ARE 213 Problem Set 1B

Becky Cardinali, Yuen Ho, Sara Johns, and Jacob Lefler

Due 10/12/2020

Question 1

In Problem Set 1a, you used linear regression to relate infant health outcomes and maternal smoking during pregnancy. Please answer the following questions.

(a) Under the assumption of random assignment conditional on the observables, what are the sources of misspecification bias in the estimates generated by the linear model estimated in Problem Set 1a?

Question 2

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?

We know that if we condition on observables, we will get a consistent estimate of the ATE under the assumption. However, if the observables are high dimensional, it might be difficult to find a comparison unit with the same values of the observables. From lecture, we know that it is sufficient instead to condition on the propensity score. Using the propensity score allows us to compare treated and control units with the same probability of being treated. The propensity score does not require that all values of the observables be the same and so therefore avoids problems of multidimensionality.

Try a few ways to estimate the effects of maternal smoking on birthweight:

2a

a) First create the propensity score. For our purposes let's use a logit specification. First specify the logit using all of the "predetermined" covariates (don't include interactions). Next, include only those "predetermined" covariates that enter significantly in the first logit specification. How comparable are the propensity scores? If they are similar does this imply that we have the "correct" set of covariates in the logit specification used for our propensity score?

```r
# get prop score using all predetermined variables
prop_all <- glm(tobacco ~ factor(stresfip) + dmage + factor(mrace3) + dmeduc +
                  dtotord + disllb + dfage + factor(birmon) + factor(orfath) +
                  factor(dmar) + dfeduc + dplural + factor(pre4000) + factor(preterm),
                family=binomial(link='logit'), data = mom_dt)
mom_dt[, prop_score_all := fitted(prop_all)]

# take a look at the output to see which are significant - omitting because takes up a lot of space
# summary(prop_all)
# only need to take out state of residence. birth month has a few months that are significant so will keep

# get prop score using significant variables from previous logit
```

```
prop_sig <- glm(tobacco ~ dmage + factor(mrace3) + dmeduc +
                  dtotord + disllb + dfage + factor(birmon) + factor(orfath) +
                  factor(dmar) + dfeduc + dplural + factor(pre4000) + factor(preterm),
                family=binomial(link='logit'), data = mom_dt)
mom_dt[, prop_score_sig := fitted(prop_sig)]

# how different are the prop scores?
prop_score_diff <- mom_dt$prop_score_all - mom_dt$prop_score_sig
summary(prop_score_diff)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## -0.2889930  0.0003105  0.0004940  0.0000000  0.0006952  0.2873649
```

```
# a few outliers but not very different
```

In the first logit, we include state of residence, mother's age, mother's race, mother's education, birth order, interval since last birth, father's age, father's Hispanic origin, father's education, birth month, plurality, previous heavy child, and previous preterm. This is a combination of the "predetermined" variables we selected and discussed in the last problem set and the ones discussed in the problem set solutions. We find that all coefficients are significant, except for state of residence. We remove that from the next regression and the propensity scores, which are the fitted values of the logit, do not change very much. This does not guarantee we are estimating the logit with the correct set of covariates. There could be important covariates that we do not have in this dataset that we would want to use in the logit specification. The fact that our propensity scores do not change much between the two specifications just gives us confidence that state of residence was not an important regressor to include. Moving forward, we will use the propensity scores from the second specification.

2b

Control directly for the estimated propensity scores using a regression analysis, and estimate an average treatment effect. State clearly the assumptions under which your estimate is correct.

```
# run regression with prop_score
prop_reg <- lm(dbrwt ~ tobacco + prop_score_sig, data = mom_dt)
```

Controlling directly for the propensity score, we get the ATE is -223.23 and is statistically significant at the 99.9% level. The assumption under which this estimate is consistent is unconfoundedness and homogeneous treatment effects. If we instead believed that there were heterogeneous treatment effects (that varied with X), then we would instead want to also interact the propensity score with the treatment status (though this is not all that helpful).

2c

```
# create propensity weights
mom_dt[,prop_weights := ifelse(tobacco == 1,
                               1/prop_score_sig,
                               1/(1 - prop_score_sig))]
# normalize the weights
mom_dt[,norm_prop_weights := ifelse(tobacco == 1,
                                    prop_weights/sum(mom_dt[tobacco == 1, prop_weights]),
                                    prop_weights/sum(mom_dt[tobacco == 0, prop_weights]))]
# estimate ATE
tau_ipw <- sum((mom_dt$tobacco*mom_dt$dbrwt)*(mom_dt$norm_prop_weights) -
                 ((1 - mom_dt$tobacco)*mom_dt$dbrwt)*(mom_dt$norm_prop_weights))

# estimate TOT - use formula from section
```

```
tot_y1 <- sum(mom_dt$tobacco*mom_dt$dbrwt) / sum(mom_dt$tobacco)
tot_y0 <- sum(mom_dt$prop_score_sig * (1 - mom_dt$tobacco) * mom_dt$dbrwt /
        (1 - mom_dt$prop_score_sig)) /
  sum(mom_dt$prop_score_sig * (1 - mom_dt$tobacco) /
        (1 - mom_dt$prop_score_sig))

tau_tot <- tot_y1 - tot_y0
```

We have from lecture that the ATE using inverse propensity weighting is

$$\hat{\tau}_{ATE} = (\sum^{N} \frac{D_i Y_i}{\hat{p}(X_i)} / \sum^{N} \frac{D_i}{\hat{p}(X_i)}) - (\sum^{N} \frac{(1-D_i)Y_i}{1-\hat{p}(X_i)} / \sum^{N} \frac{1-D_i}{1-\hat{p}(X_i)})$$

and using this formula, we get that the ATE is -225.29. We have from section that the TOT using inverse propensity weighting is

$$\hat{\tau}_{TOT} = (\sum^{N} D_i Y_i / \sum^{N} D_i - (\sum^{N} \frac{\hat{p}(X_i)(1-D_i)Y_i}{1-\hat{p}(X_i)} / \sum^{N} \frac{\hat{p}(X_i)(1-D_i)}{1-\hat{p}(X_i)})$$

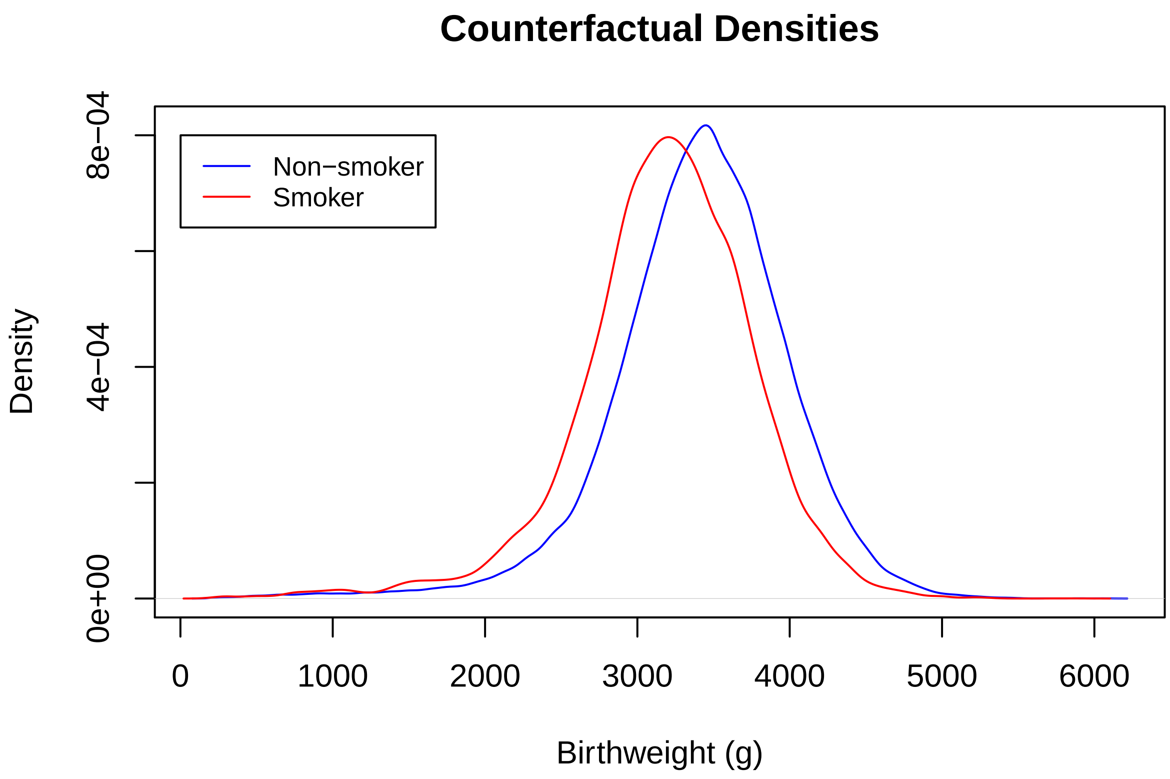
and using this formula, we get that the TOT is -224.4.

2d

Estimate the counterfactual densities relevant for the above part with a kernel density estimator. That is, estimate the density of birthweight (or log birthweight) if everyone smoked and again if no one smoked. Hint: Consider directly applying the Hirano, Imbens, and Ridder propensity score reweighting scheme in the context of estimating the densities of the treated and control groups (rather than the means of the treated and control groups). Stata has very useful preprogrammed commands. In addition to using the preprogrammed Stata command to compute/graph the kernel density over the entire range of birthweight, please also calculate by hand the kernel estimator at birthweight equals 3,000 grams (and provide the code you wrote that shows the calculation of the kernel estimator at this single point). Play around with a bandwidth starting with half the default Stata bandwidth. Choose the same bandwidth for all the pictures, and produce a (beautiful, production quality) figure depicting both densities.

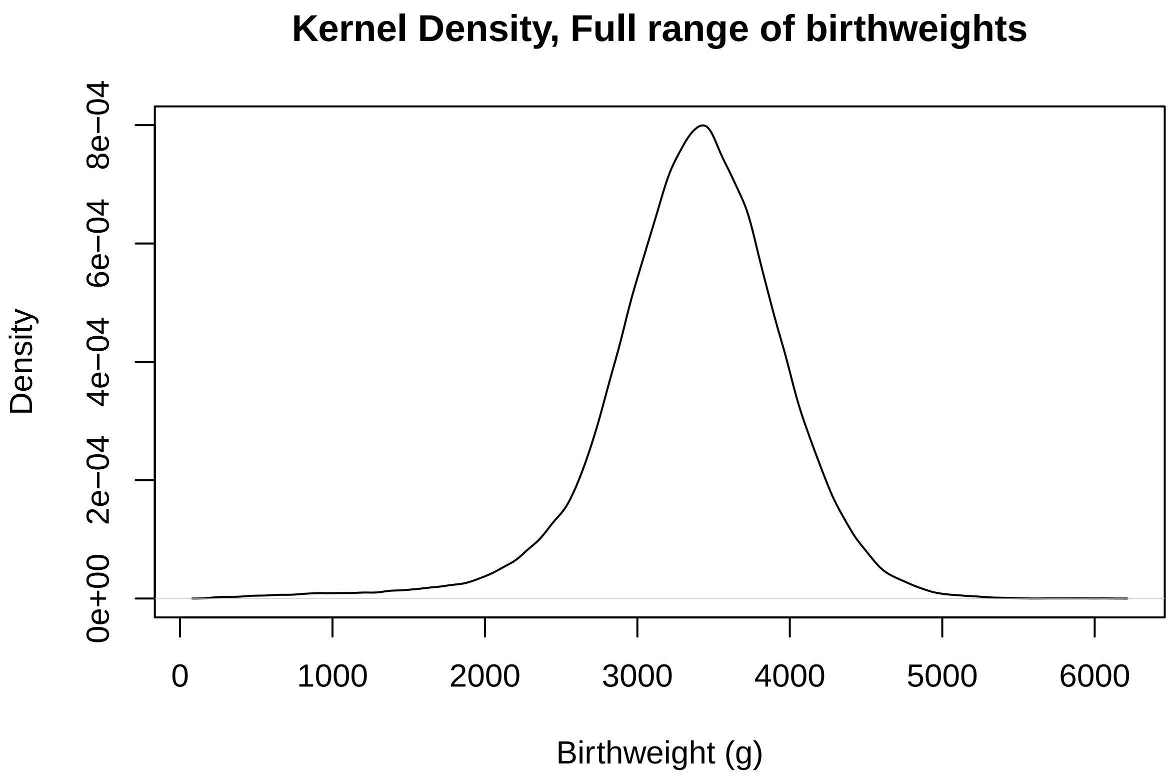
```
d0 <- density(mom_dt[tobacco==0,dbrwt], kernel="gaussian", bw="sj",
              adjust=1, weights=mom_dt[tobacco==0,norm_prop_weights])
d1 <- density(mom_dt[tobacco==1,dbrwt], kernel="gaussian", bw="sj",
              adjust=1, weights=mom_dt[tobacco==1,norm_prop_weights])

plot(d0, col="blue", main="Counterfactual Densities",
     xlab = "Birthweight (g)")
lines(d1, col="red")
legend(1, 0.0008, legend=c("Non-smoker", "Smoker"),
       col=c("blue","red"), lty=1, cex=0.8)
```



In order to get the counterfactual densities, we weight the treated and control groups by the normalized inverse propensity scores, similar to our calculation above. We use the density() function with a Gaussian kernel and a bandwidth using the methods of Sheather & Jones (1991) which uses pilot estimation of derivatives. This results in a bandwidth of 49.83 for the control and a bandwidth of 68.9 for the treated. As we expect from our estimates of the ATE, the density of the birthweights for the treated group is shifted to the left of the control group (lower birthweights).

```
d_all <- density(mom_dt$dbrwt, kernel="gaussian", bw="sj", adjust=1)
plot(d_all, col="black", main="Kernel Density, Full range of birthweights",
     xlab = "Birthweight (g)")
```



```r
# kernel estimator at 3000 g
# use bandwidth selected above: h = 49
h <- d_all$bw
ke_3000 <- 1 / (nrow(mom_dt) * h * sqrt(2*pi)) * sum(exp(-0.5 * (((3000-mom_dt$dbrwt)/h)^2 )))
```

We plot the kernel density over the entire range of birthweights, using the Gaussian kernel and selecting the bandwidth using the methods of Sheather & Jones (1991) which uses pilot estimation of derivatives. For the entire range of birthweights, this gives us a bandwidth of 49.05. We then calculate the kernel estimator by hand at a weight of 3000 grams using the Gaussian kernel and this bandwidth. Our formula is

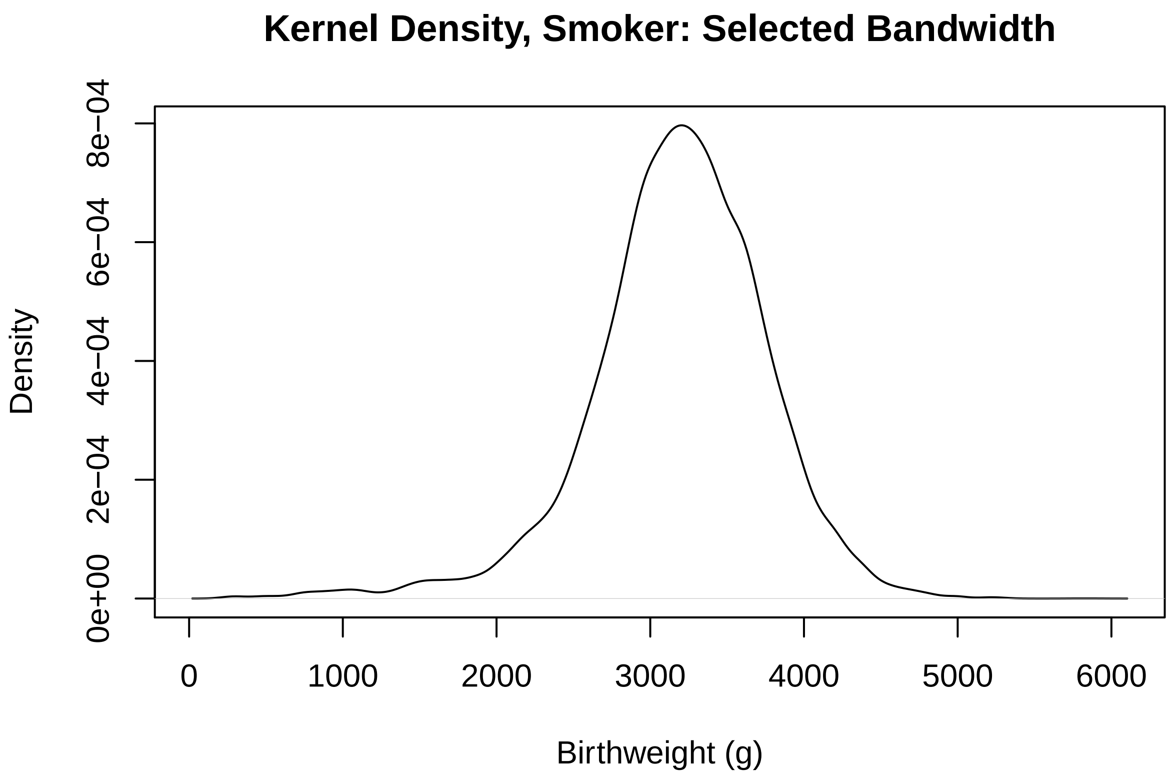
$$\hat{f}(x) = \frac{1}{nh}\frac{1}{\sqrt{2\pi}}\sum_{i}^{n} e^{-1/2(\frac{x-x_i}{h})^2}$$

where $x = 3000$, $n = 114610$, and $h = 49.05$. We get that the density at 3000g is 5.4×10^{-4}, which visually corresponds to our plot of the density.

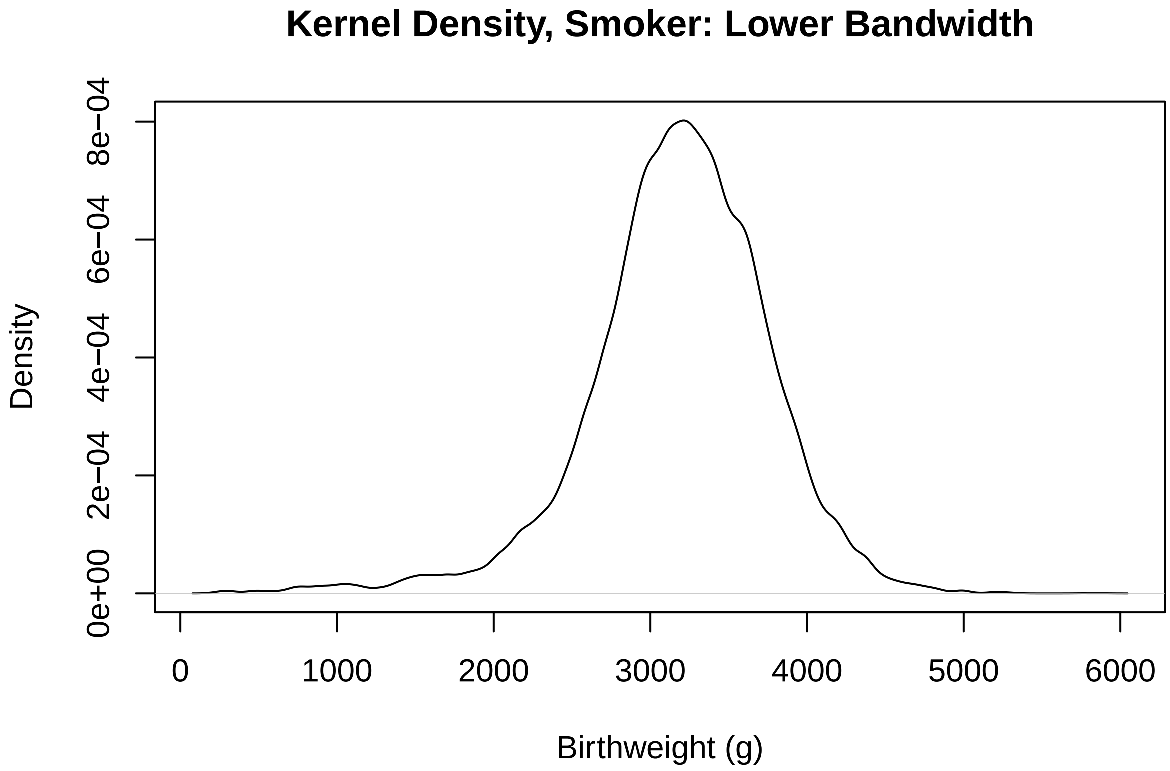
2e

Take one of your densities and display an estimate of the density using different bandwidths as well as the one you settled on. What happens with bigger (smaller) bandwidths?

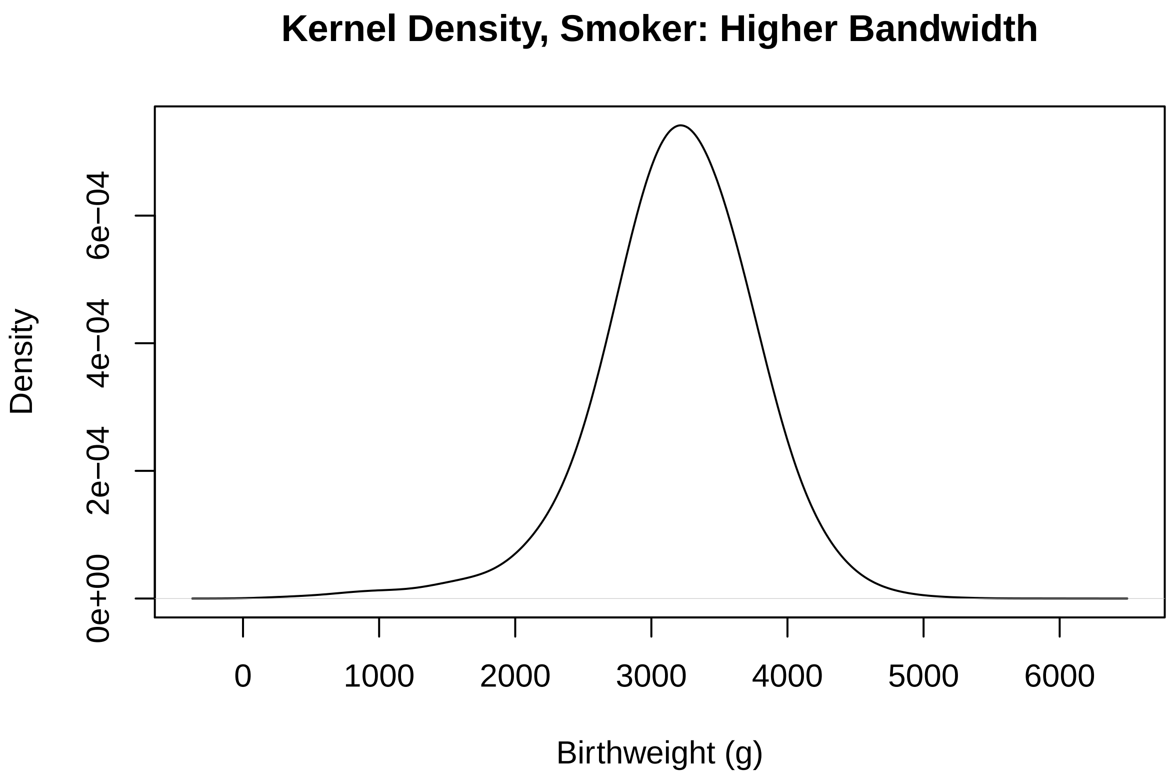
```r
# plot bandwidth selected
plot(d1, col="black", main="Kernel Density, Smoker: Selected Bandwidth",
     xlab = "Birthweight (g)")
```



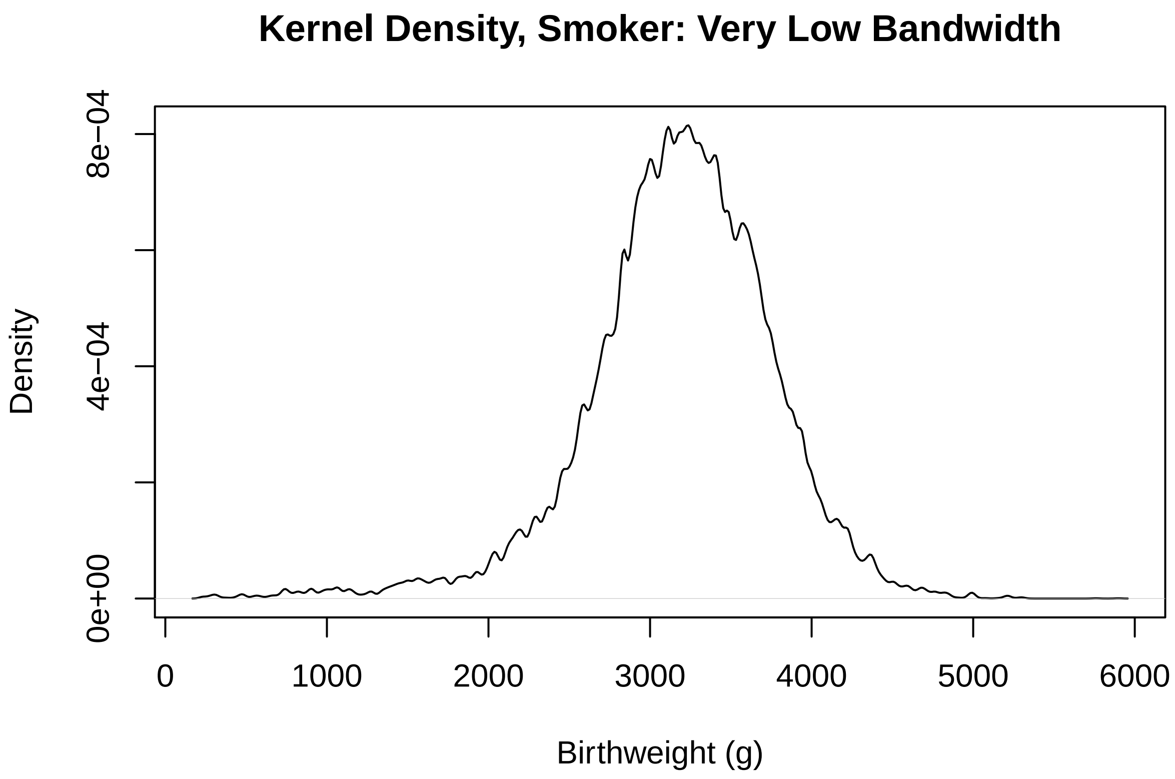
```
# bw of treated group is 68.9, what if we used 49.8 like control?
d1_low <- density(mom_dt[tobacco==1,dbrwt], kernel="gaussian", bw=49.8, adjust=1,
                  weights=mom_dt[tobacco==1,norm_prop_weights])
plot(d1_low, col="black", main="Kernel Density, Smoker: Lower Bandwidth",
     xlab = "Birthweight (g)")
```



```r
# what if we raised it?
d1_high <- density(mom_dt[tobacco==1,dbrwt], kernel="gaussian", bw=200, adjust=1,
                   weights=mom_dt[tobacco==1,norm_prop_weights])
plot(d1_high, col="black", main="Kernel Density, Smoker: Higher Bandwidth",
     xlab = "Birthweight (g)")
```



```r
# what if we made it very low
d1_verylow <- density(mom_dt[tobacco==1,dbrwt], kernel="gaussian", bw=20, adjust=1,
                      weights=mom_dt[tobacco==1,norm_prop_weights])
plot(d1_verylow, col="black", main="Kernel Density, Smoker: Very Low Bandwidth",
     xlab = "Birthweight (g)")
```



We use the treated group counterfactual density and try a few different bandwidths. First, we plot our selected bandwidth, 68.9. Next, since the treated group is smaller than the control group, the bandwidth selected had been larger than that of the control group, so we plot the density had we used the same bandwidth as the control, ~ 49.8. Next we crank up the bandwidth to 200 and finally we drop it to 20. The higher the bandwidth, the smoother the density. At a bandwidth of 20, the density becomes a lot choppier. This is what we expect as a lower bandwidth includes fewer points and thus produces less smoothing.

2f

What are the benefits of the weighting approach (from part c)? What are the potential drawbacks? Pay particular attention to to the issue of people with extremely high and extremely low values of the propensity score.

The benefits of the weighting approach in part c are that we can get a consistent estimate of the ATE and TOT that is also efficient. One potential drawback is that we had to first estimate the propensity score, which we are not certain we estimate correctly. Additionally, there are problems with overlap if the treated and control groups are very different. This could make it so people with extremely high propensity scores are only found in the treated group and people with extremely low propensity scores are only found in the control group. This will cause our weighting methods to still not perform well. However, if this is an issue, we can trim the observations above or below certain propensity scores, as Imbens did in the LaLonde data.

2g

Present your findings and interpret the results on the relationship between birthweight and smoking. For the estimates in parts (b) and (c), consider which of the following conditions must hold in order for that estimate to be valid: i. The treatment effect heterogeneity is linear in the propensity score. ii. The treatment effect heterogeneity is not linear in the propensity score. iii. The decision to smoke is completely randomly assigned. iv. Conditional on the exogenous variables the decision to smoke is randomly assigned.

Our findings are included with the relevant sections above. All findings are consistent with one another in finding that smoking causes lower birthweights – we see this both ATE's estimated and in the counterfactual densities. In part (b), we include the propensity score in a regression. In order for this estimate to be valid, we need assumption i and iv to hold (assumption i would be the case where we interact the propensity score with the treatment, as discussed in part b). In part (c), we again need iv to hold.