Final

Hunter Garfield
Microeconometrics

Problem 1

```
ols1 = lm(dbirwt~tobacco, data=d)
stargazer(ols1, type = "html", summary=F, style = "aer", single.row = T)
```

	dbirwt
tobacco	-266.029*** (4.762)
Constant	3,423.360*** (1.987)
Observations	100,000
R^2	0.030
Adjusted R^2	0.030
Residual Std. Error	571.127 (df = 99998)
F Statistic	3,121.286*** (df = 1; 99998)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

```
ols2 = lm(dbirwt~., data = d)
stargazer(ols2, type = "html", summary=F, style = "aer", single.row = T)
```

	dbirwt
alcohol	-6.066 (17.470)
anemia	-66.132*** (17.878)
cardiac	-47.541* (24.633)
chyper	-175.809*** (19.191)
dfage	-0.226 (0.416)
dfeduc	5.673*** (0.995)
diabete	50.848*** (12.730)
dislb	-0.170** (0.073)
dlivord	26.954*** (2.244)
dmage	10.023*** (2.758)
dmar	-46.760*** (5.308)
dmeduc	5.072*** (1.106)
drink	-16.060*** (3.006)
foreignb	-24.982** (10.016)
nprevist	35.990*** (0.682)
pre4000	466.768*** (16.615)

tobacco	-231.976*** (4.729)
mblack	-144.013*** (13.748)
motherr	-70.489*** (22.161)
mhispan	-84.102*** (15.705)
fblack	-67.351*** (13.410)
fotherr	-120.813*** (21.732)
fhispan	-51.271*** (14.510)
adequac2	58.423*** (7.375)
adequac3	139.622*** (15.286)
tripre2	38.796*** (7.833)
tripre3	69.325*** (16.778)
tripre0	-119.551*** (21.648)
first	-83.797*** (6.019)
plural	-973.229*** (13.459)
dmage2	-0.225*** (0.048)
Constant	2,805.360*** (39.863)
Observations	100,000
R^2	0.155
Adjusted R^2	0.155
Residual Std. Error	533.096 (df = 99968)
F Statistic	593.197*** (df = 31; 99968)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

The estimated ATE of smoking during pregnancy on birthweight from the first regression is -266. This is only a valid ATE if we assume that the value of the intercept is the same regardless of the treatment group. The estimated ATE of smoking during pregnancy on birthweight from the second regression is -231.98. This is only a valid ATE if we assume that the intercept and the coefficients on all of the other covariates are the same regardles of the treatment group. These are extremely strong assumptions and probably do not hold up here.

Problem 2

```
dtob = d[which(d$tobacco==1),] #separates into smokers and non smokers
dtob = dtob[,-18]
dno = d[which(d$tobacco==0),]
dno = dno[,-18]
mtob = colMeans(dtob)
mno = colMeans(dno)
mdiff = mtob-mno #find difference in means

ts = numeric(ncol(dtob))
pvs = numeric(ncol(dtob))
for(i in 1:ncol(dtob)){ #perform t test for each covariate
  t = t.test(dtob[,i],dno[,i],mu=0,conf.level = .95)
  ts[i] = t$statistic
  pvs[i] = t$p.value
}

p2t = data.frame(cbind(mtob, mno, mdiff, ts, pvs))
p2t = p2t[-5,]
colnames(p2t) = c(".Mean Tobacco      ","Mean No Tobacco      ","Difference in M
eans      ","Test Statistic      ","Pr>|t|   ")
stars = add.significance.stars(p2t[,5])
p2t[,3] = as.character(p2t[,3])
p2t[,3] = strtrim(p2t[,3],width = 7)
p2t[,3] = paste(p2t[,3],stars,sep="") #add stars to diff in means column

p2t = stargazer(p2t, type = "html",title = "Summary Mean Characteristics",summary
= F,digits = 4,column.sep.width = "50pt")
```

Summary Mean Characteristics

	.Mean Tobacco	.Mean No Tobacco	.Difference in Means	.Test Statistic	Pr> | t|
alcohol	0.0470	0.0067	0.04026 * * *	24.7219	0
anemia	0.0148	0.0078	0.00691 * * *	7.1788	0
cardiac	0.0042	0.0048	-0.0005	-1.0207	0.3074
chyper	0.0059	0.0083	-0.0023 * * *	-3.5511	0.0004
dfage	28.4809	29.6767	-1.1957 * * *	-22.3416	0
dfeduc	12.0195	13.2813	-1.2617 * * *	-81.7795	0
diabete	0.0187	0.0180	0.00074	0.6593	0.5097
dislb	30.1553	24.5478	5.60743 * * *	18.3066	0
dlivord	2.1219	1.9360	0.18590 * * *	18.5628	0
dmage	25.5784	27.3965	-1.8180 * * *	-40.4420	0
dmar	0.4251	0.1865	0.23860 * * *	59.9016	0
dmeduc	11.8881	13.1881	-1.2999 * * *	-88.6500	0
drink	0.1774	0.0157	0.16174 * * *	14.5044	0
foreignb	0.0170	0.0531	-0.0361 * * *	-28.8568	0
nprevist	10.2898	11.1918	-0.9019 * * *	-27.9155	0
pre4000	0.0061	0.0115	-0.0053 * * *	-7.6424	0
mblack	0.1339	0.1095	0.02436 * * *	8.7042	0

motherr	0.0033	0.0182	-0.0148 * * *	-23.3291	0
mhispan	0.0163	0.0285	-0.0122 * * *	-10.9101	0
fblack	0.1502	0.1169	0.03326 * * *	11.3608	0
fotherr	0.0041	0.0169	-0.0127 * * *	-19.3241	0
fhispan	0.0245	0.0303	-0.0057 * * *	-4.3752	0.00001
adequac2	0.2565	0.1786	0.07788 * * *	21.8343	0
adequac3	0.1003	0.0480	0.05232 * * *	21.8481	0
tripre2	0.2015	0.1208	0.08076 * * *	24.8980	0
tripre3	0.0464	0.0244	0.02194 * * *	13.0494	0
tripre0	0.0226	0.0073	0.01535 * * *	13.1811	0
first	0.3667	0.4394	-0.0726 * * *	-17.9939	0
plural	0.0155	0.0161	-0.0006	-0.5877	0.5568
dmage2	683.0094	781.1418	-98.132 * * *	-40.4348	0

This table confirms our reservations about the assumptions made above, that the coefficients on many of these covariates obtained through OLS regression are proabably not the same between the treatment (smoking) and control (no smoking) groups. We can tell that this is proabably the case because the difference in means of almost all of these covariates between the treatment and control groups is significantly different from zero with a pretty high level of confidence (all except cardiac, diabete, and plural, so 27 in total). This table suggests that there are also likely to be many unobserved factors in which the means differ significantly between smoking and non-smoking mothers, which tells us that we can not compare these two groups on the basis of this indicator (smoking can not be thought of as randomly assigned). Therefore, running an OLS regression of birth weight on the smoking indicator would not yield an unbiased estimate of the ATE, because the smoking and non-smoking groups of mothers are not comparable.

Problem 3

The main assumption used to conclude that pscore methods estimtate the ATE is that the treatment is unconfounded conditional on the pscore. This relies on the assumption of unconfoundedness, which states that there are no unobserved variables that effect the treatment and the outcome, controlling for all observed covariates. Once we can conclude that the treatment is unconfounded conditional on the covariates, we can also conclude that the treatment is unconfounded on the pscore, which is the probability of receiving treatment conditional on the covariates (captures conditional effect of covariates). In the context of this data set, unconfoundedness would imply that there is nothing outside of our list of covariates (alcohol, dmage, etc.) that effects whether or not a mother smokes during pregnancy and her child's birthweight. Pscore methods do not control for unobserved confounders affecting smoking during pregnancy or birthweight, but they do not need to since, by unconfoundedness, we assume that there are no such unobserved confounders. The advantage of using a pscore here is that we can adjust the treatmnt using a single value rather than conditioning it on all 30 other covariates. In relation to OLS, pscore methods are useful because unlike OLS, they do not rely on the assumption that the effects of the treatment on both the outcome and the covariates are linear and additive, which is very strong and not usually plausible. Pscore methods are less parametric and allow for more flexibility in estimation.

Problem 4

```
lgt = glm(tobacco~. -dbirwt, family = binomial(link = "logit"), data = d)
  #logit of covariates on tobacco (excludes birthweight)
stargazer(lgt, type = "html", summary=F, style = "aer", single.row = T)
```

	tobacco
alcohol	1.813*** (0.082)
anemia	0.340*** (0.081)
cardiac	0.008 (0.135)
chyper	-0.270** (0.112)
dfage	0.024*** (0.002)
dfeduc	-0.114*** (0.005)
diabete	0.224*** (0.066)
dislb	0.006*** (0.000)
dlivord	-0.000 (0.011)
dmage	0.238*** (0.015)
dmar	1.103*** (0.024)
dmeduc	-0.178*** (0.006)
drink	0.064*** (0.020)
foreignb	-0.711*** (0.072)
nprevist	-0.014*** (0.004)
pre4000	-0.782*** (0.109)
mblack	-0.698*** (0.066)
motherr	-1.021*** (0.182)
mhispan	-1.083*** (0.094)
fblack	-0.085 (0.064)
fotherr	-0.636*** (0.162)
fhispan	-0.369*** (0.078)
adequac2	0.066* (0.038)
adequac3	-0.016 (0.073)
tripre2	0.115*** (0.038)
tripre3	0.148* (0.076)
tripre0	0.370*** (0.094)
first	-0.141*** (0.032)
plural	0.025 (0.073)
dmage2	-0.005*** (0.000)
Constant	-1.425*** (0.210)
Observations	100,000
Log Likelihood	-40,403.250
Akaike Inf. Crit.	80,868.490

Notes: ***Significant at the 1 percent level.
**Significant at the 5 percent level.

*Significant at the 10 percent level.

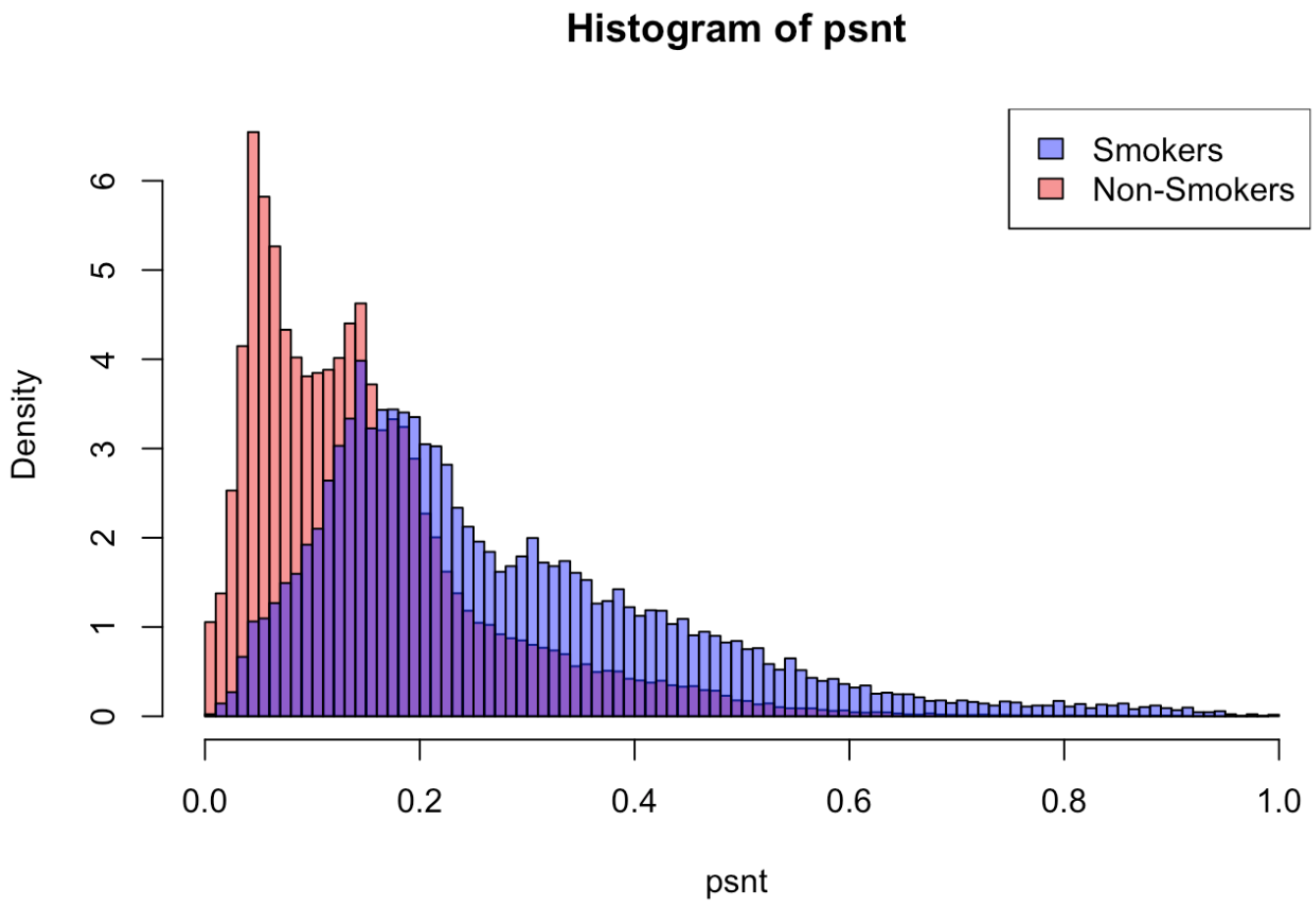
```
lcfs = lgt$coefficients
#pscores can be calculated as:
ps = plogis(as.matrix(cbind(numeric(nrow(d))+1,d[,-c(5,18)]))%*%as.matrix(lcfs))

ps = as.matrix(predict(lgt, type = "response"))
  #both of these methods give equivalent results, a built in function gives the sa
me:
pscores = pscore(data = d, formula = tobacco~. -dbirwt)
```

Problem 5

```
pwt = cbind(ps, d$tobacco)
pst = ps[which(pwt[,2]==1)]
psnt = ps[which(pwt[,2]==0)] #separates pscores into tobacco/no tobacco
hnt = hist(psnt,breaks = 75,freq = F,col = rgb(1,0,0,1/2))
ht = hist(pst, breaks = 75,freq=F, col = rgb(0,0,1,1/2),add = T)

legend("topright", legend = c("Smokers","Non-Smokers"),fill = c(rgb(0,0,1,1/2),rgb
(1,0,0,1/2)))
```



The histogram of the pscores tells us that these two groups are somewhat comparable in terms of their covariates. To be more precise, the groups are comparable on pscore in the pink-shaded region of the plot. We also can be reasonably sure that there will be very few, if any, pscores deleted from the left side of the distribution because the pscores of both treatment groups seem to overlap at very close to zero.

```
## [1] "Number of non-smokers: 82579"
```

```
## [1] "Number of smokers: 17421"
```

```
## [1] "quantiles (non-smoker vs smoker)"
```

```
##          1%          5%         10%         25%         50%         75%
## 0.009354032 0.030119840 0.041518911 0.066354607 0.128382619 0.196156501
##         90%         95%         99%
## 0.309626389 0.392645675 0.536313170
```

```
##         1%         5%        10%        25%        50%        75%
## 0.03898280 0.07323604 0.10252877 0.15125910 0.22826077 0.36704768
##        90%        95%        99%
## 0.50995154 0.61105348 0.84419892
```

```
## [1] "other descriptive stats (top row is non-smoker)"
```

```
## INDICES: 0
##    vars     n mean   sd median trimmed  mad min  max range skew kurtosis
## V1    1 82579 0.15 0.12   0.13    0.14 0.09   0 0.98  0.98 1.62     3.55
## V2    2 82579 0.00 0.00   0.00    0.00 0.00   0 0.00  0.00  NaN      NaN
##    se
## V1  0
## V2  0
## ------------------------------------------------------------
## INDICES: 1
##    vars     n mean   sd median trimmed  mad min max range skew kurtosis se
## V1    1 17421 0.28 0.17   0.23    0.26 0.14   0   1  0.99  1.2     1.45  0
## V2    2 17421 1.00 0.00   1.00    1.00 0.00   1   1  0.00  NaN      NaN  0
```

The output from these descriptive statistics (especially the min and max) tells us that the pscores of smokers and non-smokers have the same minimum (zero) so there will be no trimming of pscores from below. The max pscore of non-smokers was .98 and the max pscore of smokers was 1, so we know that we will be trimming off some observations of smokers.

Problem 6

Based on histogram, we want the min of smokers and the max of non-smokers:

```
mint = min(pst)
maxnt = max(psnt)
exclmin = which(ps<=mint)
exclmax = which(ps>=maxnt)
cps = pwt[-c(exclmin,exclmax),] #cuts off pscores below the min of tobacco
    #and above the max of no tobacco
length(cps[,1])
```

```
## [1] 99715
```

Trimming the pscores using the min-max rule results in us deleting 100,000-99,715 = 285 observations. There were actually some observations trimmed from below because the mins of the groups weren't exactly zero, and smokers had a higher min. The reason that we trim off the pscores that do not lie in the overlap region (pink region) is that these individuals do not have comparable counterparts in the opposite treatment group. Without being able to compare individuals with similar pscores across treatment groups, we can not estimate the ATE. After driopping individuals that do not lie in the overlap region, we are no longer estimating the ATE for the entire sample, instead, we estimate the ATE for the individuals that lie within the overlap region.

Problem 7

The code for this problem has not been included because it is the same as in problem two, but with the trimmed data set. Data was weighted using:

```
dovtobw = dovtob/psovt
dovnow = dovno/(1-psovnt)
```

Where dovtob and dovno are the observations in the overlap region for smokers and non-smokers, respectively, and psovt and psovnt are the pscores in the overlap region for smokers and non-smokers, respectively.

```
stargazer(p7t, type = "html", summary=FALSE,t.auto=FALSE)
```

	.Mean Tobacco	.Mean No Tobacco	.Difference in Means	.Test Statistic	.Pr> | t| .
alcohol	0.079	0.017	0.06204 * * *	18.703	0
anemia	0.053	0.011	0.04165 * * *	9.961	0
cardiac	0.023	0.006	0.01743 * * *	5.445	0.00000
chyper	0.042	0.010	0.03222 * * *	6.276	0
dfage	164.019	35.787	128.231 * * *	92.319	0
dfeduc	70.955	15.817	55.1379 * * *	89.941	0
diabete	0.101	0.022	0.07862 * * *	10.394	0
dislb	145.500	31.120	114.379 * * *	53.103	0
dlivord	11.170	2.398	8.77239 * * *	91.342	0
dmage	151.523	32.878	118.645 * * *	85.709	0
dmar	1.297	0.278	1.01931 * * *	66.710	0
dmeduc	70.741	15.701	55.0406 * * *	88.306	0
drink	0.254	0.057	0.19730 * * *	10.526	0
foreignb	0.265	0.054	0.21105 * * *	9.294	0
nprevist	60.889	13.393	47.4962 * * *	91.803	0
pre4000	0.055	0.013	0.04215 * * *	6.286	0
mblack	0.721	0.139	0.58212 * * *	26.246	0
motherr	0.072	0.016	0.05621 * * *	4.041	0.0001
mhispan	0.174	0.032	0.14179 * * *	9.453	0
fblack	0.781	0.150	0.63143 * * *	28.114	0
fotherr	0.050	0.015	0.03502 * * *	3.375	0.001
fhispan	0.197	0.036	0.16180 * * *	10.625	0
adequac2	1.129	0.234	0.89481 * * *	43.014	0
adequac3	0.358	0.071	0.28768 * * *	25.626	0
tripre2	0.814	0.165	0.64873 * * *	38.661	0
tripre3	0.179	0.035	0.14385 * * *	16.716	0
tripre0	0.062	0.012	0.05000 * * *	11.880	0
first	2.376	0.517	1.85937 * * *	51.241	0
plural	0.088	0.019	0.06892 * * *	9.825	0

dmage2	4,312.944	927.288	3385.65 * * *	67.187	0

It appears that weighting the observations by pscore actually increased the difference in means between the two groups. These two groups are far less comparable now after implementing the weights, considering that the difference in means across treatment groups for all covariates is statistically different from zero. It is not hard to see why this might be the case. For both smokers and non-smokers, most of the pscores are centered around ~ 0.15. Therefore, when we divide a smoker observation by a typical smoker's pscore (e.g. obs./.15), we obtain a much higher number than we had before. However, when we divide a non-smoker observation by one minus a typical non-smoker pscore, we are dividing by a much bigger number (e.g. 1-.15 = .85) and therefore the observation doesn't increase as much. This creates a larger disparity between the smoker and non-smoker observations, making the mean differences greater. In this case, it appears that weighting by pscore actually doesn't help to balance the covariates between treatment and control groups.

Problem 8

```
mat = Match(Y = dov$dbirwt, Tr = dov$tobacco, X = cps[,1], estimand = "ATE", M =
1, version = "fast", ties = F)
    #^requires "Matching", matches data based on pscore (X). M=1 finds the first c
losest match.             version = fast and ties = F speed up the function signif
icantly.
(mat$est)
```

```
##           [,1]
## [1,] -230.7435
```

The estimate of the ATE (overlap region) using matching on the propensity score is -231 (sometimes outputs -232 due to the function). This is very similar to the result obtained in problem 1, where we used simple linear regression with all of the covariates included (-231.98). The main drawback of using the matching approach is that it takes a very long time to find matches for all 17416 individuals in the smoking sample, even when we are matching only on pscore. This problem gets significantly worse when if we decide to find more than one closest match for each individual. Another disadvantage is having to decide how many nearest matches to obtain. Increasing the number of near matches decreases the variance of the estimated outcome of each individual, but it also increases the bias in the estimate because of the dereased quality of the matches.

Problem 9

```
psov = cps[,1]
wols = lm(dbirwt~tobacco, data=dov, weights = ((dov$tobacco/psov)+((1-dov$tobacc
o)/(1-psov))))
  #dov is overlap data, psov are overlap pscores
stargazer(wols, type = "html", summary=F, style = "aer", single.row = T)
```

	dbirwt
tobacco	-232.552*** (3.632)
Constant	3,417.750*** (2.549)
Observations	99,715
R^2	0.039
Adjusted R^2	0.039
Residual Std. Error	806.031 (df = 99713)
F Statistic	4,099.948*** (df = 1; 99713)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

The estimate of the ATE (overlap region) using weighting by pscore (specifically a weighted OLS) is -232.55.

Problem 10

```
mwols = wols = lm(dbirwt~., data=dov, weights = ((dov$tobacco/psov)+((1-dov$tobacc
o)/(1-psov))))
stargazer(mwols, type = "html", summary=F, style = "aer", single.row = T)
```

	dbirwt
alcohol	9.825 (18.458)
anemia	-85.104*** (17.768)
cardiac	-57.035** (25.676)
chyper	-197.373*** (19.569)
dfage	-0.605 (0.401)
dfeduc	6.961*** (1.013)
diabete	45.626*** (12.868)
disllb	-0.015 (0.071)
dlivord	25.571*** (2.260)
dmage	9.462*** (2.643)
dmar	-41.218*** (5.071)
dmeduc	7.234*** (1.121)
drink	-22.110*** (3.502)
foreignb	17.334* (9.555)
nprevist	36.459*** (0.676)
pre4000	452.901*** (17.021)

tobacco	-222.709*** (3.421)
mblack	-106.341*** (12.949)
motherr	22.580 (18.542)
mhispan	-89.205*** (14.679)
fblack	-90.497*** (12.625)
fotherr	-108.109*** (19.123)
fhispan	-27.570** (13.394)
adequac2	46.446*** (7.413)
adequac3	138.468*** (15.127)
tripre2	44.327*** (7.811)
tripre3	69.325*** (16.467)
tripre0	-133.259*** (21.157)
first	-72.714*** (6.096)
plural	-951.645*** (13.621)
dmage2	-0.235*** (0.046)
Constant	2,779.422*** (38.617)
Observations	99,715
R^2	0.157
Adjusted R^2	0.156
Residual Std. Error	755.356 (df = 99683)
F Statistic	597.621*** (df = 31; 99683)
Notes:	***Significant at the 1 percent level.
	**Significant at the 5 percent level.
	*Significant at the 10 percent level.

Using a mixture of weighting and OLS gives us an estimate of the ATE of -222.71 (for the overlap region). The advantage of using a mix of weighting and OLS as compared to just OLS is that it allows us to further balance the remaining unbalanced covariates not accounted for through a simpler weighted OLS. It also increases the efficiency of the estimates.

Problem 11

```
(ols2$coefficients[18])
```

```
##   tobacco
## -231.9765
```

```
(mat$est)
```

```
##           [,1]
## [1,] -230.7435
```

```
(mwols$coefficients[18])
```

```
##   tobacco
## -222.7095
```

The estimates of the ATE obtained using linear regression, matching on pscore, and mixed weighting are -231.98, -232.09, and -222.71, repsectively. Ideally, we would like to trust the estimates that are based on pscore matching and mixed weighting more than we do the estimate using simple linear regression. In this case, however, it appears that using pscore to weight observations was not very helpful because it actually made the treatment and control groups less comparable (as in Problem 7). We may be able to trust the estimates obtained from matching a bit more because this method only compares individuals on pscore rather than weighting the observations, which led to problems as discussed in problem 7. Even then, matching does not provide a much different estimate from OLS, especially considering how time consuming it is to run the program. Based on these results, I think it is prudent to use either the OLS estimate or the matching estimate, but not the estimate obtained through weighted regression since it relies on pscore weights that do not make the groups comparable.

Problems 12 & 13

*See attached calculations sheets