# ARE 213 Problem Set 1b

*Nick Depsky, Will Gorman, Peter Worley*

*October 12, 2018*

Import Data

```r
setwd("~/Dropbox/Berkeley_tings/Fall 2018/ARE213/Problem Sets/SharedFiles/are213/PS1b")
#setwd("C:\\Users\\will-\\Desktop\\are213\\PS1b")
#setwd("C:\\Users\\Will\\Desktop\\are213\\PS1b")
dat <- read.dta("ps1.dta")
#fix missing data
dat_drop <- dat %>% filter(herpes != 8 & tobacco != 9 & cigar != 99 & cigar6 != 6 &
                  alcohol != 9 & drink != 99 & drink5 != 5 & wgain != 99)

dat_drop$tobacco_p <- ifelse(dat_drop$tobacco == 2, 0, dat_drop$tobacco)

#Linear Model for reference
dat.mod <- dat_drop %>% select(dbrwt,stresfip,dmage,ormoth,mrace3,dmeduc,dmar,adequacy,
               dtotord,monpre,nprevist,disllb,birmon,dgestat,csex,dplural,
               anemia,diabetes,herpes,chyper,pre4000,
               preterm,tobacco,cigar,alcohol,drink,wgain)

lm.mod <- lm(dbrwt ~., data = dat.mod)
```

## 1a - Misspecification bias

There are a number of possible sources for misspecification bias in our linear model estimates from PS1a.

One source of misspecification bias would be omitted variables bias. The assumption that random assignment happens conditional on the observables does not protect us against non-random assignment of some unobservable covariate. Perhaps most importantly, however, is the likelihood that there exists some issue of endogeneity between the outcome variable and some of the control. For example, smoking while pregnant may contribute to hypertension, cardiac issues, or other factors that were considered in our initial model as control variables independent of the outcome variable.

A second source of misspecification bias would be in the functional form assumption of linearity. The control variables, including the treatment variable, are all assumed to have linear effects on the outcome variable, which may or may not be the correct functional form of the model, as some effects may in fact be non-linear, or there may exist important interaction effects between controls that were not considered.It could be the case that smoking has some non-linear effect on the birthweight of a baby that we would not capture in the linear model we estimated. We would want to explore nonparametric regression to evaluate the sensitivity to this misspecification.

## 1b - Higher order specifications

We explored using a series estimator of the following functional form:

```r
dat_drop$dmage2 <- dat_drop$dmage^2
dat_drop$cigar2 <- dat_drop$cigar^2
```

```
dat_drop$cigar3 <- dat_drop$cigar^3
dat_drop$cig_dmag <- dat_drop$cigar*dat_drop$dmage

lm.out <- lm(dbrwt ~ stresfip+dmage+ormoth+mrace3+dmeduc+dmar+adequacy+dfage+
               orfath+dfeduc+dtotord+monpre+nprevist+disllb+birmon+dgestat+csex+dplural+
               anemia+diabetes+herpes+chyper+
               preterm+tobacco_p+cigar+alcohol+dmage2+cigar2+cigar3+cig_dmag, data = dat_drop)
summary(lm.out)
```

```
##
## Call:
## lm(formula = dbrwt ~ stresfip + dmage + ormoth + mrace3 + dmeduc +
##     dmar + adequacy + dfage + orfath + dfeduc + dtotord + monpre +
##     nprevist + disllb + birmon + dgestat + csex + dplural + anemia +
##     diabetes + herpes + chyper + preterm + tobacco_p + cigar +
##     alcohol + dmage2 + cigar2 + cigar3 + cig_dmag, data = dat_drop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2845.5  -296.1    -7.8   291.6  2658.1
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.054e+02  8.280e+01 -10.935  < 2e-16 ***
## stresfip     7.137e-01  6.324e-01   1.128 0.259125
## dmage        1.140e+01  2.205e+00   5.169 2.35e-07 ***
## ormoth      -1.982e+01  3.477e+00  -5.701 1.19e-08 ***
## mrace3      -7.880e+01  2.348e+00 -33.565  < 2e-16 ***
## dmeduc       3.787e+00  8.651e-01   4.377 1.20e-05 ***
## dmar        -3.539e+01  4.200e+00  -8.427  < 2e-16 ***
## adequacy     7.465e+00  4.114e+00   1.815 0.069595 .
## dfage       -6.516e-03  3.379e-01  -0.019 0.984616
## orfath      -1.624e+01  3.425e+00  -4.743 2.11e-06 ***
## dfeduc       2.500e+00  8.040e-01   3.110 0.001872 **
## dtotord      1.122e+01  1.214e+00   9.240  < 2e-16 ***
## monpre       1.203e+00  1.413e+00   0.851 0.394650
## nprevist     1.088e+01  5.328e-01  20.419  < 2e-16 ***
## disllb      -1.713e-01  4.851e-03 -35.319  < 2e-16 ***
## birmon      -1.222e+00  4.031e-01  -3.032 0.002426 **
## dgestat      1.115e+02  5.986e-01 186.230  < 2e-16 ***
## csex        -1.400e+02  2.738e+00 -51.154  < 2e-16 ***
## dplural     -5.853e+02  8.127e+00 -72.013  < 2e-16 ***
## anemia      -2.296e+01  1.380e+01  -1.664 0.096122 .
## diabetes    -1.406e+02  8.536e+00 -16.468  < 2e-16 ***
## herpes      -2.518e+01  1.753e+01  -1.436 0.150883
## chyper       1.325e+02  1.573e+01   8.425  < 2e-16 ***
## preterm      2.642e+02  1.171e+01  22.560  < 2e-16 ***
## tobacco_p   -1.028e+02  1.140e+01  -9.011  < 2e-16 ***
## cigar       -1.097e+01  2.283e+00  -4.804 1.56e-06 ***
## alcohol      5.537e+01  1.412e+01   3.920 8.85e-05 ***
## dmage2      -1.786e-01  3.737e-02  -4.780 1.75e-06 ***
## cigar2       4.131e-01  8.815e-02   4.687 2.78e-06 ***
## cigar3      -3.291e-03  9.998e-04  -3.292 0.000996 ***
## cig_dmag    -1.495e-01  4.681e-02  -3.194 0.001402 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 463 on 114579 degrees of freedom
## Multiple R-squared:  0.3743, Adjusted R-squared:  0.3741
## F-statistic:  2284 on 30 and 114579 DF,  p-value: < 2.2e-16
```

The benefits of this approach is that it potentially increases the accuracy of the prediction of treatment effect by removing misspecification bias. The drawbacks of this approach are the potential for overspecification, meaning the new specification is based more on the noise inherent in the data, and less on actual relationship of treatment to outcome. It will be hard to justify that all of the nonparametric decisions make intuitive economic or real world "sense". Finally, as you approach more parameters you could run into the curse of dimensionality when wanting to interpret a causal effect.

```
## add nick's spline code
# Find knots based on equal quantiles of data
nknots <- 3
knots <- attr(bs(dat_drop$dbrwt, df = nknots+3), "knots")
cspline.mod <- lm(dbrwt ~ bs(stresfip,knots = knots)+bs(dmage,knots = knots)+bs(ormoth,knots = knots)+bs
```

Here are graphs showing the predicted vs. observed birthweights using the simple linear model and with cubic splines.

## 1c - Using LASSO

In our application of lasso, we apply the method proposed by belloni, chernozhukov, and hansen. First we apply lasso of the treatment on the covariates. Then, we apply it on the outcome variable and the covariates and keep the set of covariates that lasso selects in either 1 or 2.

```
x_lasso <- dat_drop %>%
  select(stresfip,dmage,ormoth,mrace3,dmeduc,dmar,adequacy,dfage,
             orfath,dfeduc,dtotord,monpre,nprevist,disllb,birmon,dgestat,csex,dplural,
             anemia,diabetes,herpes,chyper,
             preterm,cigar,alcohol,dmage2,cigar2,cigar3,cig_dmag) %>% as.matrix()

y_lasso <- dat_drop %>% select(dbrwt) %>% as.matrix()

d_lasso <- dat_drop %>% select(tobacco_p) %>% as.matrix()

fit <- cv.glmnet(x_lasso,d_lasso)
coef(fit, s = "lambda.1se")
```

```
## 30 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  9.250479e-02
## stresfip     .
## dmage       -8.490930e-05
## ormoth       .
## mrace3       .
## dmeduc      -1.618641e-03
## dmar         2.474879e-02
## adequacy     .
## dfage        .
## orfath       .
## dfeduc      -7.119644e-04
## dtotord      .           
## monpre       .           
## nprevist     .           
## disllb       .           
## birmon       .           
## dgestat      .           
## csex         .           
## dplural      .           
## anemia       .           
## diabetes     .           
## herpes       .           
## chyper       .           
## preterm      .           
## cigar        1.242140e-01
## alcohol     -3.149641e-02
## dmage2       .           
## cigar2      -3.954556e-03
## cigar3       3.001411e-05
## cig_dmag     .           
```

```
fit2 <- cv.glmnet(x_lasso,y_lasso)
coef(fit2, s = "lambda.1se")
```

```
## 30 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept) -5.569281e+02
## stresfip     .           
## dmage        1.231709e+00
## ormoth      -1.121544e+01
## mrace3      -6.865584e+01
## dmeduc       3.629035e+00
## dmar        -4.959588e+01
## adequacy     .           
## dfage        .           
## orfath      -8.906802e+00
## dfeduc       1.877541e+00
## dtotord      6.283663e+00
## monpre       .           
## nprevist     8.695619e+00
## disllb      -1.624610e-01
## birmon       .           
## dgestat      1.106621e+02
## csex        -1.273088e+02
## dplural     -5.463108e+02
## anemia       .           
## diabetes    -1.020847e+02
## herpes       .           
## chyper       5.713629e+01
## preterm      2.120339e+02
## cigar       -1.294992e+01
## alcohol      1.350272e+01
## dmage2       .           
## cigar2       .           
## cigar3       1.134211e-03
## cig_dmag    -2.291158e-03
```

Based on the estimates of 0 in both, we drop stresfip, adequacy, dfage, monpre, birmon, anemia, herpes, dfage2, dmage2, cigar2.

```
lm.out <- lm(dbrwt ~ dmage+ormoth+mrace3+dmeduc+dmar+
               orfath+dfeduc+dtotord+nprevist+disllb+dgestat+csex+dplural+diabetes+chyper+
               preterm+tobacco_p+cigar+alcohol+cigar3+cig_dmag, data = dat_drop)
summary(lm.out)
```

```
##
## Call:
## lm(formula = dbrwt ~ dmage + ormoth + mrace3 + dmeduc + dmar +
##     orfath + dfeduc + dtotord + nprevist + disllb + dgestat +
##     csex + dplural + diabetes + chyper + preterm + tobacco_p +
##     cigar + alcohol + cigar3 + cig_dmag, data = dat_drop)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2840.00  -296.40    -7.47   291.80  2652.15
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.157e+02  5.754e+01 -14.176  < 2e-16 ***
## dmage        1.026e+00  3.377e-01   3.038 0.002378 **
## ormoth      -1.979e+01  3.476e+00  -5.694 1.24e-08 ***
## mrace3      -7.834e+01  2.342e+00 -33.457  < 2e-16 ***
## dmeduc       4.273e+00  8.532e-01   5.008 5.50e-07 ***
## dmar        -3.923e+01  4.041e+00  -9.708  < 2e-16 ***
## orfath      -1.629e+01  3.425e+00  -4.757 1.97e-06 ***
## dfeduc       2.363e+00  8.009e-01   2.951 0.003173 **
## dtotord      1.147e+01  1.209e+00   9.487  < 2e-16 ***
## nprevist     1.001e+01  4.082e-01  24.523  < 2e-16 ***
## disllb      -1.741e-01  4.816e-03 -36.155  < 2e-16 ***
## dgestat      1.118e+02  5.916e-01 188.995  < 2e-16 ***
## csex        -1.401e+02  2.738e+00 -51.167  < 2e-16 ***
## dplural     -5.839e+02  8.121e+00 -71.902  < 2e-16 ***
## diabetes    -1.416e+02  8.532e+00 -16.592  < 2e-16 ***
## chyper       1.323e+02  1.572e+01   8.420  < 2e-16 ***
## preterm      2.632e+02  1.171e+01  22.475  < 2e-16 ***
## tobacco_p   -1.426e+02  7.477e+00 -19.071  < 2e-16 ***
## cigar       -2.804e+00  1.399e+00  -2.004 0.045107 *
## alcohol      5.157e+01  1.411e+01   3.655 0.000257 ***
## cigar3       1.108e-03  3.361e-04   3.296 0.000980 ***
## cig_dmag    -1.281e-01  4.669e-02  -2.744 0.006078 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 463.1 on 114588 degrees of freedom
## Multiple R-squared:  0.3739, Adjusted R-squared:  0.3738
## F-statistic:  3259 on 21 and 114588 DF,  p-value: < 2.2e-16
```

Some of these terms I would have thought would have mattered such as adequacy of care and anemia.

# 2 - Propensity score description

The propensity score helps solve the issue that it is hard to condition on the covariates if they are highly dimensional. However, we want to do such conditioning in order to compare between treated and control units. The propensity score helps us proxy for the probability of entering treatment and therefore after conditioning on the propensity score, the units are as good as randomly assigned.

## 2a - Create propensity score

```
Pcontrols = 'dmage+ ormoth+ mrace3+ dmeduc+ dmar+ dfage+
orfath+ dfeduc + disllb + isllb10 + anemia + diabetes+ herpes+ phyper'

# estimate propensity score
all_p <- glm( as.formula(paste( 'tobacco_p ~ ', Pcontrols)),
                              dat_drop, family = binomial(logit))

summary(all_p)
```

```
##
## Call:
## glm(formula = as.formula(paste("tobacco_p ~ ", Pcontrols)), family = binomial(logit),
##     data = dat_drop)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4737  -0.6044  -0.4261  -0.2816   3.2548
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.794e-01  3.161e-01   0.884  0.37682
## dmage       -1.787e-02  2.570e-03  -6.953 3.57e-12 ***
## ormoth      -4.093e-01  2.724e-02 -15.027  < 2e-16 ***
## mrace3      -4.676e-01  1.389e-02 -33.667  < 2e-16 ***
## dmeduc      -1.812e-01  5.485e-03 -33.030  < 2e-16 ***
## dmar         1.300e+00  2.178e-02  59.697  < 2e-16 ***
## dfage        2.602e-02  1.974e-03  13.186  < 2e-16 ***
## orfath      -2.093e-01  2.392e-02  -8.752  < 2e-16 ***
## dfeduc      -1.229e-01  5.172e-03 -23.758  < 2e-16 ***
## disllb      -1.406e-04  4.449e-05  -3.161  0.00157 **
## isllb10      6.460e-02  5.063e-03  12.758  < 2e-16 ***
## anemia      -2.895e-02  7.769e-02  -0.373  0.70939
## diabetes    -5.496e-02  5.320e-02  -1.033  0.30159
## herpes      -1.569e-01  1.072e-01  -1.464  0.14325
## phyper       4.271e-01  5.847e-02   7.305 2.77e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 100543  on 114609  degrees of freedom
## Residual deviance:  88060  on 114595  degrees of freedom
## AIC: 88090
```

```
##
## Number of Fisher Scoring iterations: 5
```

```
dat_drop$p.hat_all <- predict(glm( as.formula(paste( 'tobacco_p ~ ', Pcontrols)),
                              dat_drop, family = binomial(logit)),
                         type = "response") # this is important!

Pcontrols2 = 'dmage+ ormoth+ mrace3+ dmeduc+ dmar+
dfage+ orfath+ dfeduc + disllb + isllb10 + phyper'

dat_drop$p.hat_2 <- predict(glm( as.formula(paste( 'tobacco_p ~ ', Pcontrols2)),
                              dat_drop, family = binomial(logit)),
                         type = "response") # this is important!
```

Anemia, diabetes, and herpes were all insignificant in the calculation of the propensity score.

```
dat_drop$diff <- dat_drop$p.hat_all - dat_drop$p.hat_2

summary(dat_drop$diff)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.0018225 -0.0004369 -0.0002613  0.0000000 -0.0001460  0.0517418
```

The propensity scores were very comparable with at most a 5% difference and an average difference of 0%.

Further, we can see that excluding the non-significant covariates in the first propensity score estimation has a little overall effect on the propensity scores of being treated (R2 = 1).

This implies that we have the right covariates that are predetermined in the problem from the original set of data but does not necessarily protect us against ommitted variables.

## 2b - propensity score estimation

```
#Controlling for propensity scores
reg4 <- lm(dbrwt ~ tobacco_p + p.hat_2, data = dat_drop)
summary(reg4)
```

```
##
## Call:
## lm(formula = dbrwt ~ tobacco_p + p.hat_2, data = dat_drop)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3191.5  -311.4    26.9   359.9  2744.6
##
## Coefficients:
##             Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 3422.731      2.747 1245.870  < 2e-16 ***
## tobacco_p   -231.375      4.952  -46.721  < 2e-16 ***
## p.hat_2      -78.835     14.319   -5.506 3.68e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 578.4 on 114607 degrees of freedom
## Multiple R-squared:  0.02288,    Adjusted R-squared:  0.02287
## F-statistic:  1342 on 2 and 114607 DF,  p-value: < 2.2e-16
```

Using this approach, we estimate an average treatment effect of -231 grams as a result of smoking while pregnant. This treatment effect is consistent under unconfoundedness (i.e. random assignment of treatment conditional on the covariates) and the assumption of a constant (i.e., homogenous) treatment effect. Furthermore, it assumes that there is sufficient overlap of treatment and controls in the covariate space.

## 2c - propensity score reweighting

```
ATE <- (sum(dat_drop$tobacco_p*dat_drop$dbrwt/dat_drop$p.hat_2)/
          sum(dat_drop$tobacco_p/dat_drop$p.hat_2)) -
  (sum((1 - dat_drop$tobacco_p)*dat_drop$dbrwt/(1 - dat_drop$p.hat_2))/
             sum((1 - dat_drop$tobacco_p) /(1 - dat_drop$p.hat_2)) )

print(ATE)
```

```
## [1] -232.2479
```

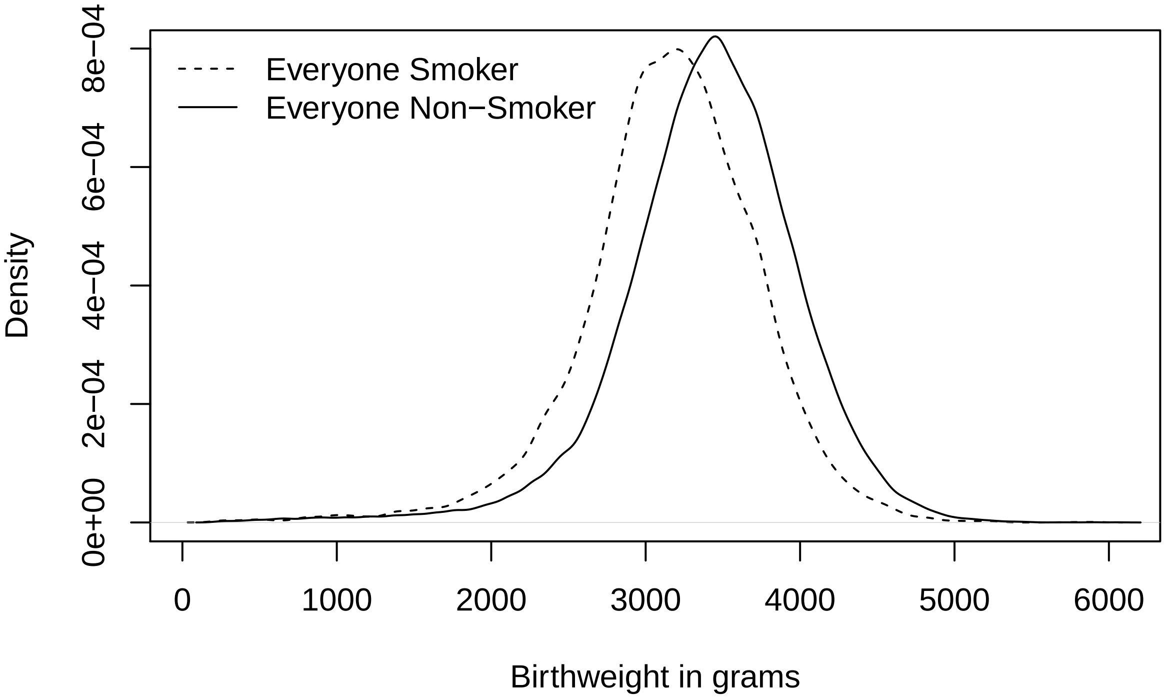
```
TOT <- ( sum(dat_drop$tobacco_p*dat_drop$dbrwt)/sum(dat_drop$tobacco_p)) -
  sum(dat_drop$p.hat_2 *(1 - dat_drop$tobacco_p)*dat_drop$dbrwt / (1 - dat_drop$p.hat_2) ) /
  sum(dat_drop$p.hat_2 * (1 - dat_drop$tobacco_p) /
        (1 - dat_drop$p.hat_2))

print(TOT)
```

```
## [1] -229.6262
```

## 2d kernel density estimator

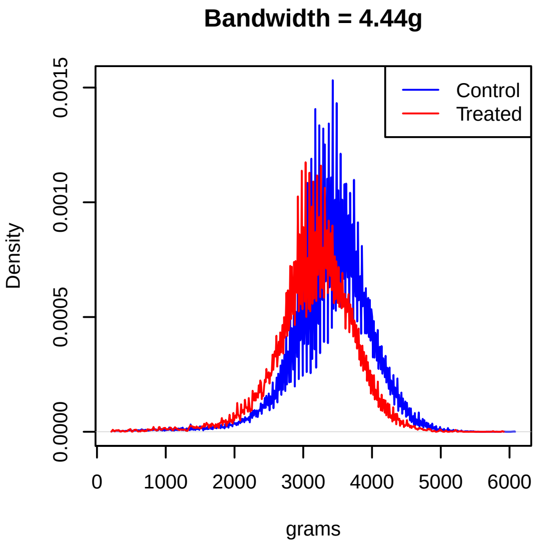
Plotting the densities here for the full sample

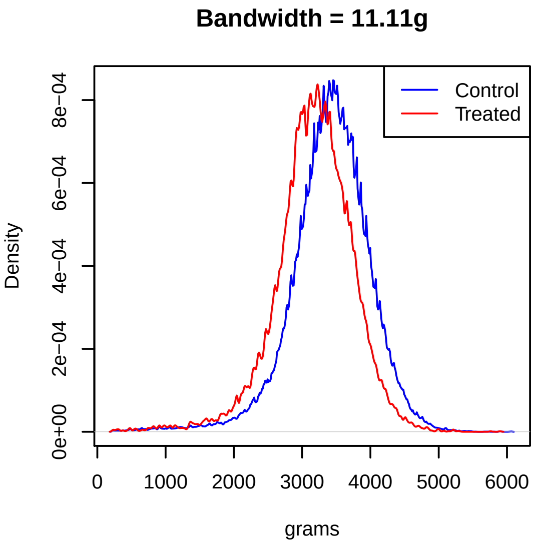
```
treatment <- density(dat_drop$dbrwt[dat_drop$tobacco_p==1], kernel = "gaussian",
                     weights = dat_drop$p.hat_2[dat_drop$tobacco_p==1]/
                       sum(dat_drop$p.hat_2[dat_drop$tobacco_p==1]))

control <- density(dat_drop$dbrwt[dat_drop$tobacco_p==0], kernel = "gaussian",
                   weights = (1 - dat_drop$p.hat_2[dat_drop$tobacco_p==0])/
                     sum((1 - dat_drop$p.hat_2[dat_drop$tobacco_p==0])))

plot(treatment,
lty = 'dashed',
main = "",
xlab = "Birthweight in grams",
ylab = "Density")
lines(control)
legend(x = 'topleft', legend = c('Everyone Smoker', 'Everyone Non-Smoker'),
lty = c('dashed', 'solid'), bty = "n")
```

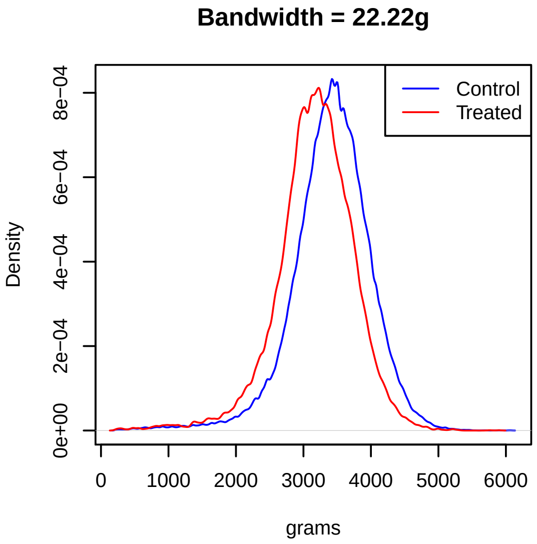


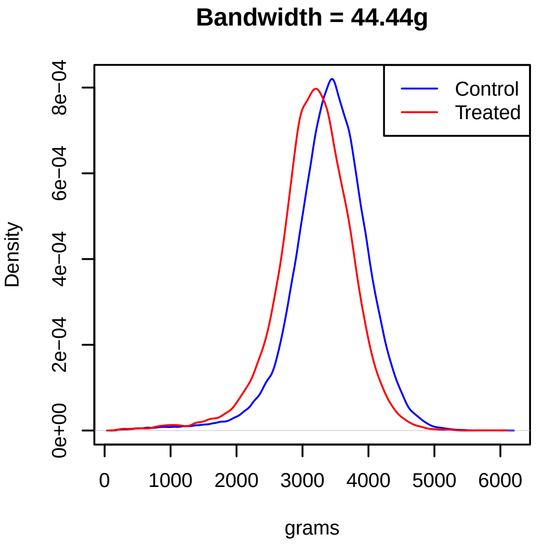
## 2e - kernel density bandwidth adjustments

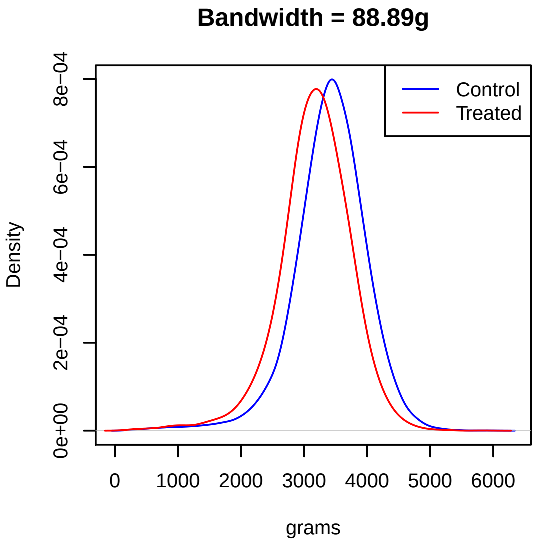
```
n <- nrow(dat_drop)
y <- dat_drop$dbrwt
D <- dat_drop$tobacco_p
#Number of Control Obs
nC <- length(D[D==0])
nT <- length(D[D==1])
par(mfrow = c(3,2))
for(adj in c(0.1,0.25,0.5,1,2,10)){
# Define Bandwidth
h0 <- density(y)$bw
bw_adjust <- adj
h <- h0*bw_adjust

# Control & Treatment Densities
Cdens <- density(y[D==0], adjust = bw_adjust)
Tdens <- density(y[D==1], adjust = bw_adjust)

#Plot densities
plot(Cdens, col = "blue", main = paste0("Bandwidth = ",round(h,2),"g"), xlab = "grams")
lines(Tdens, col = "red")
legend('topright', legend = c("Control", "Treated"), col = c("blue", "red"), lty = 1)
}
```

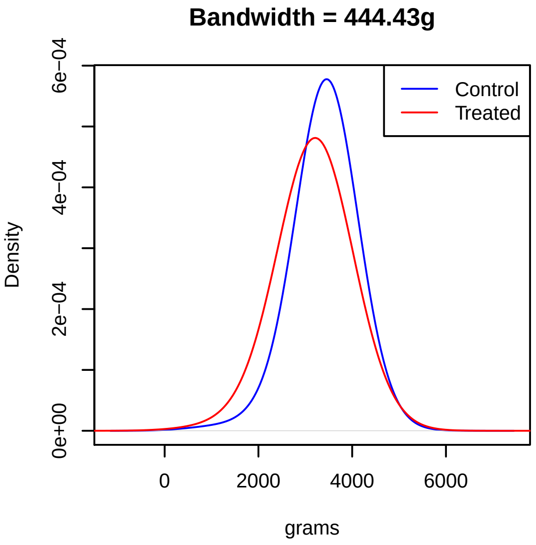

Bandwidth = 4.44g
Density
0.0000
0.0005
0.0010
0.0015
0
1000
2000
3000
4000
5000
6000
grams
Control
Treated


Bandwidth = 11.11g
Density
0e+00
2e−04
4e−04
6e−04
8e−04
0
1000
2000
3000
4000
5000
6000
grams
Control
Treated


Bandwidth = 22.22g
Density
0e+00
2e−04
4e−04
6e−04
8e−04
0
1000
2000
3000
4000
5000
6000
grams
Control
Treated


Bandwidth = 44.44g
Density
0e+00
2e−04
4e−04
6e−04
8e−04
0
1000
2000
3000
4000
5000
6000
grams
Control
Treated

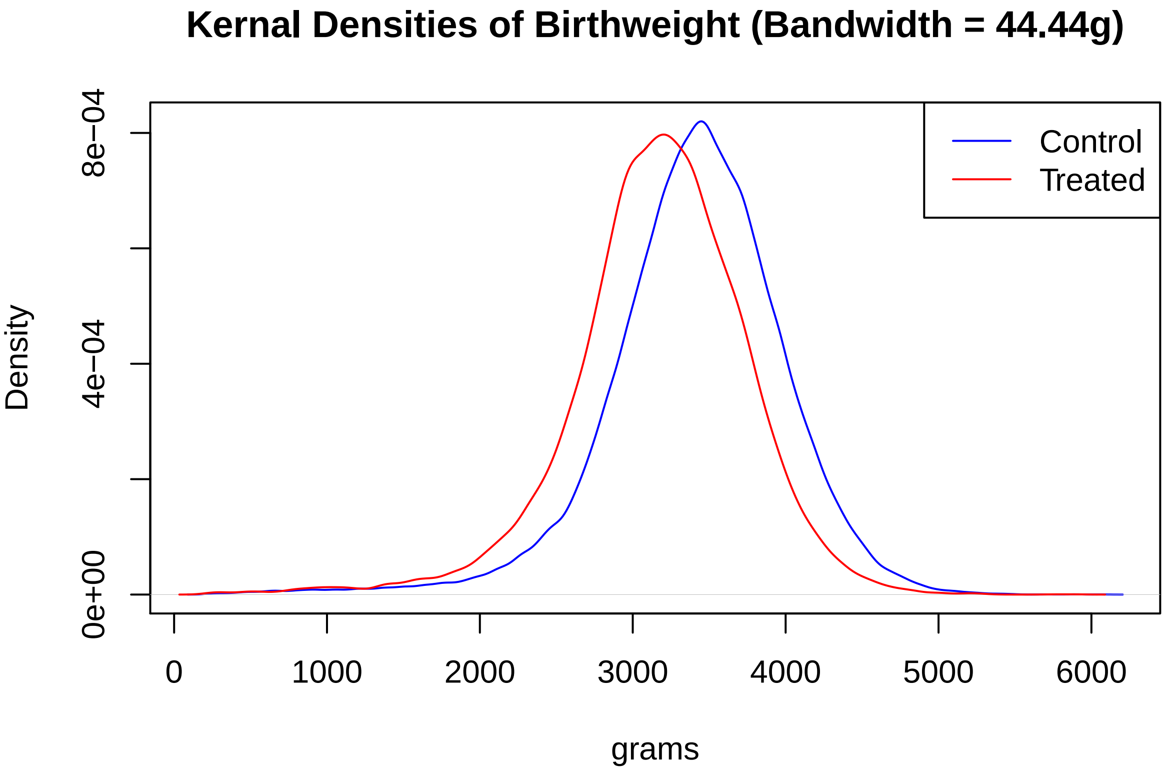

Bandwidth = 88.89g
Density
0e+00
2e−04
4e−04
6e−04
8e−04
0
1000
2000
3000
4000
5000
6000
grams
Control
Treated


Bandwidth = 444.43g
Density
0e+00
2e−04
4e−04
6e−04
0
2000
4000
6000
grams
Control
Treated

In the above figure, we could see that with smaller bandwidths, the kernal densities get much more jagged, or variable. However, as we increase the bandwidth and the curves become smoother, it's also evident that there is more bias in the estimates because the magnitudes of the curves begin to change relative to one another (i.e. the peak of the control curve starts to decline as the bandwidth is increased).

The default bandwidth of ~44 grams seems to produce a fairly smooth kernal density, and was selected as the preferred bandwidth.

```r
h0 <- density(y)$bw
bw_adjust <- 1
h <- h0*bw_adjust

# Control & Treatment Densities
Cdens <- density(y[D==0], adjust = bw_adjust)
Tdens <- density(y[D==1], adjust = bw_adjust)

#Plot densities
plot(Cdens, col = "blue", main = paste0("Kernal Densities of Birthweight (Bandwidth = ",round(h,2),"g")"
lines(Tdens, col = "red")
legend('topright', legend = c("Control", "Treated"), col = c("blue", "red"), lty = 1)
```



Solving for kernel estimator at birthweight = 3000 grams.

```r
# Uniform kernal function
k.uni <- function(u){
  return(ifelse(abs(u) <= 1,0.5,0))
}

# Triangular kernal function
k.tri <- function(u){
  return(ifelse(abs(u) < 1,1-abs(u),0))
}

# Epanechnikov kernal function
k.epan <- function(u){
  return(ifelse(abs(u) < 1,0.75*(1-u^2),0))
}

# Gaussian kernal function
k.gauss <- function(u){
  return((1/sqrt(2*pi))*exp(-0.5*u^2))
}

# Kernal Density Estimator for Control Group
k.dens.C <- function(ystar, k.func){
  ks <- rep(NA,nC)
  for(i in 1:n){
    u <- (ystar - y[i])/h
    ks[i] <- k.func(u)*(1-D[i])/(1-dat_drop$p.hat_2[i])
  }
  return((1/(nC*h))*(sum(ks)))
}

# Kernal Density Estimator for Treatment Group
k.dens.T <- function(ystar, k.func){
  ks <- rep(NA,nT)
  for(i in 1:n){
    u <- (ystar - y[i])/h
    ks[i] <- k.func(u)*(D[i]/dat_drop$p.hat_2[i])
  }
  return((1/(nT*h))*(sum(ks)))
}
```

Kernal Estimates at 3000g

```
h <- density(y)$bw

est.uni.C <- k.dens.C(3000,k.func = k.uni) %>% round(6)
est.tri.C <- k.dens.C(3000,k.func = k.tri) %>% round(6)
est.epan.C <- k.dens.C(3000,k.func = k.epan) %>% round(6)
est.gauss.C <- k.dens.C(3000,k.func = k.gauss) %>% round(6)

est.uni.T <- k.dens.T(3000,k.func = k.uni) %>% round(6)
est.tri.T <- k.dens.T(3000,k.func = k.tri) %>% round(6)
est.epan.T <- k.dens.T(3000,k.func = k.epan) %>% round(6)
est.gauss.T <- k.dens.T(3000,k.func = k.gauss) %>% round(6)

est.df <- data.frame(Kernal = c("Uniform","Triangular","Epanechnikov","Gaussian"), NonSmoker = c(est.un
kable(est.df, caption = "Kernal Estimates for Birthweight of 3000g")
```

Table 1: Kernal Estimates for Birthweight of 3000g

| Kernal | NonSmoker | Smoker |
|---|---:|---:|
| Uniform | 0.000584 | 0.004304 |
| Triangular | 0.000614 | 0.004835 |

| Kernal | NonSmoker | Smoker |
|---|---|---|
| Epanechnikov | 0.000615 | 0.004735 |
| Gaussian | 0.000602 | 0.004486 |

We can see that as expected, the point kernal estimates for the smoker (treated) at 3000g are higher relative to non-smokers (control), which is inline with what the kernal density functions demonstrate above. Therefore, this can be interpreted, that on average, being a smoker means your overall likelihood of having a child of exactly 3000g is higher than being a non-smoker. 3000g, while not considered especially low, is below average for both the control and treated birthweight distributions seen above.

## 2f - benefits and drawbacks of propensity method

The benefits of the propensity weighting approach in part c is that it allows for conditioning solely on the likelihood of selecting into treatment, rather than on all predetermined control variables, which rectifies issues that arise when matching across many variables (Curse of Dimensionality).

Furthermore, we know that propensity scores are not balanced across treated and control groups. The weighting scheme ensures that each observation is equally represented (in expectation) in the treated and control groups. According to to Hirano, Imbens,and Ridder, the weighting estimator we use is efficient. However, very low and very high propensity scores may weight observations to 0 or infinity and bias the results.

## 2g - Present and discuss results

i. Must hold. The inverse weighting of propensity scores seems more reasonable if the treatment effect heterogeneity is linear in the propensity scores. If it is non-linear we might need to use a differenet weighting mechanism.

ii. Need not hold. If it is non-linear we might need to use a differenet weighting mechanism. Our weighhting mechanism in 2c might not be accurate.

iii. Need not hold. Without conditioning for propensity scores we saw a systemic difference between smokers and non smokers who selected into treatment. The systematic dfference was accounted for when we calclated the propensity scores.

iv. Must hold. On taking propensity scores, we account for the exogenous variation and the decision to smoke is as good as randomly assigned. If conditional on the exogenous variables the decision to smoke is still not randomly assigned we will get biased estimnates for ATE and TOT.

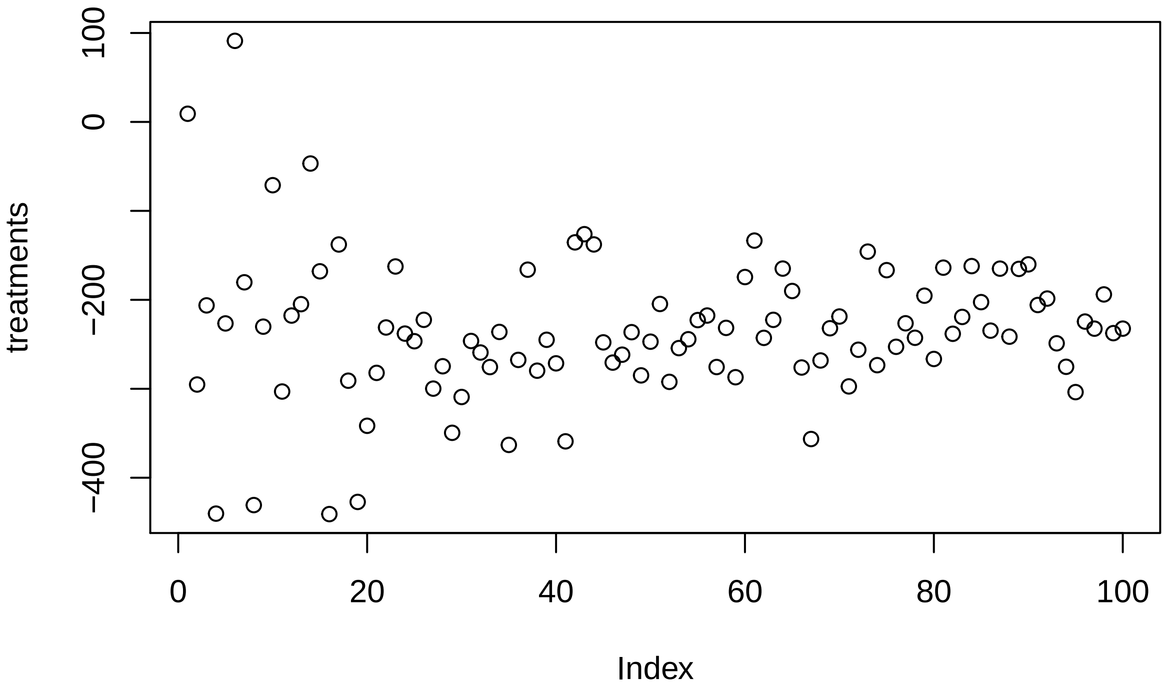
## 3 - Blocking non-parametric approach

```r
# Create 100 equally sized bins based on propensity scores

dat_drop$bin <- cut(dat_drop$p.hat_2, quantile(dat_drop$p.hat_2, seq(0,1,0.01)),
                    include.lowest = TRUE, labels = FALSE)

# Find difference between birth weights of non-smokers and smokers in each bin, and assign weights to t

treatment_effect = 0
```

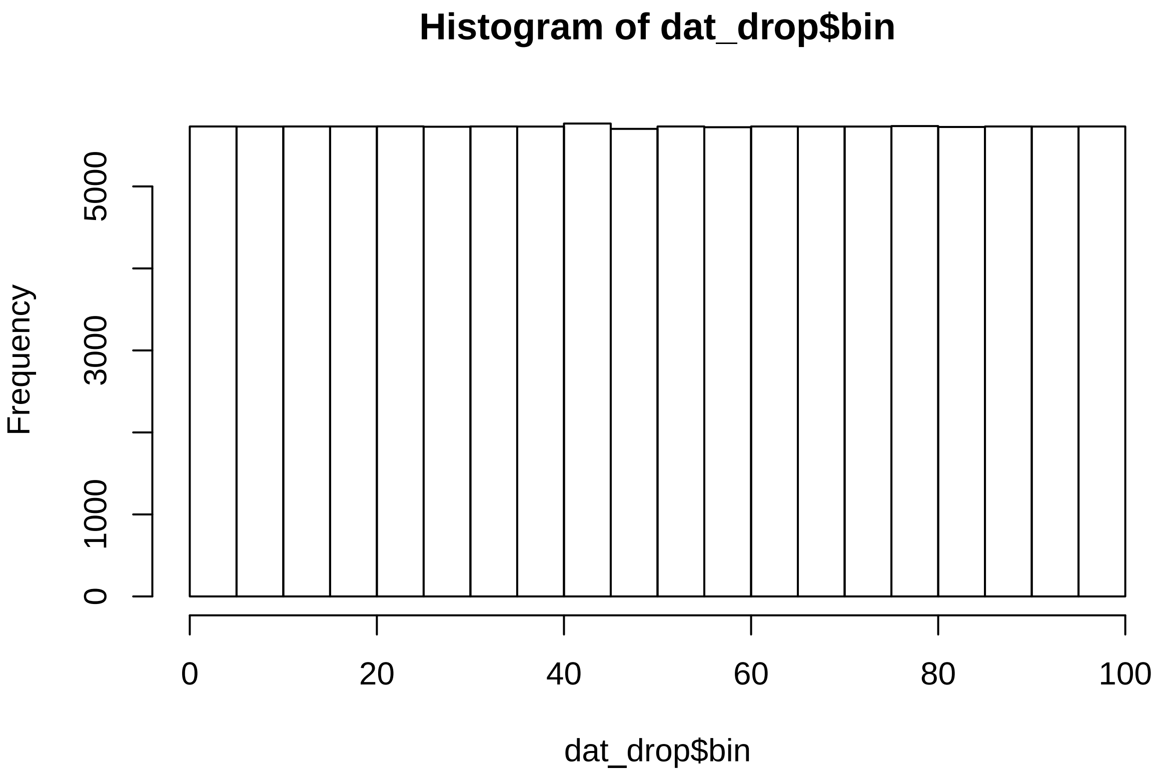
```
treatments <- c()
for(i in 1:100)
{

mean_diff = mean(dat_drop$dbrwt[dat_drop$tobacco_p==1 & dat_drop$bin==i]) -
  mean(dat_drop$dbrwt[dat_drop$tobacco_p==0 & dat_drop$bin==i])
treatments <- c(treatments, mean_diff)
#print(i)
#print(mean_diff)

N <- nrow(dat_drop)
N1 <- nrow(dat_drop[dat_drop$tobacco_p==1 & dat_drop$bin == i,])
N0 <- nrow(dat_drop[dat_drop$tobacco_p==0 & dat_drop$bin == i,])

weighted_mean <- mean_diff*((N1+N0)/N)
#print(weighted_mean)

treatment_effect = weighted_mean + treatment_effect
}

print(treatment_effect)
```

```
## [1] -234.3858
```

```
plot(treatments)
```



```
hist(dat_drop$bin)
```



The result from this weighted mean difference method shows that on average, babies of smokers have lower birth weights by 234 grams as compared to babies of non-smokers.This is almost equal to the ATE of -232 grams found in the previous question where we did not block as per propensity scores. These equal results show that there are no systematic differences between smokers and non-smokers based on the other covariates included in the study.

A quick check of the number of data points within the bins shows that we have balanced blocking as well since the data is uniformly distributed across the bins. Furthermore, a plot of all of the mean treatment effects in each bin shows a pretty stable line for each bin.

## 4 - low birth weights

```
# Create a low birth weight indicator variable
dat_drop$low <- 0
dat_drop$low[dat_drop$dbrwt < 2500] <- 1

# Find difference between proportion of low birth weights between non-smokers and smokers in each bin,

treatment_effect = 0
treatments <- c()

for(i in 1:100)
{

mean_diff = mean(dat_drop$low[dat_drop$tobacco_p==1 & dat_drop$bin==i]) -
  mean(dat_drop$low[dat_drop$tobacco_p==0 & dat_drop$bin==i])
treatments <- c(treatments, mean_diff)
#print(i)
```

```r
#print(mean_diff)

N <- nrow(dat_drop)
N1 <- nrow(dat_drop[dat_drop$tobacco_p==1 & dat_drop$bin == i,])
N0 <- nrow(dat_drop[dat_drop$tobacco_p==0 & dat_drop$bin == i,])

weighted_mean <- mean_diff*((N1+N0)/N)
#print(weighted_mean)

treatment_effect = weighted_mean + treatment_effect

}

print(treatment_effect)
```

```
## [1] 0.04452343
```

```r
# Difference in proportion of low birthweight babies between smokers and non-smokers
low_diff <- (nrow(dat_drop[dat_drop$low ==1 & dat_drop$tobacco_p==1,])/
               nrow(dat_drop[dat_drop$tobacco_p==1,]) - nrow(dat_drop[dat_drop$low ==1 &

low_diff
```

```
## [1] 0.04514949
```

The result shows that smokers have a 4.45 pc higher proportion of low birthweight babies than do non-smokers. This is almost equal to the absolute difference of 4.52 pc in the proportion of low birth weight babies between smokers and non-smokers. Hence, we can say that there are no systematic differences in birth weights between the bins/blocks created, indicating that we have a uniform distribution of data across the bins.

# 5 - Summarize results

Drawing on our results from the previous assignment where we found that birth weights of smokers were on average 231.8 gms lower than the birth weights of non-smokers, we now allow the relationship between birth weight and the explanatory variables to assume a non-linear functional form. In order to ensure that the selection into treatment does not differ in some meaningful way from the units that do not select into treatment, we condition the regression on propensity scores. The propensity scores are calculated using all the pre-determined covariates and quadratic terms for age of mother and father, based on observation of the functional form and results of the LASSO regression estimate. The results show that the average treatment effect of smoking increased to 234 gms on controlling for the propensity scores, showing that some of the difference in birth weights between smokers and non-smokers was due to systematic differences in how smokers selected into the treatment. In order to balance the propensity scores across treatment and control groups, we use a weighting method which ensures that smokers and non-smokers are equally represented across both groups. The results for this regression show that the ATE is 232 gms which is the same as our previous result from the conditional regression. However, the treatment on treated effect is found to be 229 gms showing that the difference in birth weights is lower for those who actually selected into treatment and control versus those who had the highest likelihood of being in either group. We also use a blocking method based on propensity scores to confirm that there are no block-wise differences between smokers and non-smokers with similar propensity scores. The results indicate an ATE of 234.3 gms which is almost equal to the ATE found previously, showing that there are no systematic differences between smokers and non-smokers based on the other covariates included in the study. Using the blocking method, we also find that smokers have a 4.45 pc higher proportion of low birthweight babies than do non-smokers.