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?

Even if our assumption of "selection on observables" holds, our estimates from Pset1a may be biased if $E[D|X]$ is not linear in X. That is our estimates from Pset1a impose a linear functional form on the relationship between Y and X, or between $E[D|X]$ and X, causing our estimates to be biased if the true relationship is not linear. To address the potential misspecification bias, we can instead use a nonparametric method to control for X, or estimate $E[D|X]$ nonparametrically.

(b) Consider a series estimator. 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?

In our series estimator, we include higher order terms of mother's and father's age and education, to allow for a more flexible functional form. We also include interaction terms between mother's race and mother's age and splines for mother's age with knots at 18 and 35 (because pregnant women under the age of 18 or above the age of 35 may have higher health risks with pregnancy) and splines for mother's education with a knot at 12, that is we allow for a differential relationship depending on whether the mother is a highschool dropout or not. Specifically, we estimate the model:

$$\begin{gathered}
birthweight_i = \alpha_i + \beta smoking_i + \delta_1 state_i + \delta_2 countypop_i + \delta_3 mother\_race_i + \delta_4 mother\_age_i + \delta_5 mother\_age_i^2 + \\
\delta_6 mother\_age_i^3 + \delta_7 mother\_educ_i + \delta_8 mother\_educ_i^2 + \delta_9 mother\_educ_i^3 + \delta_{10} marital\_status_i + \delta_{11} father\_age_i + \\
\delta_{12} father\_age_i^2 + \delta_{13} father\_age_i^3 + \delta_{14} father\_educ_i + \delta_{15} father\_educ_i^2 + \delta_{16} father\_educ_i^3 + \\
\delta_{17} mother\_race_i X mother\_age_i + \delta_{18} mother\_race_i X mother\_age_i^2 + \delta_{19} I(mother\_age_i > 18)(mother\_age_i - 18)^3 + \\
\delta_{20} I(mother\_age_i > 35)(mother\_age_i - 35)^3 + \delta_{21} I(mother\_educ_i >= 12)(mother\_educ_i - 12)^3 + \epsilon_i
\end{gathered}$$

```r
# Generate higher order terms
mom_dt[,c("dmage2", "dmage3", "dfage2", "dfage3", "dmeduc2", "dmeduc3", "dfeduc2", "dfeduc3") := .(dmage^2, d

# Generate splines for mother's age above 18 and above 35
mom_dt[dmage >= 18 , dmage_adult := (dmage - 18)^3]
mom_dt[dmage < 18, dmage_adult := 0]
mom_dt[dmage > 35, dmage_ger := (dmage-35)^3]
mom_dt[dmage <= 35, dmage_ger := 0]
#View(mom_dt[,.(dmage, dmage_adult, dmage_ger)])

# Generate splines for mother's education highschool graduate or more
mom_dt[dmeduc >= 12, dmeduc_hs := (dmeduc-12)^3]
mom_dt[dmeduc < 12, dmeduc_hs := 0]
#View(mom_dt[,.(dmeduc, dmeduc_dropout, dmeduc_hs)])
```

```
lm1 <- lm(dbrwt ~ tobacco + factor(stresfip) + factor(cntocpop) + factor(mrace3) + dmage + dmage2 + dmage3 +

summary(lm1)
```

```
##
## Call:
## lm(formula = dbrwt ~ tobacco + factor(stresfip) + factor(cntocpop) +
##     factor(mrace3) + dmage + dmage2 + dmage3 + dmeduc + dmeduc2 +
##     dmeduc3 + factor(dmar) + dfage + dfage2 + dfage3 + dfeduc +
##     dfeduc2 + dfeduc3 + factor(mrace3) * dmage + factor(mrace3) *
##     dmage2 + dmage_adult + dmage_ger + dmeduc_hs, data = mom_dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3267.2  -304.9    26.6   354.9  2831.3
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -7.899e+03  3.238e+03  -2.440 0.014708 *
## tobacco                -2.148e+02  4.895e+00 -43.873  < 2e-16 ***
## factor(stresfip)4       1.055e+02  4.245e+02   0.248 0.803806
## factor(stresfip)6       2.404e+02  2.123e+02   1.132 0.257436
## factor(stresfip)8      -1.081e+02  4.245e+02  -0.255 0.798990
## factor(stresfip)9      -1.112e+02  3.149e+02  -0.353 0.723965
## factor(stresfip)10      1.162e+02  1.393e+02   0.835 0.403913
## factor(stresfip)11     -5.332e+01  4.245e+02  -0.126 0.900056
## factor(stresfip)12     -2.602e+01  1.747e+02  -0.149 0.881576
## factor(stresfip)13      9.353e+01  2.326e+02   0.402 0.687555
## factor(stresfip)17      3.134e+02  2.880e+02   1.088 0.276509
## factor(stresfip)19     -3.402e+01  4.246e+02  -0.080 0.936141
## factor(stresfip)21      6.107e+02  2.880e+02   2.121 0.033942 *
## factor(stresfip)23      3.063e+02  5.852e+02   0.523 0.600639
## factor(stresfip)24      1.620e+02  1.406e+02   1.152 0.249504
## factor(stresfip)25      2.763e+02  2.880e+02   0.959 0.337323
## factor(stresfip)26      2.302e+02  3.552e+02   0.648 0.516909
## factor(stresfip)27      4.829e+02  5.852e+02   0.825 0.409268
## factor(stresfip)29      7.286e+02  5.852e+02   1.245 0.213145
## factor(stresfip)31      4.433e+02  5.852e+02   0.758 0.448693
## factor(stresfip)32      1.395e+02  5.853e+02   0.238 0.811588
## factor(stresfip)34      1.014e+02  1.349e+02   0.752 0.452216
## factor(stresfip)36      3.376e+01  1.439e+02   0.235 0.814454
## factor(stresfip)37      8.072e+01  2.123e+02   0.380 0.703783
## factor(stresfip)38     -3.263e+02  5.852e+02  -0.558 0.577097
## factor(stresfip)39      6.587e+01  1.378e+02   0.478 0.632559
## factor(stresfip)40      1.650e+02  4.246e+02   0.389 0.697531
## factor(stresfip)42      1.490e+02  1.343e+02   1.110 0.267067
## factor(stresfip)44     -1.068e+02  5.852e+02  -0.182 0.855222
## factor(stresfip)45     -4.561e+02  3.149e+02  -1.449 0.147455
## factor(stresfip)46      1.352e+01  5.852e+02   0.023 0.981568
## factor(stresfip)47      2.028e+02  2.685e+02   0.755 0.450147
## factor(stresfip)51     -7.483e+00  1.830e+02  -0.041 0.967379
## factor(stresfip)53     -1.005e+02  3.552e+02  -0.283 0.777228
## factor(stresfip)54      8.864e+01  1.486e+02   0.596 0.550938
## factor(stresfip)55      7.845e+01  3.149e+02   0.249 0.803253
## factor(cntocpop)1       2.782e+01  5.358e+00   5.191 2.09e-07 ***
## factor(cntocpop)2       4.363e+01  4.505e+00   9.686  < 2e-16 ***
## factor(cntocpop)3       3.850e+01  4.996e+00   7.707 1.30e-14 ***
## factor(mrace3)2         1.928e+02  2.924e+02   0.659 0.509594
## factor(mrace3)3         2.471e+01  8.967e+01   0.276 0.782900
## dmage                      1.784e+03  5.433e+02   3.283 0.001027 **
## dmage2                    -9.695e+01  3.037e+01  -3.192 0.001412 **
## dmage3                     1.774e+00  5.647e-01   3.142 0.001681 **
## dmeduc                     6.845e+01  2.702e+01   2.533 0.011306 *
## dmeduc2                   -9.172e+00  2.867e+00  -3.199 0.001378 **
## dmeduc3                    3.581e-01  9.862e-02   3.631 0.000282 ***
## factor(dmar)2             -5.825e+01  5.216e+00 -11.166  < 2e-16 ***
## dfage                      6.042e+00  7.978e+00   0.757 0.448827
## dfage2                    -1.322e-01  2.294e-01  -0.576 0.564528
## dfage3                     9.445e-04  2.137e-03   0.442 0.658586
## dfeduc                    -5.522e+01  1.934e+01  -2.855 0.004300 **
## dfeduc2                    5.421e+00  1.748e+00   3.101 0.001931 **
## dfeduc3                   -1.562e-01  5.088e-02  -3.071 0.002136 **
## dmage_adult               -1.754e+00  5.687e-01  -3.083 0.002047 **
## dmage_ger                 -1.313e-01  6.628e-02  -1.982 0.047529 *
## dmeduc_hs                 -1.318e+00  2.756e-01  -4.783 1.73e-06 ***
## factor(mrace3)2:dmage  -3.197e+01  2.017e+01  -1.584 0.113094
## factor(mrace3)3:dmage  -1.331e+01  6.954e+00  -1.914 0.055621 .
## factor(mrace3)2:dmage2  6.043e-01  3.430e-01   1.762 0.078106 .
## factor(mrace3)3:dmage2  1.915e-01  1.299e-01   1.474 0.140520
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 569.5 on 114549 degrees of freedom
## Multiple R-squared:  0.05321,    Adjusted R-squared:  0.05271
## F-statistic: 107.3 on 60 and 114549 DF,  p-value: < 2.2e-16
```

A benefit to this approach is that it is straightforward and relatively easy to implement and interpret. However, a drawback to this approach is that the choice of which covariates, higher order terms, interaction terms, and knots to include is arbitrary, so our series estimator may not accurately capture the true data generating process and we run the risk of overfitting our model or omitting an important control variable.

(c) Use the LASSO to determine which covariates (and higher order terms) to include in your regression from part (b). Do you end up dropping some covariates that you had thought might be necessary to include?

We follow the procedure suggested by Belloni, Chernozhukov, and Hansen by first applying the LASSO to the equation

$$birthweight_i = \alpha_i + \beta smoking_i + \delta_1 state_i + \delta_2 countypop_i + \delta_3 mother\_race_i + \delta_4 mother\_age_i + \delta_5 mother\_age_i^2 + \\ \delta_6 mother\_age_i^3 + \delta_7 mother\_educ_i + \delta_8 mother\_educ_i^2 + \delta_9 mother\_educ_i^3 + \delta_{10} marital\_status_i + \delta_{11} father\_age_i + \\ \delta_{12} father\_age_i^2 + \delta_{13} father\_age_i^3 \delta_{14} father\_educ_i + \delta_{15} father\_educ_i^2 + \delta_{16} father\_educ_i^3 + \delta_{17} mother\_race_i X mother\_age_i + \\ \delta_{18} mother\_race_i X mother\_age_i^2 + \delta_{19} I(mother\_age_i > 18)(mother\_age_i - 18)^3 + \\ \delta_{20} I(mother\_age_i > 35)(mother\_age_i - 35)^3 + \delta_{21} I(mother\_educ_i >= 12)(mother\_educ_i - 12)^3 + \epsilon_i$$

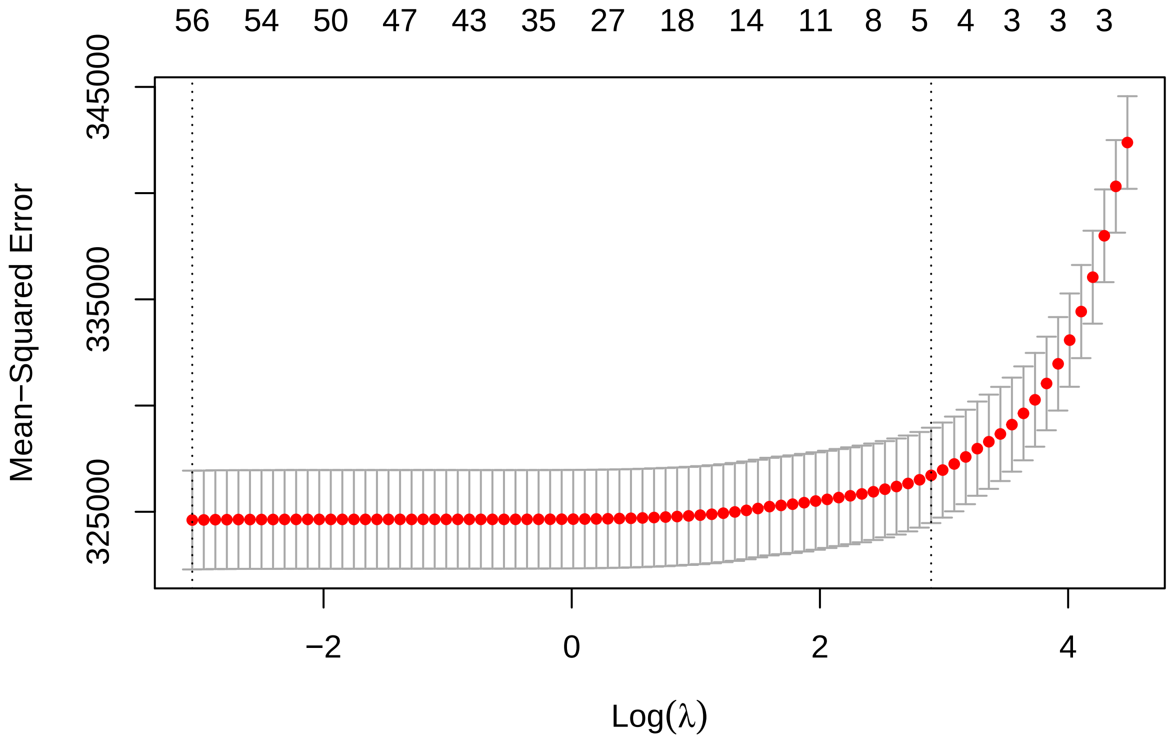
We then apply the LASSO to the equation

$$smoking_i = \gamma_0 + \gamma_1 state_i + \gamma_2 countypop_i + \gamma_3 mother\_race_i + \gamma_4 mother\_age_i + \gamma_5 mother\_age_i^2 + \\ \gamma_6 mother\_age_i^3 + \gamma_7 mother\_educ_i + \gamma_8 mother\_educ_i^2 + \gamma_9 mother\_educ_i^3 + \gamma_{10} marital\_status_i + \gamma_{11} father\_age_i + \\ \gamma_{12} father\_age_i^2 + \gamma_{13} father\_age_i^3 \gamma_{14} father\_educ_i + \gamma_{15} father\_educ_i^2 + \gamma_{16} father\_educ_i^3 + \gamma_{17} mother\_race_i X mother\_age_i + \\ \gamma_{18} mother\_race_i X mother\_age_i^2 + \gamma_{19} I(mother\_age_i > 18)(mother\_age_i - 18)^3 + \\ \gamma_{20} I(mother\_age_i > 35)(mother\_age_i - 35)^3 + \gamma_{21} I(mother\_educ_i >= 12)(mother\_educ_i - 12)^3 + \epsilon_i$$

Then we regress $birthweight_i$ on $smoking_i$ and all of the covariates selected by LASSO in either equation 1 or 2 above. In each case, we cross-validate the LASSO and choose the value of lambda such that the mean cross-validated error is within 1 standard error of the minimum.

```r
#First convert factor variables in dataset to factor variables
mom_dt[,stresfip := as.factor(stresfip)]
mom_dt[,cntocpop := as.factor(cntocpop)]
mom_dt[,mrace3 := as.factor(mrace3)]
mom_dt[,dmar := as.factor(dmar)]

#Lasso for lm1
x1 <- model.matrix(lm1, data = mom_dt) #create X matrix
lasso_y <- cv.glmnet(x1, mom_dt$dbrwt, family = "gaussian", standardize = TRUE, intercept = TRUE, alpha = 1,
plot(lasso_y)
```



```r
lasso_y$glmnet.fit
```

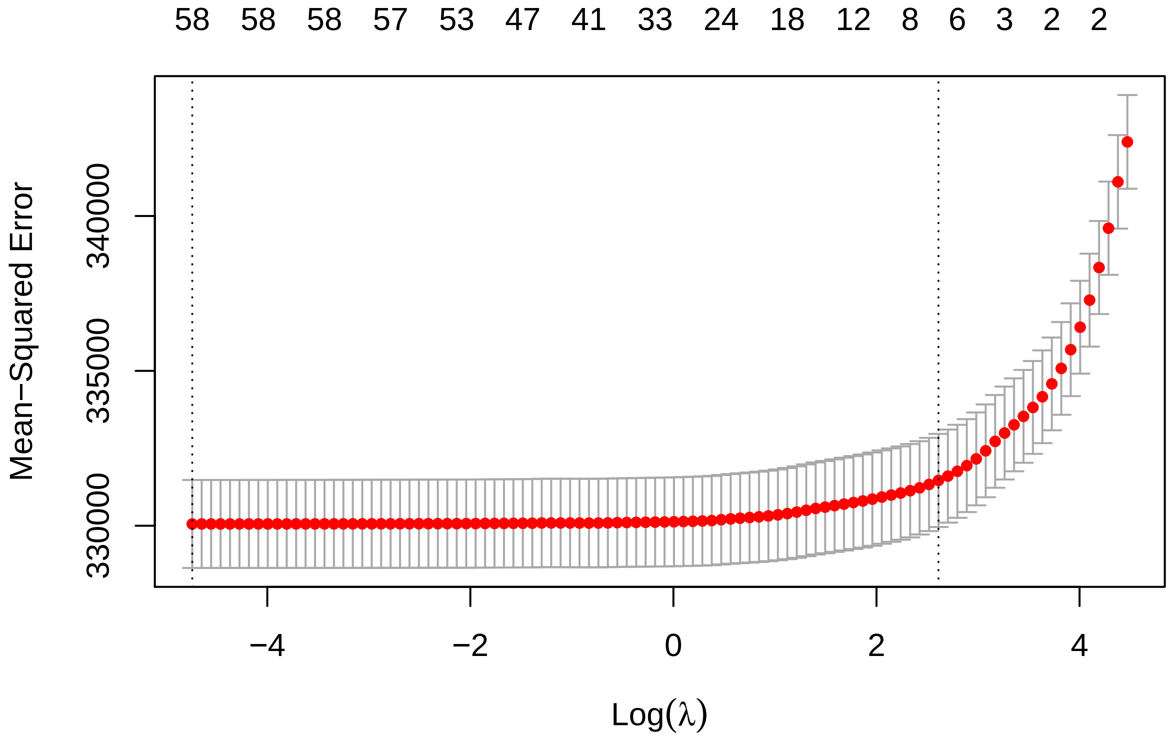
```
##
## Call:  glmnet(x = x1, y = mom_dt$dbrwt, family = "gaussian", standardize = TRUE,      intercept = TRUE, al
##
##    Df %Dev Lambda
## 1   0 0.00 88.020
## 2   3 0.61 80.200
## 3   3 1.30 73.080
## 4   3 1.87 66.580
## 5   3 2.34 60.670
## 6   3 2.73 55.280
## 7   3 3.06 50.370
## 8   3 3.33 45.890
## 9   3 3.55 41.820
## 10  3 3.74 38.100
## 11  3 3.90 34.720
## 12  3 4.02 31.630
## 13  3 4.13 28.820
## 14  4 4.23 26.260
## 15  4 4.34 23.930
## 16  5 4.44 21.800
## 17  5 4.53 19.870
## 18  5 4.60 18.100
## 19  5 4.66 16.490
## 20  5 4.71 15.030
## 21  5 4.75 13.690
## 22  7 4.79 12.480
## 23  8 4.83 11.370
## 24  8 4.86 10.360
## 25  8 4.88  9.438
## 26 10 4.91  8.600
## 27 10 4.94  7.836
## 28 11 4.96  7.140
## 29 12 4.99  6.505
## 30 12 5.01  5.927
## 31 12 5.02  5.401
## 32 14 5.04  4.921
## 33 14 5.07  4.484
## 34 14 5.09  4.086
## 35 14 5.12  3.723
## 36 14 5.13  3.392
## 37 17 5.15  3.091
## 38 17 5.16  2.816
## 39 17 5.18  2.566
## 40 18 5.19  2.338
## 41 20 5.20  2.130
## 42 21 5.20  1.941
## 43 22 5.21  1.769
## 44 24 5.22  1.611
## 45 25 5.22  1.468
## 46 27 5.23  1.338
## 47 29 5.23  1.219
## 48 29 5.23  1.111
## 49 31 5.24  1.012
## 50 33 5.24  0.922
## 51 35 5.24  0.840
## 52 35 5.24  0.766
## 53 36 5.25  0.698
## 54 37 5.25  0.636
## 55 41 5.25  0.579
## 56 41 5.25  0.528
## 57 44 5.25  0.481
## 58 43 5.25  0.438
## 59 44 5.25  0.399
## 60 44 5.26  0.364
## 61 46 5.26  0.331
## 62 46 5.26  0.302
## 63 45 5.26  0.275
## 64 47 5.26  0.251
## 65 47 5.26  0.228
## 66 47 5.26  0.208
## 67 47 5.26  0.190
## 68 47 5.26  0.173
## 69 48 5.26  0.157
## 70 50 5.26  0.144
## 71 51 5.26  0.131
## 72 53 5.26  0.119
## 73 53 5.26  0.108
## 74 53 5.26  0.099
## 75 53 5.26  0.090
## 76 54 5.26  0.082
## 77 54 5.26  0.075
## 78 54 5.26  0.068
## 79 55 5.26  0.062
## 80 55 5.26  0.057
## 81 56 5.28  0.052
## 82 56 5.28  0.047
```

```
coef(lasso_y, s = "lambda.1se") #Coefficients from lasso for lambda that gets error within 1 se of the minimu
```

```
## 62 x 1 sparse Matrix of class "dgCMatrix"
##                                    1
## (Intercept)            3427.4911084
## (Intercept)               .
## tobacco                -168.7874259
## factor(stresfip)4         .
## factor(stresfip)6         .
## factor(stresfip)8         .
## factor(stresfip)9         .
## factor(stresfip)10        .
## factor(stresfip)11        .
## factor(stresfip)12        .
## factor(stresfip)13        .
## factor(stresfip)17        .
## factor(stresfip)19        .
## factor(stresfip)21        .
## factor(stresfip)23        .
## factor(stresfip)24        .
## factor(stresfip)25        .
## factor(stresfip)26        .
## factor(stresfip)27        .
## factor(stresfip)29        .
## factor(stresfip)31        .
## factor(stresfip)32        .
## factor(stresfip)34        .
## factor(stresfip)36        .
## factor(stresfip)37        .
## factor(stresfip)38        .
## factor(stresfip)39        .
## factor(stresfip)40        .
## factor(stresfip)42        .
## factor(stresfip)44        .
## factor(stresfip)45        .
## factor(stresfip)46        .
## factor(stresfip)47        .
## factor(stresfip)51        .
## factor(stresfip)53        .
## factor(stresfip)54        .
## factor(stresfip)55        .
## factor(cntocpop)1         .
## factor(cntocpop)2         .
## factor(cntocpop)3         .
## factor(mrace3)2         -68.1185019
## factor(mrace3)3        -155.3097234
## dmage                     0.4367174
## dmage2                    .
## dmage3                    .
## dmeduc                    .
## dmeduc2                   .
## dmeduc3                   .
## factor(dmar)2           -77.5924121
## dfage                     .
## dfage2                    .
## dfage3                    .
## dfeduc                    .
## dfeduc2                   .
## dfeduc3                   .
## dmage_adult               .
## dmage_ger                 .
## dmeduc_hs                 .
## factor(mrace3)2:dmage     .
## factor(mrace3)3:dmage     .
## factor(mrace3)2:dmage2    .
## factor(mrace3)3:dmage2    .
```

```
#Lasso for lm2
lm2 <- lm(tobacco ~ factor(stresfip) + factor(cntocpop) + factor(mrace3) + dmage + dmage2 + dmage3 + dmeduc +

x2 <- model.matrix(lm2, data = mom_dt) #create X matrix
lasso_d <- cv.glmnet(x2, mom_dt$dbrwt, family = "gaussian", standardize = TRUE, intercept = TRUE, alpha = 1,
plot(lasso_d)
```



```
lasso_d$glmnet.fit
```

```
##
## Call:  glmnet(x = x2, y = mom_dt$dbrwt, family = "gaussian", standardize = TRUE,      intercept = TRUE, al
##
##      Df %Dev Lambda
## 1     0 0.00 87.460
## 2     2 0.38 79.690
## 3     2 0.83 72.610
## 4     2 1.20 66.160
## 5     2 1.51 60.280
## 6     2 1.76 54.930
## 7     2 1.97 50.050
## 8     2 2.15 45.600
## 9     2 2.30 41.550
## 10    2 2.42 37.860
## 11   2 2.52 34.500
## 12   2 2.60 31.430
## 13   3 2.68 28.640
## 14   3 2.76 26.100
## 15   5 2.84 23.780
## 16   5 2.93 21.660
## 17   5 3.01 19.740
## 18   6 3.07 17.990
## 19   6 3.13 16.390
## 20   6 3.17 14.930
## 21   7 3.22 13.610
## 22   7 3.25 12.400
## 23   8 3.29 11.300
## 24   8 3.31 10.290
## 25   9 3.34  9.378
## 26   9 3.36  8.545
## 27  11 3.38  7.786
## 28  11 3.40  7.094
## 29  11 3.42  6.464
## 30  12 3.43  5.890
## 31  12 3.45  5.367
## 32  13 3.47  4.890
## 33  13 3.48  4.455
## 34  14 3.49  4.060
## 35  14 3.51  3.699
## 36  14 3.53  3.370
## 37  18 3.55  3.071
## 38  19 3.56  2.798
## 39  21 3.57  2.550
## 40  22 3.58  2.323
## 41  22 3.59  2.117
## 42  23 3.60  1.929
## 43  24 3.61  1.757
## 44  24 3.62  1.601
## 45  24 3.63  1.459
## 46  25 3.63  1.329
## 47  27 3.64  1.211
## 48  29 3.64  1.104
## 49  33 3.64  1.006
## 50  32 3.65  0.916
## 51  33 3.65  0.835
## 52  35 3.65  0.761
## 53  36 3.66  0.693
## 54  37 3.66  0.632
## 55  37 3.66  0.575
## 56  37 3.66  0.524
## 57  40 3.67  0.478
## 58  41 3.67  0.435
## 59  41 3.67  0.397
## 60  42 3.67  0.361
## 61  43 3.67  0.329
## 62  46 3.67  0.300
## 63  46 3.67  0.273
## 64  47 3.68  0.249
## 65  47 3.68  0.227
## 66  47 3.68  0.207
## 67  48 3.68  0.188
## 68  49 3.68  0.172
## 69  49 3.68  0.156
## 70  51 3.68  0.142
## 71  53 3.68  0.130
## 72  53 3.69  0.118
## 73  55 3.69  0.108
## 74  55 3.69  0.098
## 75  55 3.69  0.090
## 76  56 3.69  0.082
## 77  57 3.69  0.074
## 78  57 3.69  0.068
## 79  57 3.69  0.062
## 80  57 3.69  0.056
## 81  57 3.69  0.051
## 82  57 3.69  0.047
## 83  57 3.69  0.043
## 84  58 3.69  0.039
## 85  58 3.69  0.035
## 86  58 3.69  0.032
## 87  57 3.69  0.029
## 88  57 3.69  0.027
## 89  57 3.69  0.024
## 90  58 3.69  0.022
## 91  58 3.69  0.020
## 92  58 3.69  0.018
## 93  58 3.69  0.017
## 94  58 3.69  0.015
## 95  58 3.69  0.014
## 96  58 3.69  0.013
## 97  58 3.69  0.012
## 98  58 3.69  0.011
## 99  58 3.69  0.010
## 100 58 3.69  0.009
```

```
coef(lasso_d, s = "lambda.1se") #Coefficients from lasso for lambda that gets error within 1 se of the minimu
```

```
## 61 x 1 sparse Matrix of class "dgCMatrix"
##                                    1
## (Intercept)            3343.38457847
## (Intercept)               .
## factor(stresfip)4         .
## factor(stresfip)6         .
## factor(stresfip)8         .
## factor(stresfip)9         .
## factor(stresfip)10        .
## factor(stresfip)11        .
## factor(stresfip)12        .
## factor(stresfip)13        .
## factor(stresfip)17        .
## factor(stresfip)19        .
## factor(stresfip)21        .
## factor(stresfip)23        .
## factor(stresfip)24        .
## factor(stresfip)25        .
## factor(stresfip)26        .
## factor(stresfip)27        .
## factor(stresfip)29        .
## factor(stresfip)31        .
## factor(stresfip)32        .
## factor(stresfip)34        .
## factor(stresfip)36        .
## factor(stresfip)37        .
## factor(stresfip)38        .
## factor(stresfip)39        .
## factor(stresfip)40        .
## factor(stresfip)42        .
## factor(stresfip)44        .
## factor(stresfip)45        .
## factor(stresfip)46        .
## factor(stresfip)47        .
## factor(stresfip)51        .
## factor(stresfip)53        .
## factor(stresfip)54        .
## factor(stresfip)55        .
## factor(cntocpop)1         .
## factor(cntocpop)2         1.82807490
## factor(cntocpop)3         .
## factor(mrace3)2         -88.97808437
## factor(mrace3)3        -145.73330270
## dmage                     0.33575112
## dmage2                    .
## dmage3                    .
## dmeduc                    3.89585373
## dmeduc2                   .
## dmeduc3                   .
## factor(dmar)2          -113.11457902
## dfage                     .
## dfage2                    .
## dfage3                    .
## dfeduc                    .
## dfeduc2                   0.08884019
## dfeduc3                   .
## dmage_adult               .
## dmage_ger                 .
## dmeduc_hs                 .
## factor(mrace3)2:dmage     .
## factor(mrace3)3:dmage     .
## factor(mrace3)2:dmage2    .
## factor(mrace3)3:dmage2    .
```

Based on these results, we choose to keep the following covariates: an intercept, tobacco, both levels of mother's race, mother's age, mother's education, marital status, and the square of father's education. Thus, we drop a significant number of covariates that we had initially thought would be necessary to include, such as state of residence, county population, higher order terms of mother's age and education, most of the terms for father's age and education other than the square of father's education, and all of our spline and interaction terms. After our LASSO procedure, our final model is:

$$\begin{aligned} birthweight_i = \alpha_i + \beta Smoking_i + \delta_1 mother\_race_i + \delta_2 mother\_age_i + \delta_3 mother\_educ_i + \\ \delta_4 marital\_status_i + \delta_5 father\_educ_i^2 + \epsilon_i \end{aligned}$$

```r
#New model with covariates chosen from 2-stage Lasso above
lm3 <- lm(dbrwt ~ tobacco + factor(mrace3) + dmage + dmeduc + factor(dmar) + dfeduc2, mom_dt)
summary(lm3)
```

```
##
## Call:
## lm(formula = dbrwt ~ tobacco + factor(mrace3) + dmage + dmeduc +
##     factor(dmar) + dfeduc2, data = mom_dt)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3266.2  -304.5    27.5   355.6  2828.1
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3363.47151   13.26200 253.617  < 2e-16 ***
## tobacco          -211.06360    4.84356 -43.576  < 2e-16 ***
## factor(mrace3)2 -216.32638   12.15321 -17.800  < 2e-16 ***
## factor(mrace3)3 -201.64518    5.72732 -35.208  < 2e-16 ***
## dmage              2.07897    0.35726   5.819 5.93e-09 ***
## dmeduc             1.58151    1.02122   1.549   0.1215
## factor(dmar)2    -82.38579    4.92862 -16.716  < 2e-16 ***
## dfeduc2            0.07687    0.03684   2.086   0.0369 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 570.5 on 114602 degrees of freedom
## Multiple R-squared:  0.04968,    Adjusted R-squared:  0.04963
## F-statistic: 855.9 on 7 and 114602 DF,  p-value: < 2.2e-16
```

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?

```
# 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
```

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.

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)

# 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
```

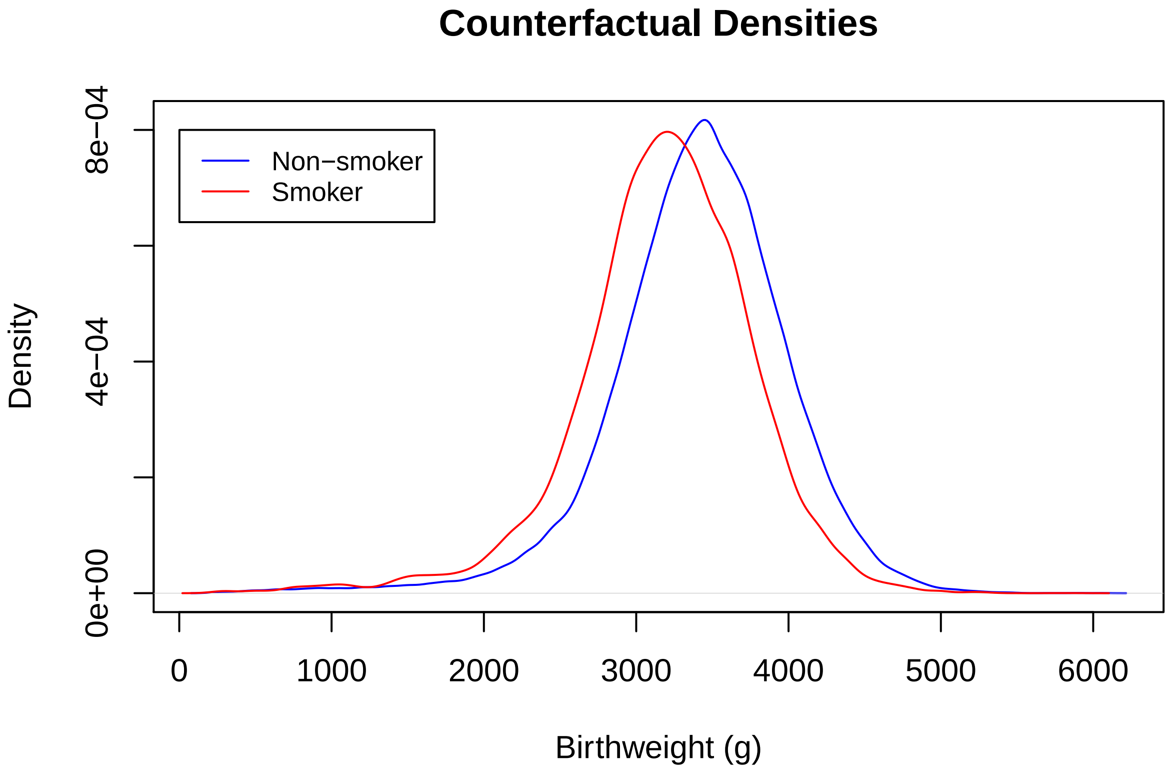
Using inverse propensity score weighting, we get that the ATE is now -225.29 and the TOT is -224.4.

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.

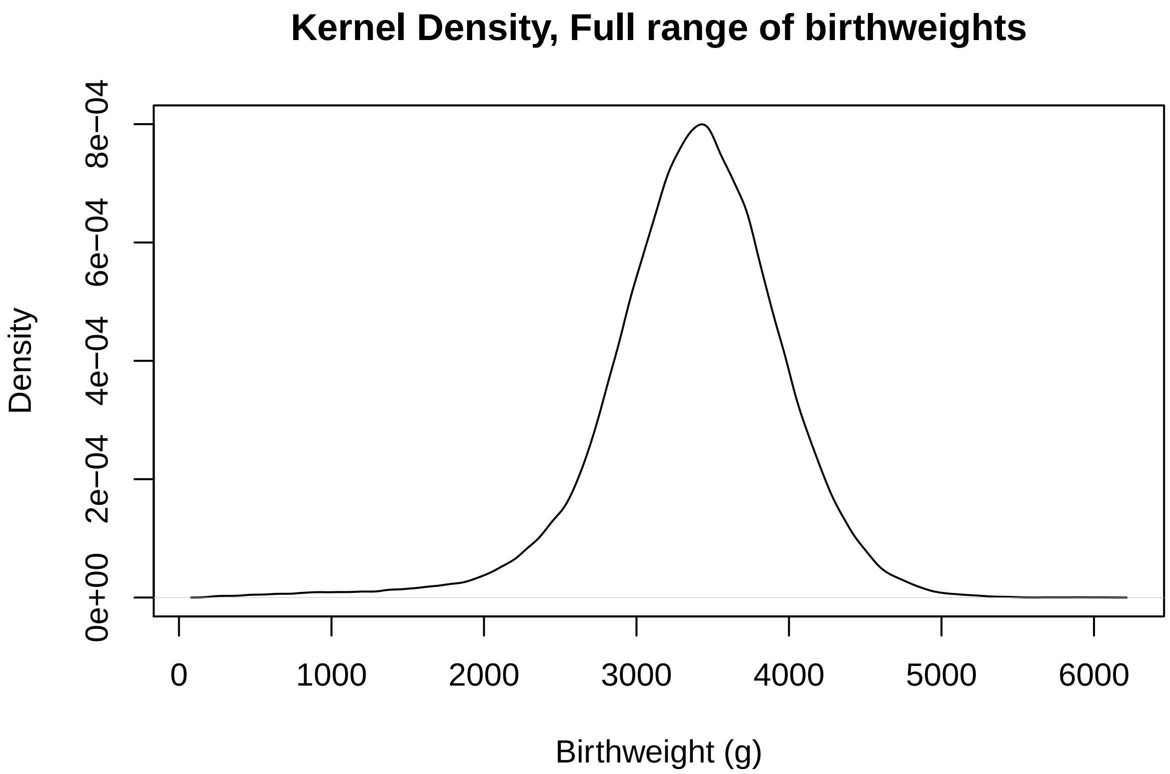
```
d0 <- density(mom_dt[tobacco==0,dbrwt], kernel="gaussian", bw="sj",
              adjust=1, weights=mom_dt[tobacco==0,norm_prop_weights])
d1 <- density(mom_dt[tobacco==1,dbrwt], kernel="gaussian", bw="sj",
              adjust=1, weights=mom_dt[tobacco==1,norm_prop_weights])

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)
```



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)")
```



```
# kernel estimator at 3000 g
# use bandwidth selected above: h = 49
h <- 49
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. We then calculate the kernel estimator by hand at a weight of 3000 grams using the Gaussian kernel and this bandwidth. Our formula is

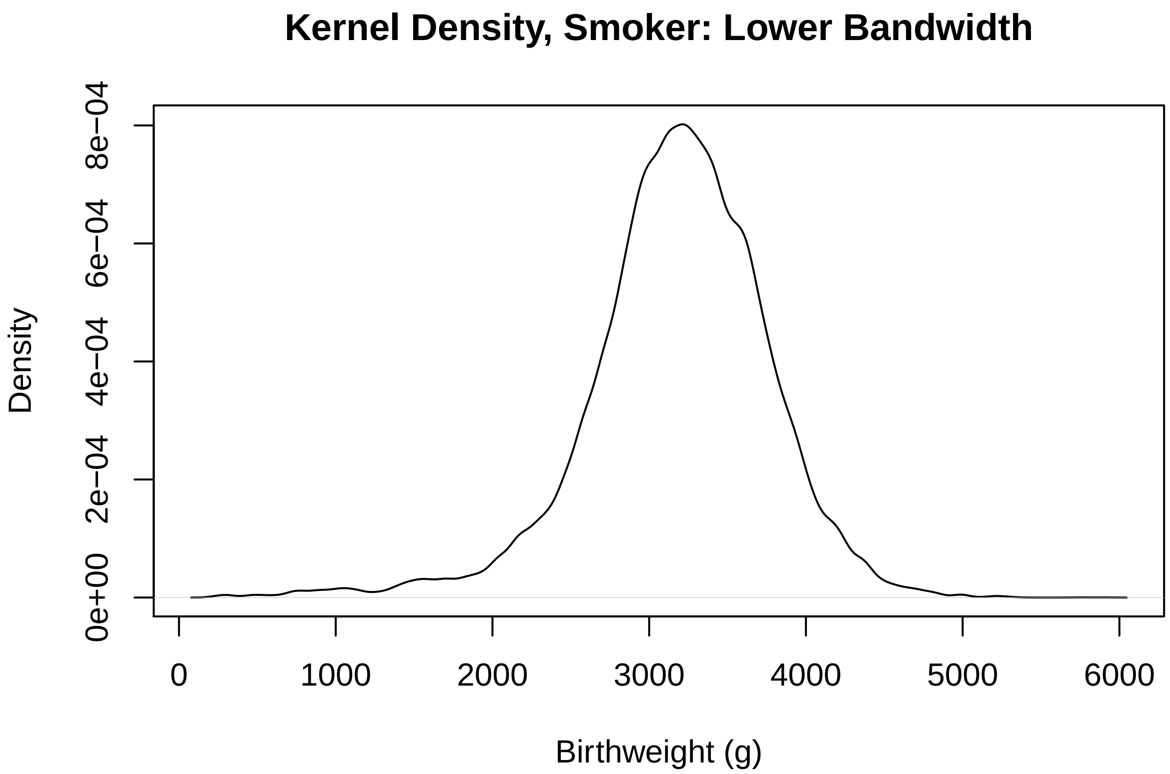
$$\hat{f}(x) = \frac{1}{nh}\frac{1}{\sqrt{2\pi}}\sum_{i}^{n} e^{-1/2(\frac{x-x_i}{h})^2}$$

where $x = 3000$ and $h = 49$. We get that the density at 3000g is 5.4×10^{-4}, which visually corresponds to our plot of the density.

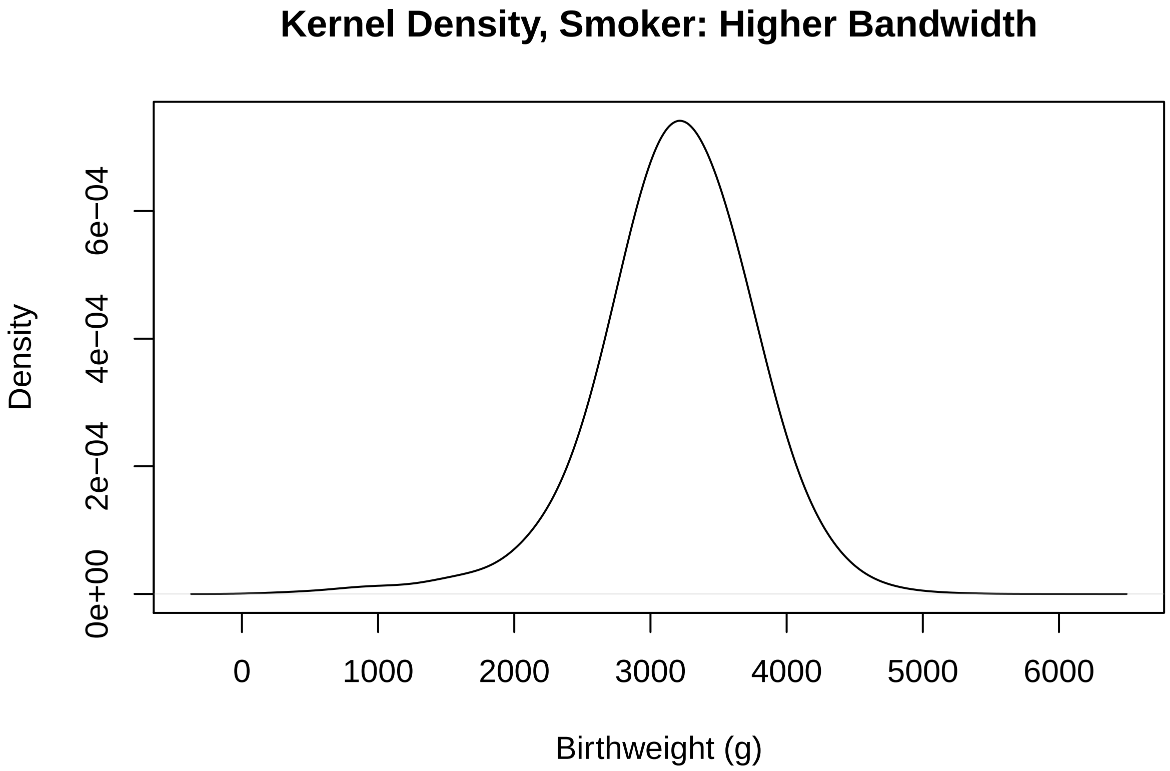
2e

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?

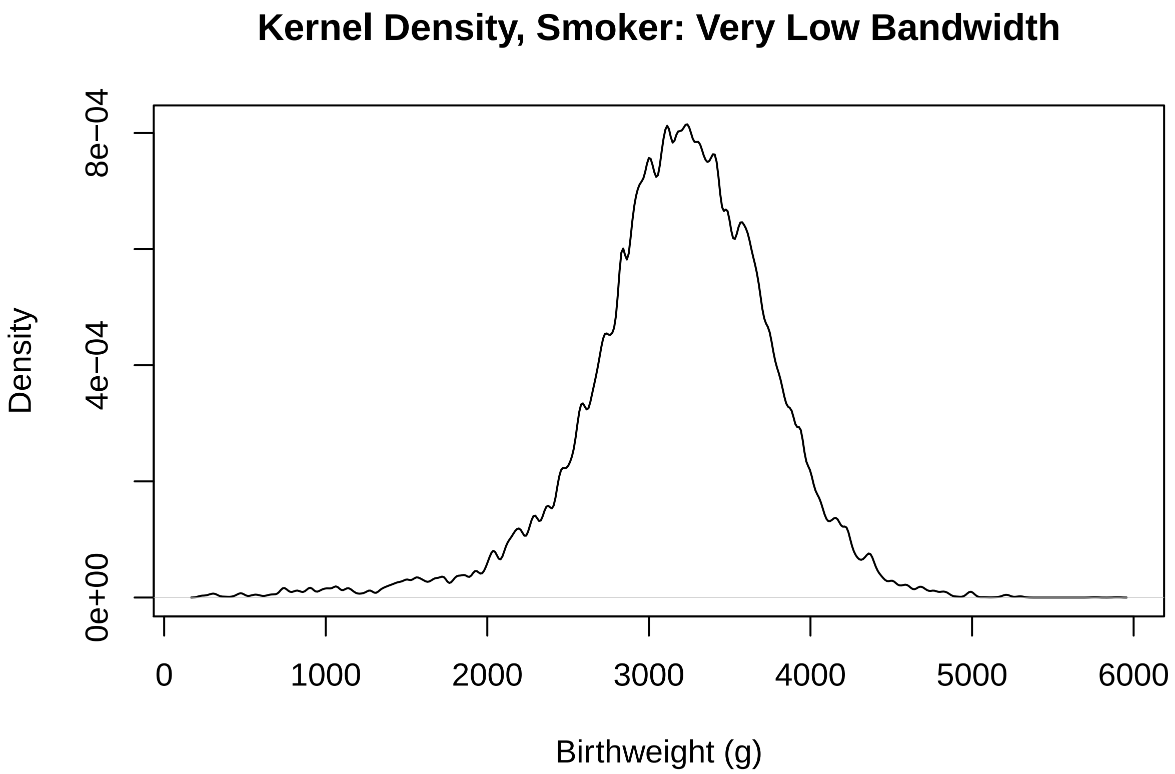
```
# bw of treated group is 68.9, what if we used 49.8 like control?
d1_low <- density(mom_dt[tobacco==1,dbrwt], kernel="gaussian", bw=49.8, adjust=1, weights=mom_dt[tobacco==1,n
plot(d1_low, col="black", main="Kernel Density, Smoker: Lower Bandwidth",
     xlab = "Birthweight (g)")
```



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



```
# 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
plot(d1_verylow, col="black", main="Kernel Density, Smoker: Very Low Bandwidth",
     xlab = "Birthweight (g)")
```



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.

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.