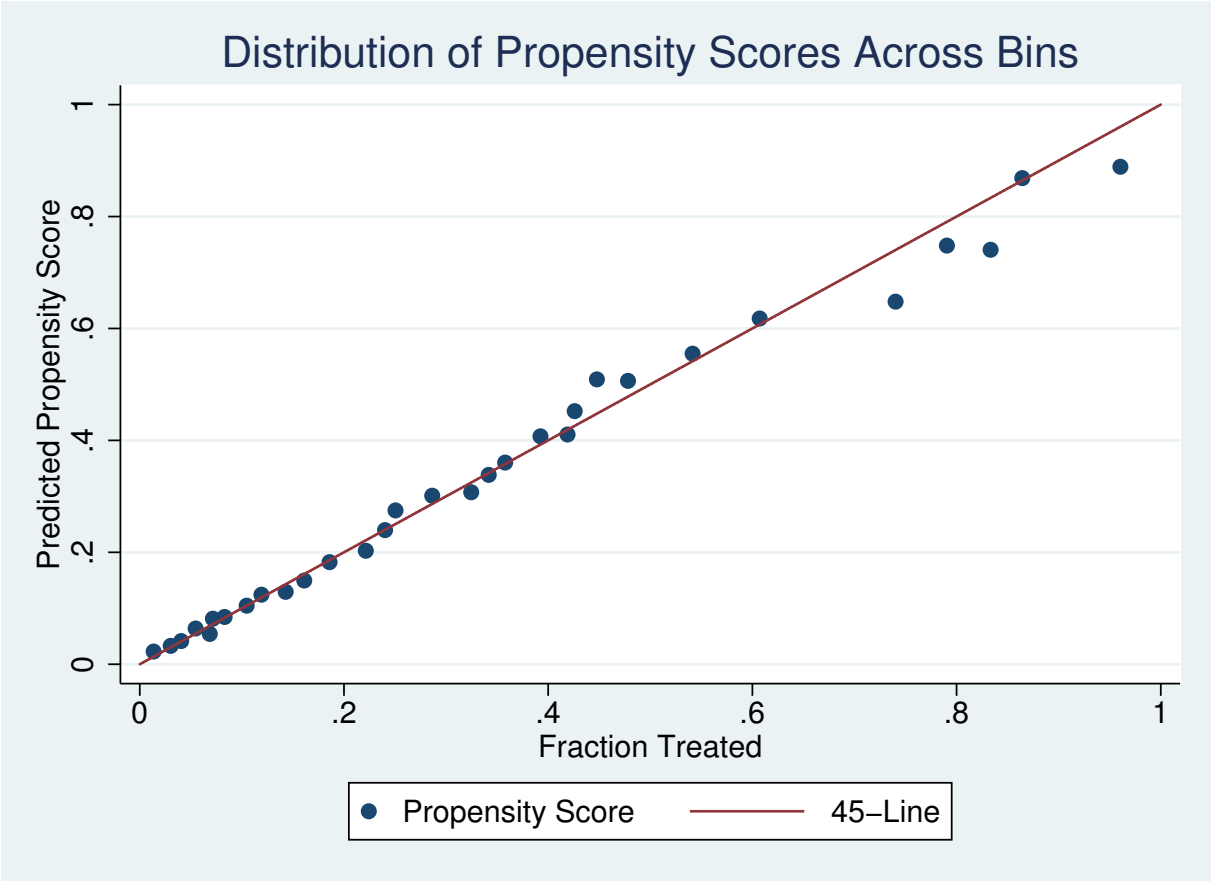

Figure 2: Distribution of Propensity Scores by Bin

Question (f)

Now use the estimated propensity scores as individual weights (and normalize the weights to one) to estimate: i) the average treatment effect. Compare your estimates to those in part (e) and interpret your findings. What are the benefits and drawbacks of approaches that use the estimated propensity scores as individual weights? Use a graph to provide evidence on the appropriateness of the propensity score weighting estimator ? i.e., the sensitivity of the estimated propensity scores. In other words, plot the mean estimated propensity scores (y-axis) against the actual fraction of smokers (x-axis) for 200 equal sized cells of the estimated propensity score (you should produce a graph along the lines of the figure in PAGE 11 of attached PDF Lecture 3). Interpret the results.

Using weights and comparison

Based on the instructions provided, we follow the following instructions.

```
use smoking.ps200.dta, clear
logit tobacco dmage dmeduc dmar dlivord nprevist disllb dfage dfeduc
   anemia diabete phyper pre4000 preterm alcohol drink foreignb plural
   deadkids mblack motherr mhispan fblack fotherr fhispan tripre1 tripre2
    tripre3 tripre0 first death dmeduc2 dfeduc2 dmage_mblack dmage_dmar
   dfage_fblack dmage_alcohol dmage_drink
```

```
predict phat
gen phat_prime = 1-phat
gen wt=sum(1/phat) if tobacco==1
replace wt=sum(1/phat_prime) if tobacco==0
egen maxwt=max(wt)
replace wt=wt/maxwt
reg dbirwt tobacco wt, vce(robust)
esttab using ‘imagepath’f1.tex, title("Regression with individual level
   weights\label{f1}") mtitle("Birth Weight") replace
```

Table 8 provides us with the regression results. What we find is that assuming that no unobservable factors account for smoking choice, that on average smoking during pregnance has a negative impact on birth weight by 210 to 240 across various models.

Table 8: Regression with individual level weights

	(1)
	Birth Weight
tobacco	-257.3***
	(-64.58)
wt	-97.80***
	(-17.80)
_cons	3473.0***
	(1118.98)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Benefits and drawbacks of weighing by estimated propensity scores

Over the past few questions, we experimented with using propensity scores in a number of ways. Specifically, we initially used it as a regressor, then interacted it with treatment and used it as part of a subclassifying mechanism. Finally we have now experimented with using adjustment weights. The advantages of the latter mechanisms are that they are more robust to highly heterogenous samples, whereas the downside is that the individual weighting makes the estimates sensitive to higher values of propensity scores.

Evidence on appropriateness of propensity score weighing estimator

Figure 3 provides us a visual respresentation of the appropriateness of the propensity score weighing estimator, as we see the points clustered around the 45 line.

The Stata code for this question is as below

```
set more off
use smoking_labels, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/

set more off
```

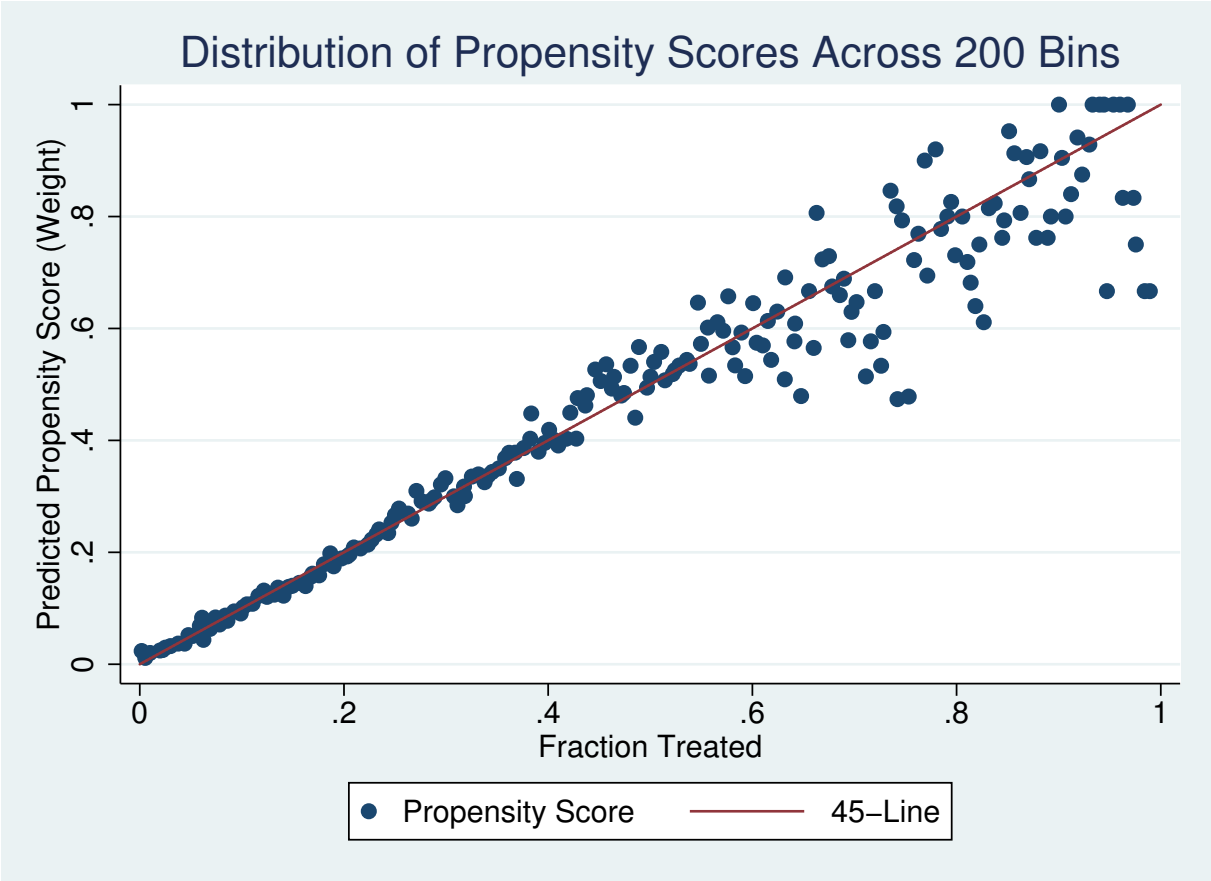

Figure 3: Distribution of Propensity Scores by Bin

```
use smoking_labels, clear
pscore tobacco dmage dmeduc dmar dlivord nprevist disllb dfage dfeduc
    anemia diabete phyper pre4000 preterm alcohol drink foreignb plural
    deadkids mblack motherr mhispan fblack fotherr fhispan tripre1 tripre2
     tripre3 tripre0 first death dmeduc2 dfeduc2 dmage_mblack dmage_dmar
    dfage_fblack dmage_alcohol dmage_drink, pscore(ps200) blockid(blo200)
    logit level(0.005) numblo(201)
save smoking.ps200.dta, replace

set more off
use smoking.ps200.dta, clear
logit tobacco dmage dmeduc dmar dlivord nprevist disllb dfage dfeduc
    anemia diabete phyper pre4000 preterm alcohol drink foreignb plural
    deadkids mblack motherr mhispan fblack fotherr fhispan tripre1 tripre2
     tripre3 tripre0 first death dmeduc2 dfeduc2 dmage_mblack dmage_dmar
    dfage_fblack dmage_alcohol dmage_drink
predict phat
gen phat_prime = 1-phat
gen wt=sum(1/phat) if tobacco==1
replace wt=sum(1/phat_prime) if tobacco==0
egen maxwt=max(wt)
replace wt=wt/maxwt
```

```
reg dbirwt tobacco wt, vce(robust)
esttab using ‘imagepath’f1.tex, title("Regression with individual level
   weights\label{f1}") mtitle("Birth Weight") replace

sort blo200
by blo200, sort: egen smoke=count(tobacco) if tobacco==1
by blo200, sort: egen total=count(tobacco)
gen fraction=smoke/total
drop if smoke==.
duplicates drop blo200, force
graph twoway (scatter fraction ps200) (function y=x, range(0 1)), ytitle
   ("Predicted Propensity Score (Weight)") xtitle("Fraction Treated") ///
       title("Distribution of Propensity Scores Across 200 Bins") legend(
          label(1 Propensity Score) label(2 45-Line))
graph2tex, epsfile(‘imagepath’f2) ht(5) caption(Fraction treated by bin)
```

Question (g)

A more general (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, for smokers and non-smokers. To do this, stratify the smokers into 100 equal-sized cells based on their estimated propensity scores and calculate the mean birth weight and the mean estimated propensity score in each cell. Do the same for the non-smokers. Plot these two conditional mean functions on the same graph, with the mean estimated propensity scores on the x-axis and the mean birth weight on the y-axis. Interpret your findings and relate them to the results in part (e) and (f).

The above instructions have been followed in the following code.

```
use smoking.ps100.dta, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
egen mps100=mean(ps100), by(blo100)
bysort blo100: egen m_smok=mean(dbirwt) if tobacco==1
bysort blo100: egen m_nosmok=mean(dbirwt) if tobacco==0
label variable m_smok "Smokers"
label variable m_nosmok "Non-Smokers"
tw (scatter m_smok mps100) (scatter m_nosmok mps100), ytitle(birth weight
    in grams) xtitle(mean propensity scores in 100 bins)
graph2tex, epsfile(‘imagepath’g1) ht(5) caption(Comparison of Treatment
   and Control by Propensity Score (100 bins))
```

Figure 4 is a self-explanatory graph, that confirms the effect of smoking on birth weight across propensity scores.

Interpretation of Findings

From Figure 4, we note that the propensity score method is better at correcting for selection bias at lower values of the propensity score, but it is less so at the upper end. Therefore we may reasonably conclude that the estimation of average treatment effect on the treated may be less reliable and more vulnerable at higher propensity score levels. Finally in using the results from questions (e) and (f) we may calculate the gain from using the propensity score by taking the difference of the average treatment effect and the effect from the classification scheme used in question (e).

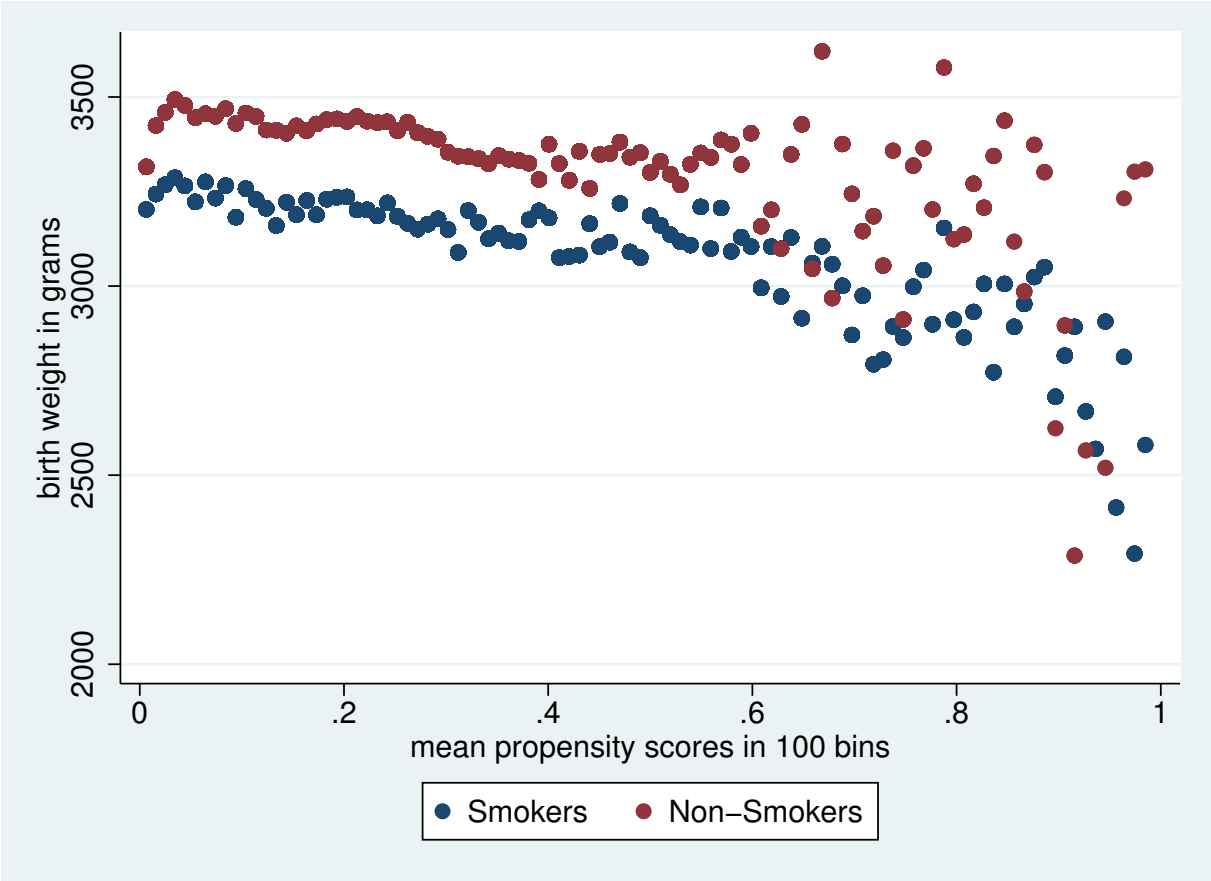

Figure 4: Distribution of Propensity Scores by Bin

Question (h)

Low birth weight births (less than 2,500 grams) are considered particularly undesirable since they comprise a large share of infant deaths. Redo part (f) and (g) using an indicator for low birth weight as the outcome of interest. Interpret your findings.

Tables 9 and 10 lay out the results of the estimation of maternal smoking on low birth weight. We see that both the 100 bin and the 200 bin sample show a statistically significant coefficient estimate of 0.0387, implying that there is a 3.87% impact on the probability of low birth weight among smoking mothers as compared those that did not smoke during pregnancy.
The Stata code for this question is as below

```
use smoking.ps100.dta, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
gen lowbirwt = 1 if dbirwt <= 2500
replace lowbirwt = 0 if dbirwt > 2500
atts lowbirwt tobacco, pscore(ps100) blockid(blo100)
reg lowbirwt tobacco ps100, vce(robust)
esttab using ‘imagepath’h1.tex, title("Propensity Score Regression (100
   bins)\label{h1}") mtitle("Low Birth Weight") replace
```

Table 9: Propensity Score Regression (100 bins)

	(1) Low Birth Weight
tobacco	0.0387***
	(18.81)
ps100	0.0966***
	(16.99)
_cons	0.0334***
	(31.45)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Propensity Score Regression (200 bins)

	(1) Low Birth Weight
tobacco	0.0387***
	(18.81)
ps200	0.0966***
	(16.99)
_cons	0.0334***
	(31.45)
N	139149

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

```
use smoking.ps200.dta, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
gen lowbirwt = 1 if dbirwt <= 2500
replace lowbirwt = 0 if dbirwt > 2500
atts lowbirwt tobacco, pscore(ps200) blockid(blo200)
reg lowbirwt tobacco ps200, vce(robust)
esttab using `imagepath'h2.tex, title("Propensity Score Regression (200
   bins)\label{h2}") mtitle("Low Birth Weight") replace
```

Question (i)

Estimate the impact of maternal smoking on infant death (indicator for death) using the methods in parts (b) and (f). Interpret your findings. From your results, what might you conclude about the relationship between smoking-induced low birth weight and infant death?

As indicated in question (b), we first estimate the effect of maternal smoking on infant death under the assumption that maternal smoking is randomly assigned conditional on the observable variables. Table 11 lays out the results, where we observe a statistically significant but close to zero estimate (the estimate is -0.006).

Table 11: tobacco Randomly Assigned Conditional on Observables

	(1) Infant Death
tobacco	-0.00666***
	(-9.24)
dbirwt	-0.0000307***
	(-25.28)
dmage	-0.0000982
	(-1.29)
dmeduc	0.0000718
	(0.50)
dmar	0.00121
	(1.59)
dlivord	0.000667
	(1.87)
nprevist	-0.000894***
	(-9.03)
disllb	0.000000535
	(0.05)
dfage	0.0000568
	(0.94)
dfeduc	-0.0000454
	(-0.34)
anemia	0.00183
	(0.66)

diabete	0.00506**
	(2.74)
phyper	-0.00294*
	(-1.97)
pre4000	0.0127***
	(7.83)
preterm	0.00280
	(0.84)
alcohol	-0.00227
	(-1.15)
drink	-0.000776*
	(-2.41)
foreignb	-0.00409***
	(-3.72)
plural	-0.00369
	(-1.03)
deadkids	0.00193***
	(5.04)
mblack	-0.00826***
	(-3.38)
motherr	-0.00270
	(-0.87)
mhispan	-0.00432*
	(-2.25)
fblack	0.00268
	(1.12)
fotherr	0.00199
	(0.59)
fhispan	0.000860
	(0.44)
tripre1	0.00658***
	(4.02)
tripre2	0.00433**

	(2.68)
tripre3	0
	(.)
tripre0	0.0298***
	(5.22)
first	-0.00108
	(-1.29)
_cons	0.115***
	(23.58)
N	139149

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

As indicated in question (f), we first estimate the effect of maternal smoking on infant death taking individual weights for calculating propensity scores. Table 12 lists the results, where we observe a statistically significant but close to zero estimate (the estimate is +0.002).

Table 12: Regression with individual level weights

	(1)
	Infant Death
tobacco	0.00272***
	(4.26)
wt	0.00263**
	(3.18)
_cons	0.00520***
	(11.48)
N	139149

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Unlike in the case of estimating the effects of maternal smoking on birth weight where we saw consistent results, this is not quite the case with the impact of maternal smoking on infant mortality. Presumably we need to rethink the model, and if our assumptions of exogeneity of unobservable factors continues to hold in this context as well.
The Stata code for this question is as below

```
use smoking_labels, clear
local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
set more off
reg death tobacco dbirwt dmage dmeduc dmar dlivord nprevist disllb dfage
   dfeduc anemia diabete phyper pre4000 preterm alcohol drink foreignb
```

```
   plural deadkids mblack motherr mhispan fblack fotherr fhispan tripre1
   tripre2 tripre3 tripre0 first , vce(robust)
esttab using ‘imagepath’i1.tex, title("tobacco Randomly Assigned
   Conditional on Observables\label{i1}") mtitle("Infant Death")
   longtable replace


local imagepath /Users/anu/OneDrive/code/articles/adv-eco-hw4-images/
set more off
use smoking.ps200.dta, clear
logit tobacco dmage dmeduc dmar dlivord nprevist disllb dfage dfeduc
   anemia diabete phyper pre4000 preterm alcohol drink foreignb plural
   deadkids mblack motherr mhispan fblack fotherr fhispan tripre1 tripre2
    tripre3 tripre0 first death dmeduc2 dfeduc2 dmage_mblack dmage_dmar
   dfage_fblack dmage_alcohol dmage_drink
predict phat
gen phat_prime = 1-phat
gen wt=sum(1/phat) if tobacco==1
replace wt=sum(1/phat_prime) if tobacco==0
egen maxwt=max(wt)
replace wt=wt/maxwt
reg death tobacco wt, vce(robust)
esttab using ‘imagepath’i2.tex, title("Regression with individual level
   weights\label{i2}") mtitle("Infant Death") replace
```

Question (j)

Concisely and coherently summarize all of your findings. In this summary, describe the estimated effects of maternal smoking on birth weight and infant mortality and whether the causal effect of maternal smoking is credibly identified. State why or why not.

Using the sample of 139149 samples of data about mothers and infants, we initially estimated a -260 gram effect on infant birth weight due to maternal smoking, assuming that maternal smoking was randomly assigned. We then took several steps to correct for that, eventually narrowing down on an individual level weighted propensity score that provided us with an estimate of -210 gram effect on infant birth weight due to maternal smoking. However, we make some strong assumptions in reaching this estimate, including that maternal smoking is not affected by any unobserved variables. Additionally we also assume that the outcome of infant birth weight is not endogenous to the choice of smoking choice by the mother. The propensity score is one way to overcome the problem of potential heterogeniety in the treatment effect across mothers with different propensity score levels. While this seems like a step forward, the large question remains if our assumption about conditionality of selection only upon the observables is a valid one. In the affirmative, the propensity score matching approach is valuable in explicating the causal relationship here between maternal smoking and infant birth weight. However, the impact on infant mortality seems less clear.