ARE 213 PS 1b

S. Sung, H. Husain, T. Woolley, A. Watt

2021-10-18

Contents

Packages (omitted)

Data cleaning from PS 1a

```
data = read.dta('ps1.dta')
missing_codes = read.csv('missing_codes.csv')
mvars = as.character(missing_codes$varname)
```

```
missing_codes$num_missing = as.integer(0)
for (row in 1:nrow(missing_codes)) {
  var = as.character(missing_codes[row, "varname"])
  code  = as.numeric(missing_codes[row, "missing_code"])
  nmissing = as.integer(sum(data[, var] == code))
  missing_codes$num_missing[missing_codes$varname==var] = nmissing
  data[, var] = na_if(data[, var], code)
}
# Convert all variables with <7 unique values to factor (and 3 additional variables)
factor_vars = c("isllb10", "birmon", "weekday")
for (var in colnames(data)) {
  if (length(unique(data[!is.na(data[, var]), var])) < 7 || var %in% factor_vars) {
    data[, var] = factor(data[, var])
  }
}
# label data
variable_labels_df = read.csv('variable_labels.csv')
variable_labels <- setNames(as.character(variable_labels_df$label), variable_labels_df$varname)
data <- Hmisc::upData(data, labels = variable_labels)

# Dataframe with missing dropped
df = data[complete.cases(data), ]
# Treatment reference level
df$tobacco <- relevel(df$tobacco, ref = "2")  # reference level: 2 = no for tobacco use during pregnanc
```

Problem 1

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

Part (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?

In problem set 1(a), we input observable characteristic in a linear fashion into our crude estimation with OLS. However, it is possible that CEF isn't linear in parameters, leading to bias. Also, even if we have a linear CEF, it's possible we didn't include important interaction terms.

In order to assume linearity, we need (1) joint normality with observables, indicator for maternal smoking, and our outcome variable of interest or (2) saturated model. In their absence, we could have used other non-linear estimation methods discussed in class.

Part (b)

Now, 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?

```
df1b = df %>% select(dbrwt, fmaps, omaps, tobacco,
                     csex, mrace3, preterm, dmage, dfage, dmeduc, dfeduc,
                     ormoth, orfath, disllb, dtotord, dmar, adequacy, nprevist)
```

```
# indicator vars (no higher order terms)
vars1 = names(Filter(is.factor, select(df1b, -c(dbrwt, fmaps, omaps))))
# quantitative vars (need to create higher order terms)
vars2 = names(Filter(is.integer, select(df1b, -c(dbrwt, fmaps, omaps))))

birthweight = df1b$dbrwt
fiveminapgar = df1b$fmaps
oneminapgar = df1b$omaps

x = df1b %>% select(-c(dbrwt, fmaps, omaps))
# Create dummies from factor variables, all interactions, and squared continuous vars
formula1 = as.formula(paste("~ .^2 +", paste0("I(",vars2,"^2)", collapse=' + ') ))
xx <- model.matrix(formula1, x)[, -1]

# Series Regression
reg_1b_dbrwt = lm(birthweight ~ xx)
reg_1b_fmaps = lm(fiveminapgar ~ xx)
reg_1b_omaps = lm(oneminapgar ~ xx)

# rename the coefficients
names(reg_1b_dbrwt$coefficients) <- gsub("xx","",names(reg_1b_dbrwt$coefficients))
names(reg_1b_fmaps$coefficients) <- gsub("xx","",names(reg_1b_fmaps$coefficients))
names(reg_1b_omaps$coefficients) <- gsub("xx","",names(reg_1b_omaps$coefficients))
```

```
      keep = "tobacco1(?!:)",
      perl=TRUE,
```

```
stargazer(reg_1b_dbrwt, reg_1b_fmaps, reg_1b_omaps,
          type = 'latex',
          keep = "tobacco1(?!:)",
          perl=TRUE,
          title = "Series Regressions",
          align = TRUE, no.space = TRUE, font.size = "small",
          notes = c("For brevity, only the original variables are presented;",
                    "300+ second order / interaction terms are not shown."))
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sun, Oct 24, 2021 - 08:58:38 PM % Requires LaTeX packages: dcolumn

REVIEW THIS BEFORE SUBMISSION: Need to change to longtable before final submission**

```
A table enviroment cannot be broken across pages. Delete \begin{table}\centering and \end{table}, repla
```

Part (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?

For brevity, I continue on with the exercise with only one dependent variable (dbrwt), skipping the others (fmaps and omaps). Running a simple lasso with all the included varibles from 1(b), and arbitrarily setting number of variables to be included as 20, we chose lambda = 14.23. This is the lowest lambda iteration that results in a model with 20 non-zero coefficients.

Table 1: Series Regressions

	Dependent variable:		
	birthweight	fiveminapgar	oneminapgar
	(1)	(2)	(3)
tobacco1	-214.419^{***}	0.123	0.040
	(60.526)	(0.077)	(0.138)
Observations	114,610	114,610	114,610
R^2	0.132	0.029	0.022
Adjusted R^2	0.130	0.026	0.020
Residual Std. Error (df = 114304)	545.824	0.698	1.247
F Statistic (df = 305; 114304)	57.135^{***}	11.215^{***}	8.474^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01
			For brevity, only the original variables are presented;
			300+ second order / interaction terms are not shown.

Under this setting, predicted impact of smoking on birthweight is -76.9.

We ended up dropping `dmage` and 'dfage', or age of parents, which we expected to have some correlation with infant's general health and/or weight. On the other hand, variables such as 'csex' and 'preterm', which we expected to have a stronger relationship with infant weight (i.e. male infant might be bigger on average, and previous baby's birth weight might be predictive of the baby of interest), are not dropped as expected.

One problem, however, with what we have done below is that while an interaction term that includes 'dmage' is included, the standalone variable 'dmage' is dropped, which might be undesirable.

```
# use glmnet with alpha=1 for lasso
reg_1c = glmnet(xx, birthweight, family="gaussian", alpha=1)
# print results (Df = # of variables, %Dev = R^2)
print(reg_1c)
```

```
##
## Call:  glmnet(x = xx, y = birthweight, family = "gaussian", alpha = 1)
##
##       Df  %Dev  Lambda
## 1      0  0.00 100.400
## 2      1  0.50  91.460
## 3      3  1.11  83.340
## 4      4  1.95  75.930
## 5      4  2.74  69.190
## 6      5  3.48  63.040
## 7      5  4.24  57.440
## 8      6  4.91  52.340
## 9      7  5.59  47.690
## 10     8  6.17  43.450
## 11     8  6.65  39.590
## 12     8  7.05  36.070
## 13     9  7.38  32.870
## 14     9  7.68  29.950
## 15    11  7.92  27.290
## 16    12  8.14  24.860
## 17    12  8.34  22.660
## 18    14  8.53  20.640
## 19    15  8.71  18.810
```

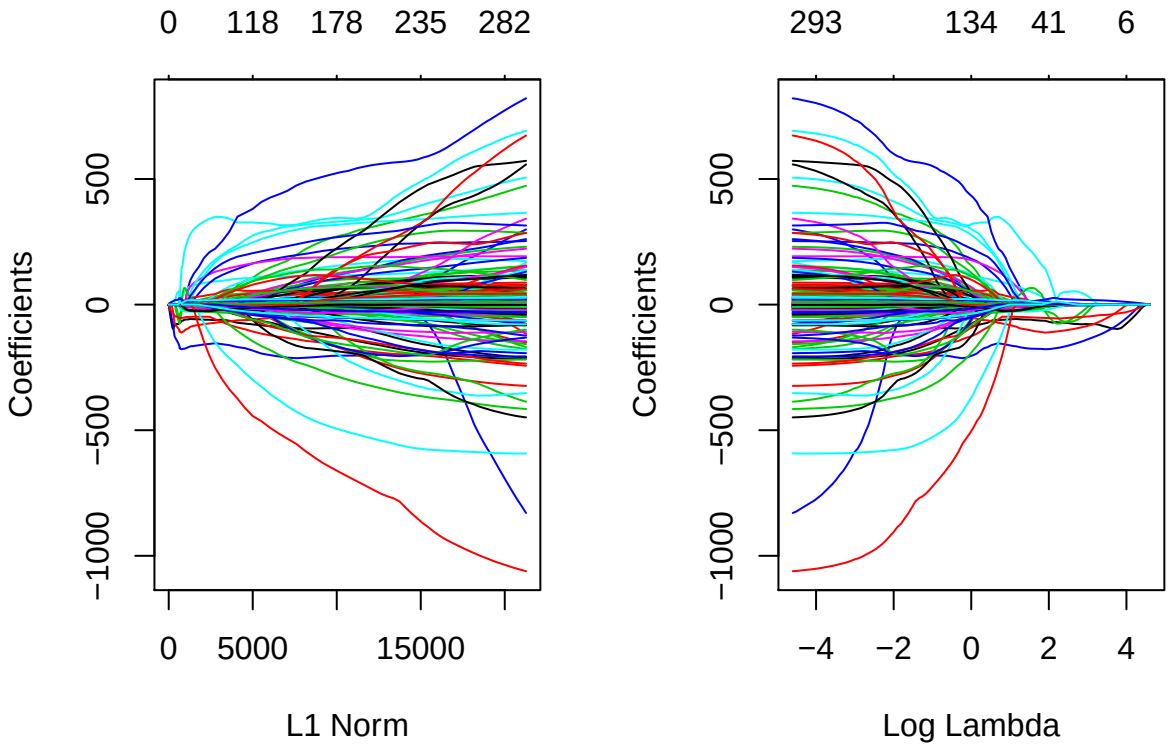
```
## 20   17  8.89  17.140
## 21   18  9.04  15.620
## 22   20  9.19  14.230
## 23   24  9.33  12.960
## 24   25  9.45  11.810
## 25   25  9.55  10.760
## 26   28  9.65   9.807
## 27   34  9.87   8.936
## 28   40 10.15   8.142
## 29   39 10.48   7.419
## 30   41 10.75   6.760
## 31   42 10.98   6.159
## 32   41 11.15   5.612
## 33   46 11.29   5.113
## 34   48 11.42   4.659
## 35   49 11.52   4.245
## 36   52 11.61   3.868
## 37   58 11.69   3.525
## 38   65 11.76   3.211
## 39   67 11.85   2.926
## 40   70 11.92   2.666
## 41   71 12.00   2.429
## 42   77 12.07   2.213
## 43   83 12.13   2.017
## 44   90 12.23   1.838
## 45   97 12.32   1.674
## 46  102 12.41   1.526
## 47  105 12.47   1.390
## 48  112 12.53   1.267
## 49  118 12.59   1.154
## 50  126 12.63   1.052
## 51  134 12.66   0.958
## 52  136 12.70   0.873
## 53  142 12.75   0.796
## 54  150 12.78   0.725
## 55  158 12.82   0.660
## 56  169 12.85   0.602
## 57  176 12.88   0.548
## 58  178 12.91   0.500
## 59  184 12.93   0.455
## 60  189 12.96   0.415
## 61  196 12.98   0.378
## 62  198 13.00   0.344
## 63  201 13.02   0.314
## 64  207 13.04   0.286
## 65  207 13.05   0.260
## 66  216 13.06   0.237
## 67  218 13.07   0.216
## 68  224 13.09   0.197
## 69  231 13.09   0.180
## 70  235 13.10   0.164
## 71  240 13.11   0.149
## 72  245 13.12   0.136
## 73  248 13.13   0.124
```

```
## 74  249 13.14   0.113
## 75  251 13.15   0.103
## 76  255 13.15   0.094
## 77  258 13.16   0.085
## 78  261 13.17   0.078
## 79  265 13.17   0.071
## 80  267 13.17   0.065
## 81  272 13.18   0.059
## 82  274 13.18   0.054
## 83  275 13.18   0.049
## 84  277 13.19   0.044
## 85  277 13.19   0.041
## 86  279 13.19   0.037
## 87  280 13.19   0.034
## 88  282 13.20   0.031
## 89  284 13.20   0.028
## 90  286 13.20   0.025
## 91  289 13.20   0.023
## 92  291 13.20   0.021
## 93  291 13.20   0.019
## 94  293 13.20   0.018
## 95  294 13.20   0.016
## 96  294 13.20   0.015
## 97  294 13.21   0.013
## 98  296 13.21   0.012
## 99  298 13.21   0.011
## 100 298 13.21   0.010
```

```
# Limit the model to num_vars number of variables
num_vars = 20
# choose lowest lambda iteration that results in num_vars non-zero variables
i = max(which(abs(reg_1c$df - num_vars) == min(abs(reg_1c$df - num_vars))))
lambda = reg_1c$lambda[[i]]
print(paste0('# of variables in ', i, 'th iteration: ', sum(reg_1c$beta[, i] != 0)))
```

```
## [1] "# of variables in 22th iteration: 20"
```

```
# print the i'th lasso regression coefficients
# print(reg_1c$beta[, i])

# Plot what the coefficients are doing as we increase lambda
op <- par(mfrow=c(1,2))
plot(reg_1c, "norm", label=TRUE)
plot(reg_1c, "lambda", label=TRUE)
```




```
par(op)
```

In both graphs, each curve corresponds to a variable. They show the path of its coefficient against (in the left plot) the L1-norm of the whole coefficient vector as lambda varies and (in the right plot) the values of log-lamba. The top axis indicates the number of nonzero coefficients at the current lambda, which is the effective degrees of freedom (df) for the lasso. We see clearly that as lambda increases more coefficients goes to zero, as intended in model selection.

Here are the non-zero coefficients:

```
knitr::kable(reg_1c$beta[, i][reg_1c$beta[, i] != 0],
             caption=paste0("Lasso Regression for top ", num_vars, " vars (lambda = ", lambda, ")"),
             col.names = 'Non-zero Coefficients', align = "l", digits = 3)
```

Table 2: Lasso Regression for top 20 vars (lambda = 14.2285096201913)

	Non-zero Coefficients
tobacco1	-76.896
csex2	-96.432
mrace32	-65.449
mrace33	-156.481
preterm2	53.255
ormoth2	-50.644
dmar2	-51.005
I(disllb^2)	0.000
tobacco1:dmage	-3.770

	Non-zero Coefficients
tobacco1:dfage	-0.083
mrace32:dfage	-0.304
mrace32:dmeduc	-0.001
preterm2:dmage	0.254
preterm2:dmeduc	3.175
preterm2:nprevist	20.146
dmage:dislb	-0.003
dislb:dtotord	-0.007
dtotord:adequacy2	4.431
adequacy2:nprevist	0.661
adequacy3:nprevist	1.996

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

Part (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?

```
# create the propensity score using logit
# using all of the "predetermined" covariates
df2a = df %>% select(tobacco, csex, mrace3, preterm,
                     dmage, dfage, dmeduc, dfeduc, ormoth, orfath,
                     dislb, dtotord, dmar, adequacy, nprevist)
reg_2a1 <- glm(tobacco ~ ., data = df2a, family = "binomial")
```

```
stargazer(reg_2a1, title="First logit regression", header=FALSE, single.row=TRUE, type=table_type)
```

Removing sex of the child (csex) and prenatal adequacy (adequacy) from regressors (not significant).

```
# Try logit with only the significant covariates
df2a2 = df2a %>% select(-csex, -adequacy)
reg_2a2 <- glm(tobacco ~ ., data = df2a2, family = "binomial")
```

```
stargazer(reg_2a2, title="Second logit regression", header=FALSE, single.row=TRUE, type=table_type)
```

```
labs <- paste("Actual smoking status:", c("Yes", "No"))
p1_df <- data.frame(p1_score = predict(reg_2a1, type = "response"), tobacco = reg_2a1$model$tobacco) %>%
  mutate(tobacco = ifelse(tobacco == 1, labs[1], labs[2]))
p2_df <- data.frame(p2_score = predict(reg_2a2, type = "response"), tobacco = reg_2a2$model$tobacco) %>%
  mutate(tobacco = ifelse(tobacco == 1, labs[1], labs[2]))

ggplot(p1_df, aes(x=p1_score, fill = tobacco)) +
  geom_histogram(position = "identity", alpha = 0.5,
                 mapping = aes(y = stat(density))) +
```

Table 3: First logit regression

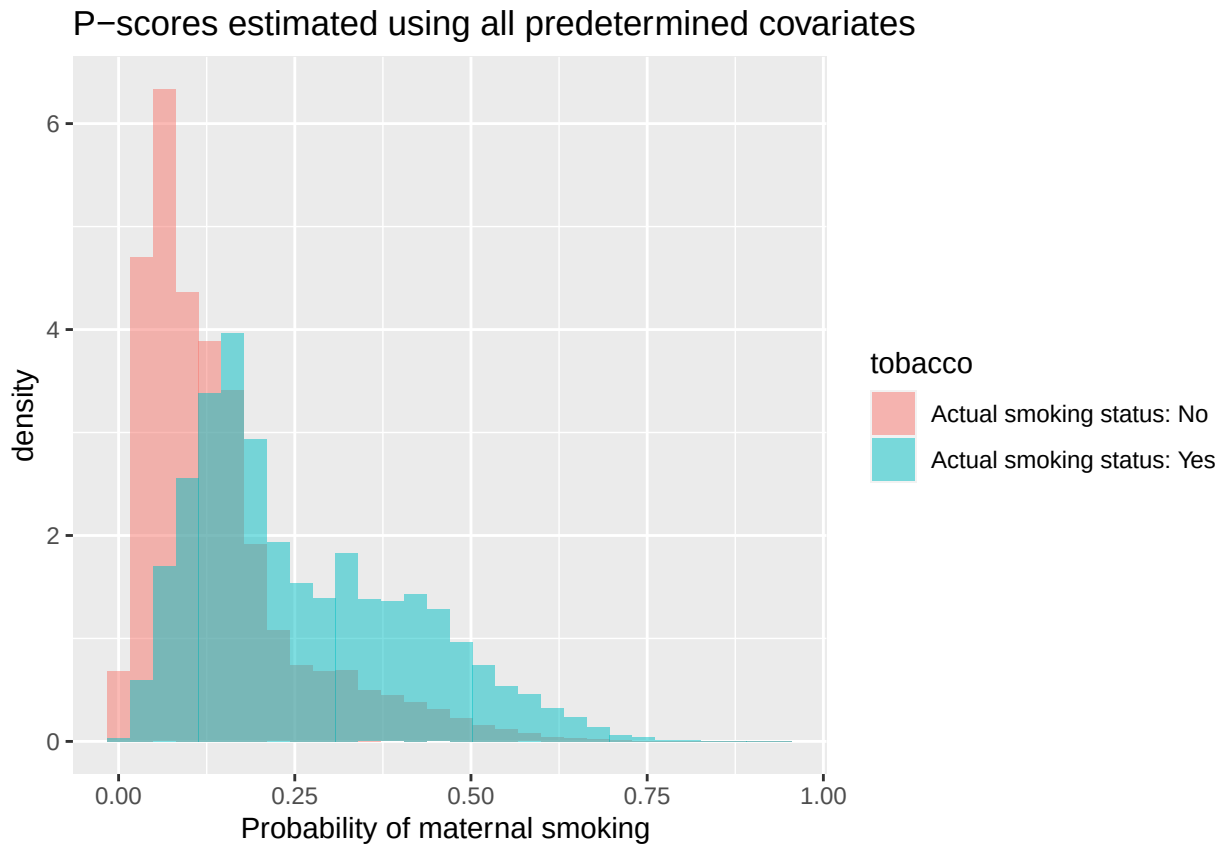
	Dependent variable:
	tobacco
csex2	-0.019 (0.017)
mrace32	-1.883^{***} (0.133)
mrace33	-0.979^{***} (0.028)
preterm2	-0.409^{***} (0.062)
dmage	-0.020^{***} (0.003)
dfage	0.027^{***} (0.002)
dmeduc	-0.178^{***} (0.006)
dfeduc	-0.122^{***} (0.005)
ormoth1	-2.082^{***} (0.292)
ormoth2	-1.225^{***} (0.083)
ormoth3	-1.390^{**} (0.606)
ormoth4	-1.724^{***} (0.304)
ormoth5	-0.891^{***} (0.143)
orfath1	-0.921^{***} (0.189)
orfath2	-0.530^{***} (0.076)
orfath3	-0.213 (0.372)
orfath4	-1.011^{***} (0.216)
orfath5	-0.312^{**} (0.140)
disllb	-0.0003^{***} (0.00003)
dtotord	0.104^{***} (0.007)
dmar2	1.287^{***} (0.022)
adequacy2	-0.024 (0.025)
adequacy3	0.003 (0.046)
nprevist	-0.009^{***} (0.003)
Constant	2.035^{***} (0.105)
Observations	114,610
Log Likelihood	$-43{,}745.970$
Akaike Inf. Crit.	87,541.940
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Second logit regression

	Dependent variable:
	tobacco
mrace32	-1.884^{***} (0.133)
mrace33	-0.978^{***} (0.028)
preterm2	-0.409^{***} (0.062)
dmage	-0.020^{***} (0.003)
dfage	0.027^{***} (0.002)
dmeduc	-0.178^{***} (0.006)
dfeduc	-0.122^{***} (0.005)
ormoth1	-2.083^{***} (0.292)
ormoth2	-1.225^{***} (0.083)
ormoth3	-1.390^{**} (0.606)
ormoth4	-1.724^{***} (0.304)
ormoth5	-0.891^{***} (0.143)
orfath1	-0.921^{***} (0.189)
orfath2	-0.530^{***} (0.076)
orfath3	-0.213 (0.372)
orfath4	-1.013^{***} (0.216)
orfath5	-0.313^{**} (0.140)
disllb	-0.0003^{***} (0.00003)
dtotord	0.104^{***} (0.007)
dmar2	1.286^{***} (0.022)
nprevist	-0.008^{***} (0.002)
Constant	2.006^{***} (0.097)
Observations	114,610
Log Likelihood	$-43{,}747.220$
Akaike Inf. Crit.	87,538.450
Note:	*p<0.1; **p<0.05; ***p<0.01

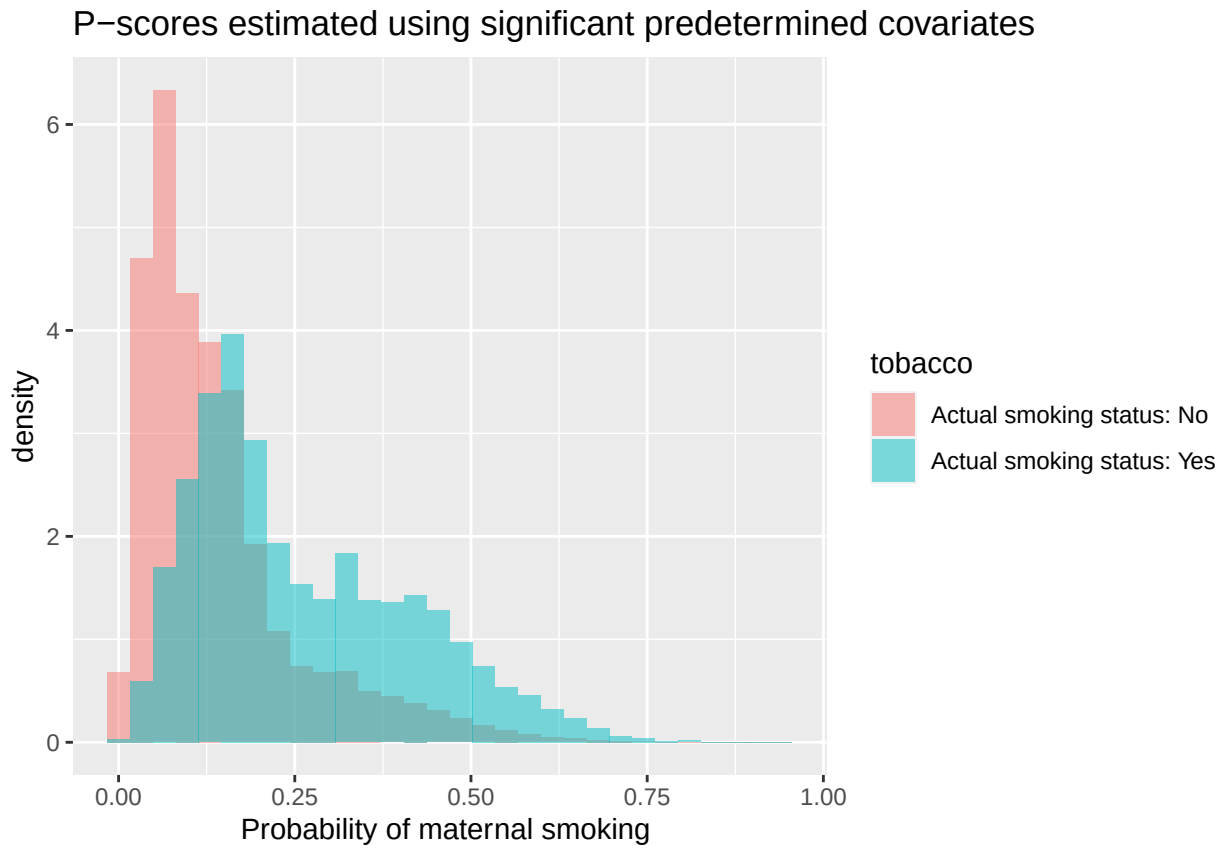
```
  xlab("Probability of maternal smoking ") +
  ggtitle("P-scores estimated using all predetermined covariates")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggplot(p1_df, aes(x=p1_score, fill = tobacco)) +
  geom_histogram(position = "identity", alpha = 0.5,
                 mapping = aes(y = stat(density))) +
  xlab("Probability of maternal smoking ") +
  ggtitle("P-scores estimated using significant predetermined covariates")
```

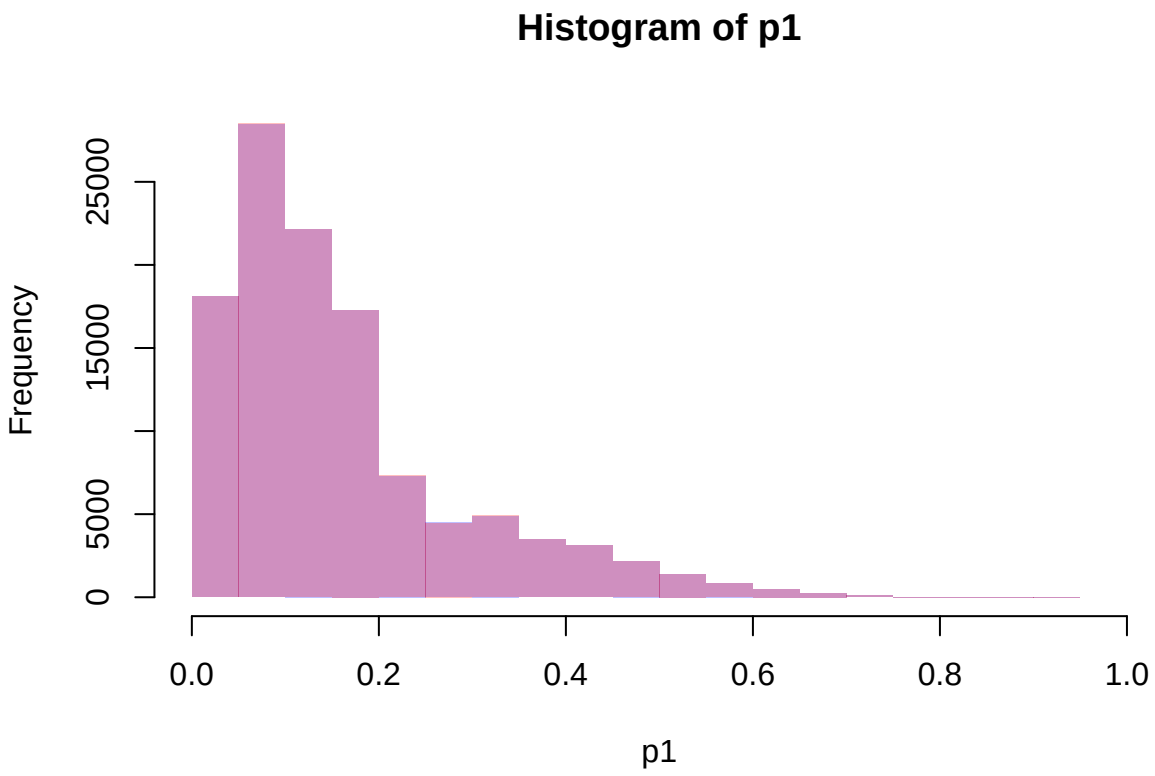
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# Compare histograms of p-scores
p1 = reg_2a1 %>% predict(df2a, type = "response")
p2 = reg_2a2 %>% predict(df2a2, type = "response")
```

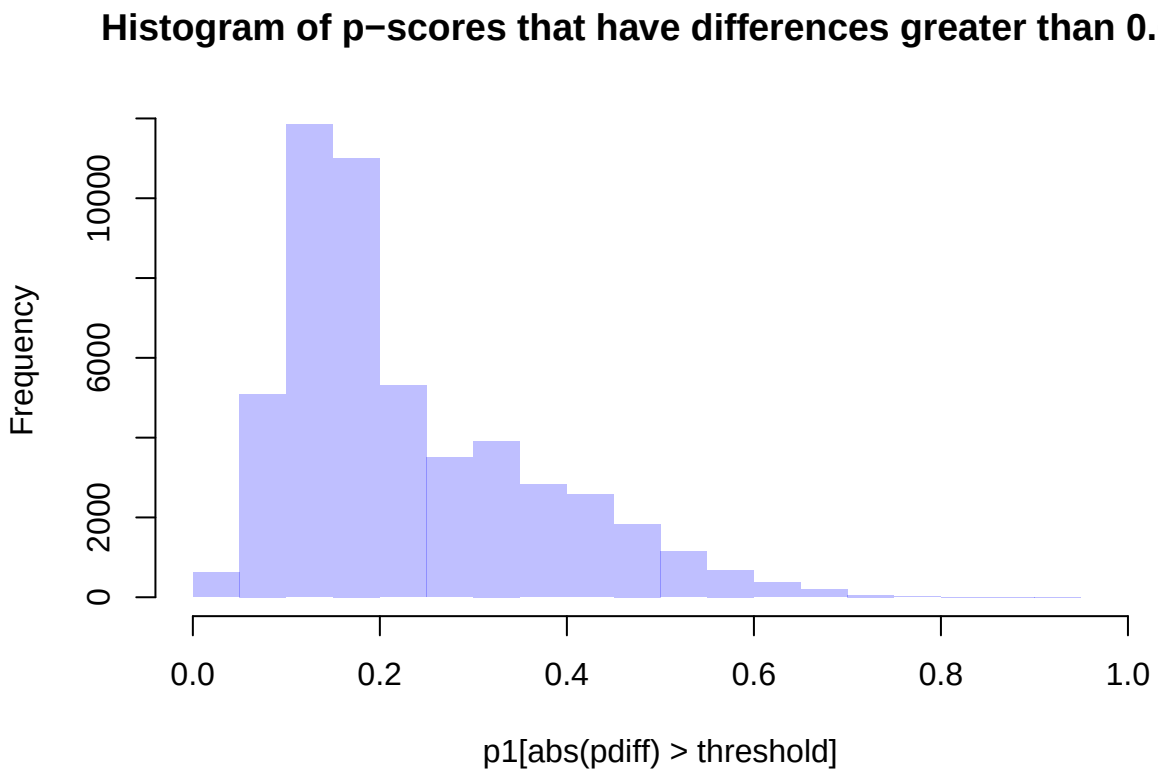
The histogram of the first predicted p-score is in orange, the second is in purple. So this fully-pink histogram is meant to show that there is nearly complete overlap in this histograms of the predicted p-scores.

```
plot(hist(p1, plot=F), col=rgb(0,0,1,1/4), border=NA, xlim=c(0,1))  # first histogram
plot(hist(p2, plot=F), col=rgb(1,0,0,1/4), border=NA, xlim=c(0,1), add=T)  # second
```

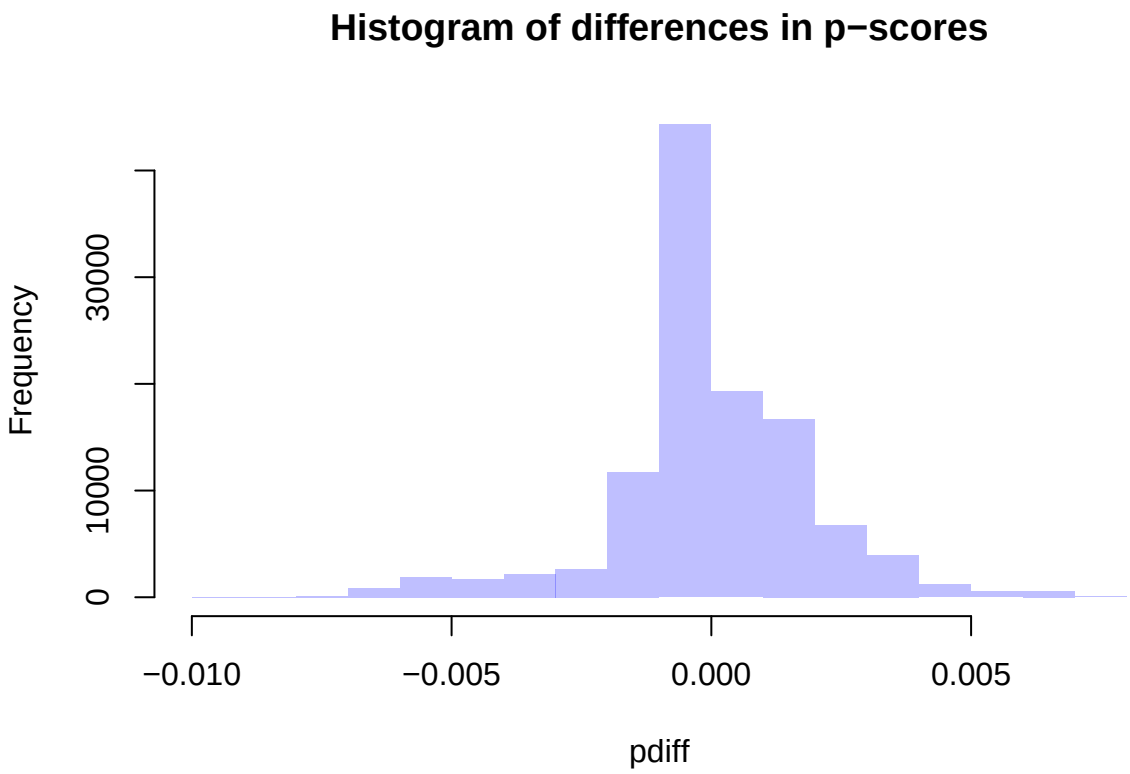


We can see where the differences are that are more than 0.001 and the shape of the differences.

```
threshold = 0.001
pdiff = p1-p2
plot(hist(p1[abs(pdiff) > threshold], plot=F), col=rgb(0,0,1,1/4), border=NA, xlim=c(0,1),
     main=paste('Histogram of p-scores that have differences greater than', threshold))
```

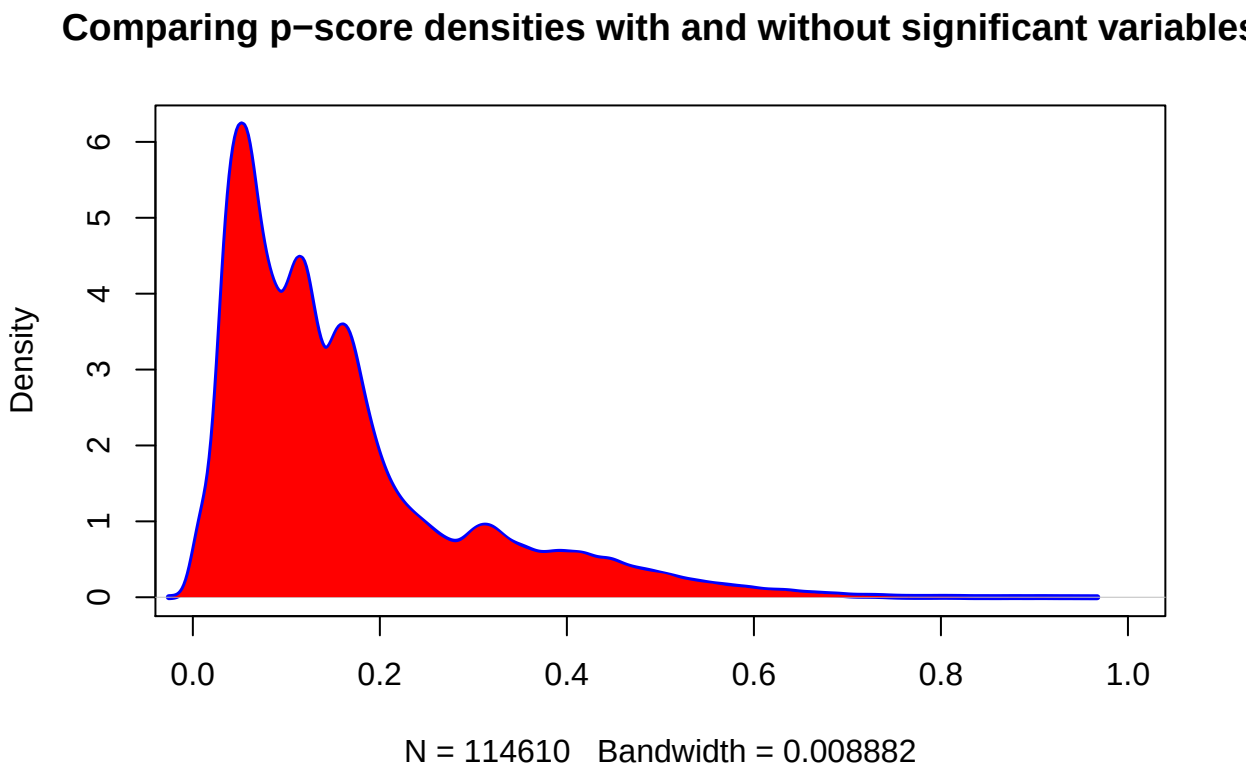



```
plot(hist(pdiff, plot=F), col=rgb(0,0,1,1/4), border=NA, main='Histogram of differences in p-scores')
```



Here's two density plots to reinforce the idea. The blue line is the density plot for the first p-score and the red filling is the density plot for the second.

```
plot(density(p1), col=rgb(0,0,1), xlim=c(0,1), lwd=3,
     main='Comparing p-score densities with and without significant variables')  # first kernel
polygon(density(p2), col="red", border=NA, xlim=c(0,1))  # second kernel filled in
```



Note that the maximum difference between any two predicted p-scores is 0.0095441.

Overall, the propensity scores are very similar. However, this does not imply that we have the "correct" set of covariates in the logit specification used for our propensity score. It merely shows that excluding the non-significant variables from the propensity score estimation won't affect the propensity score estimates. This doesn't suggest that we aren't missing any potentially important covariates.

Part (b)

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.

```
# Control for p-score in regression analysis
df2b = df %>%
  select(dbrwt, tobacco, csex, mrace3, preterm,
         dmage, dfage, dmeduc, dfeduc, ormoth, orfath,
         disllb, dtotord, dmar, adequacy, nprevist) %>%
  mutate(pscore = p1)
reg_2b = lm(dbrwt ~ ., data=df2b)

# Estimate ATE
ATE_2b = reg_2b$coefficients["tobacco1"]

reg_2bb = lm(dbrwt ~ tobacco + pscore, data=df2b)
ATE_2bb = reg_2bb$coefficients["tobacco1"]
```

The estimated ATE of tobacco use during pregnancy when including covariates and the pscore is -222.402

– smoking during pregnancy causes a 222.402 drop in grams of the birthweight of the child. When only including the pscore, the ATE is -223.145. This is correct if we have unconfoundedness and we assume a constant treatment effect.

Part (c)

As discussed in class, one can use the estimated propensity scores to reweight the outcomes of non-smokers and estimate the average treatment effect. Compute an estimate of the average treatment effect and the "effect of the treatment on the treated" by appropriate reweighting of the data.

Imbens tells us that, after dividing by the sum of the weights in each treatment group, the feasible propensity-score-weighted ATE is:

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

```
# Reweight data using p-score to weight
df2c = df2b %>%
  mutate(pscore = predict(reg_2a1, ., type = "response")) %>%
  mutate(weight = ifelse(tobacco == 1, 1/pscore, 1/(1-pscore)))
t_norm1 = sum((df2c$tobacco==1)*df2c$weight)  # smokers
c_norm1 = sum((df2c$tobacco==2)*df2c$weight)  # non-smokers
df2c = df2c %>% mutate(weight3 = ifelse(tobacco == 1, 1/pscore/t_norm1, 1/(1-pscore)/c_norm1))

# Estimate ATE
weight_mean_smoker_all = sum((df2c$tobacco==1) * df2c$weight * df2c$dbrwt) / t_norm1
weight_mean_nonsmoker_all = sum((df2c$tobacco==2) * df2c$weight * df2c$dbrwt) / c_norm1
ATE2c = weight_mean_smoker_all - weight_mean_nonsmoker_all
ATE2c
```

```
## [1] -222.0921
```

The propensity-score-weighted estimated average treatment effect is -222.09.

Imbens (2004) in a review article states that the efficient TOT estimator of $\tau_{treated}$ is derived by weighting each observation by it's propensity score. This makes intuitive sense because we want to weight-up the observations that look like the treated observations and weight-down the observations that look like the untreated. After canceling out the propensity score in the first term and dividing by the sum of the weights for each treatment group, we get:

$$\hat{\tau}_{ATT} = \frac{1}{N_T} \sum_{i:D_i=1} Y_i - \frac{\sum_{i:D_i=0} Y_i \frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}}{\sum_{i:D_i=0} \frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}} = \frac{\sum_{i=1}^{N} D_i Y_i}{\sum_{i=1}^{N} D_i} - \frac{\sum_{i=1}^{N} \frac{(1-D_i)\hat{p}(X_i)Y_i}{1-\hat{p}(X_i)}}{\sum_{i=1}^{N} \frac{(1-D_i)\hat{p}(X_i)}{1-\hat{p}(X_i)}}$$

```
# Reweight data for ATT
df2c = df2c %>%
  mutate(weight2 = ifelse(tobacco == 1, 1, pscore/(1-pscore)))
t_norm2 = sum((df2c$tobacco==1)*df2c$weight2)  # smokers
c_norm2 = sum((df2c$tobacco==2)*df2c$weight2)  # non-smokers

# Estimate ATT
weight_mean_smoker_treat = sum((df2c$tobacco==1) * df2c$weight2 * df2c$dbrwt) / t_norm2
weight_mean_nonsmoker_treat = sum((df2c$tobacco==2) * df2c$weight2 * df2c$dbrwt) / c_norm2
```

```
ATT2c = weight_mean_smoker_treat - weight_mean_nonsmoker_treat
ATT2c
```

```
## [1] -225.272
```

The propensity-score-weighted estimated average treatment effect on the treated is -225.27. If we take this result as significantly different from the ATE, then the effect of smoking on birthweight is stronger for those who look like smokers than it is in the general population. This seems to suggest that there are other, possibly unobserved, characteristics that decrease birthweight that are more common in the population that look like non-smokers than in population that looks like smokers.

Part (d)

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.

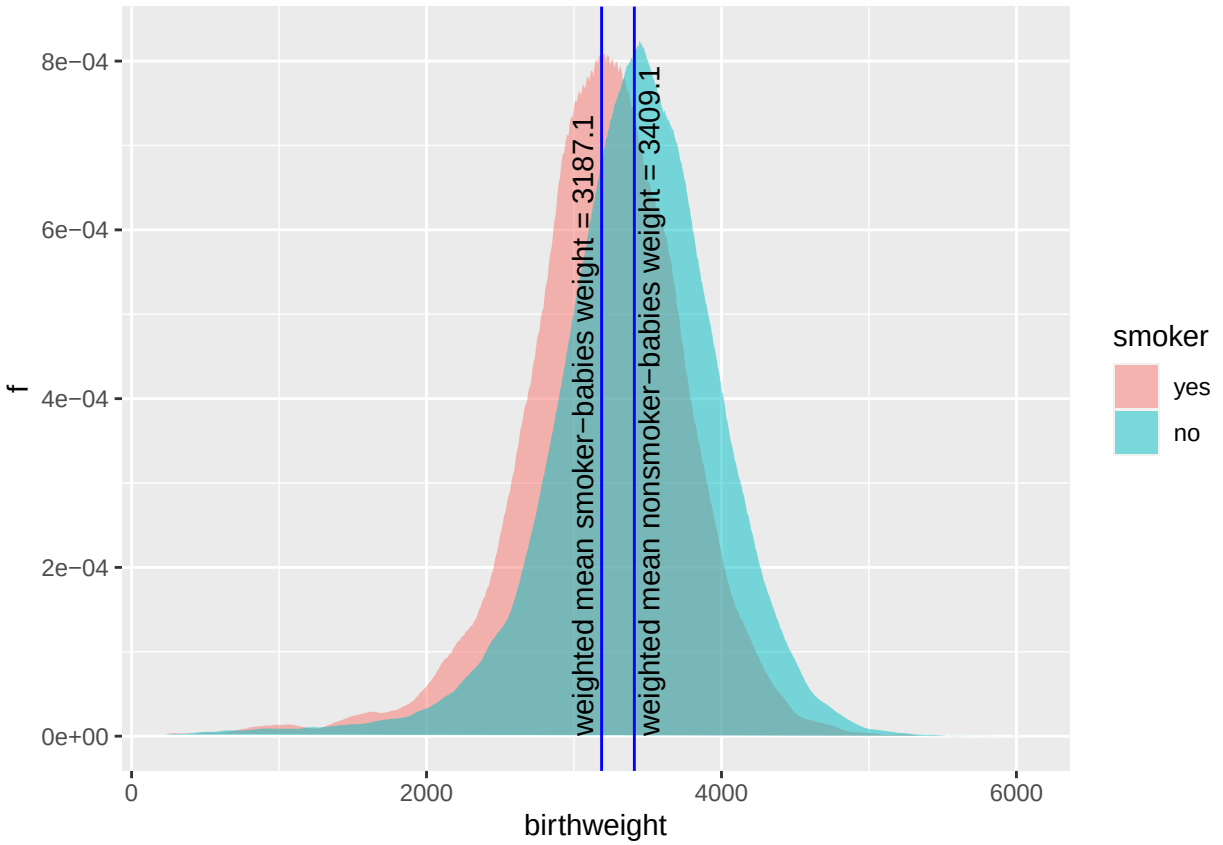
Morgan and Todd (2008) gives us the weights for ATT and ATC (average treatment effect on the controls). These come from weighting observations by $1 - p(X_i)$ instead of $p(X_i)$ to weight toward the control group:

$$\begin{aligned}
&\text{For } D_i = 1: \quad w_{i,ATT} = 1 \\
&\text{For } D_i = 0: \quad w_{i,ATT} = \frac{\hat{p}(X_i)}{1 - \hat{p}(X_i)} \\
&\qquad \text{and} \\
&\text{For } D_i = 1: \quad w_{i,ATC} = \frac{1 - \hat{p}(X_i)}{\hat{p}(X_i)} \\
&\text{For } D_i = 0: \quad w_{i,ATC} = 1
\end{aligned}$$

We can use weighted kernel density estimation with the weights above:

```
# Estimate the counterfactual birthweight densities with a kernel density estimator
# See Joel's notes for kernel density estimator
# Play around with a bandwidth starting with half the default Stata bandwidth
# For stata bandwidht, see rkdensity.pdf page 9 in this ps1b github folder.
# You can also run on stata with no bandwidth specified, then print the
# default bandwidth used using -display r(bwidth)-
# Choose the same bandwidth for all the pictures

# Stata default bandwidth
m = min(var(df2c$dbrwt), IQR(df2c$dbrwt)/1.349)
h_stata = 0.9*m/nrow(df2c)^(1/5)  # 44.14637 grams
df2d = df2c

# Create a vector of birthwieght values for plotting the density
x_vec = seq(min(df2d$dbrwt), max(df2d$dbrwt), length.out=1000)
```

```r
# Define the kernel
kernel_fun <- function(x, h){ # Epanechnikov Kernel with bandwidth h
  weight = (3/4)*(1-(x/h)^2)*as.numeric(abs(x/h)<1)
  return(weight)
}

# Define the kernel density
density_fun <- function(x, df, h) { # Kernel density estimate with data X and weights w
  withCallingHandlers({
    # h = h_stata*2
    n = nrow(df)
    df = df %>% filter(between(dbrwt, x-h, x+h))
    if (nrow(df) == 0) {return(0)}
    kernel_vec = sapply(x - df$dbrwt, kernel_fun, h=h)
    f = sum(df$weight * kernel_vec) / (n*h) #
    return(f)
  }, warning=function(w) {
      if (startsWith(conditionMessage(w), "between() called on numeric"))
          invokeRestart("muffleWarning")
  })
}

# Calculate starting bandwidths (based on Stata formula)
m1 = min(var(filter(df2d, tobacco==1)$dbrwt), IQR(filter(df2d, tobacco==1)$dbrwt)/1.349)
h1 = 0.9*m/nrow(filter(df2d, tobacco==1))^(1/5)
m2 = min(var(filter(df2d, tobacco==2)$dbrwt), IQR(filter(df2d, tobacco==2)$dbrwt)/1.349)
h2 = 0.9*m/nrow(filter(df2d, tobacco==2))^(1/5)
# estimate counterfactual densities (takes about 16 seconds for both over 1000 point x_vec)
dens1_custom = sapply(x_vec, density_fun, df=filter(df2d, tobacco==1), h=2*h1)
dens2_custom = sapply(x_vec, density_fun, df=filter(df2d, tobacco==2), h=2*h2)

# Caclulate area under the curve to use as normalization constants for the PDFs
norm_cons1 = sum(diff(x_vec) * zoo::rollmean(dens1_custom, 2))
norm_cons2 = sum(diff(x_vec) * zoo::rollmean(dens2_custom, 2))

# Use normalization constants to scale up to true PDFs
dens_df = rbind(data.frame(birthweight=x_vec, f=(dens1_custom / norm_cons1), smoker="yes"),
                data.frame(birthweight=x_vec, f=(dens2_custom / norm_cons2), smoker="no"))

# Plot
mean_lines = data.frame(avg = c(weight_mean_smoker_all, weight_mean_nonsmoker_all),
    name = c(paste("weighted mean smoker-babies weight =", round(weight_mean_smoker_all, 1)),
             paste("weighted mean nonsmoker-babies weight =", round(weight_mean_nonsmoker_all, 1))))
p <- ggplot(dens_df, aes(x = birthweight, y = f)) +
  geom_polygon(aes(fill=smoker, group = smoker), alpha = 0.5) +
  geom_vline(data=mean_lines, mapping=aes(xintercept=avg), color="blue") +
  geom_text(data=mean_lines, mapping=aes(x=avg, y=0, label=name), size=4, angle=90, vjust=c(-0.4, 1.2),
p
```



Calculate the kernel estimator at birthweight equals 3,000 grams:

```
# for smokers
density_fun(3000, filter(df2d, tobacco==1), 2*h1) / norm_cons1
```

```
## [1] 0.0007465044
```

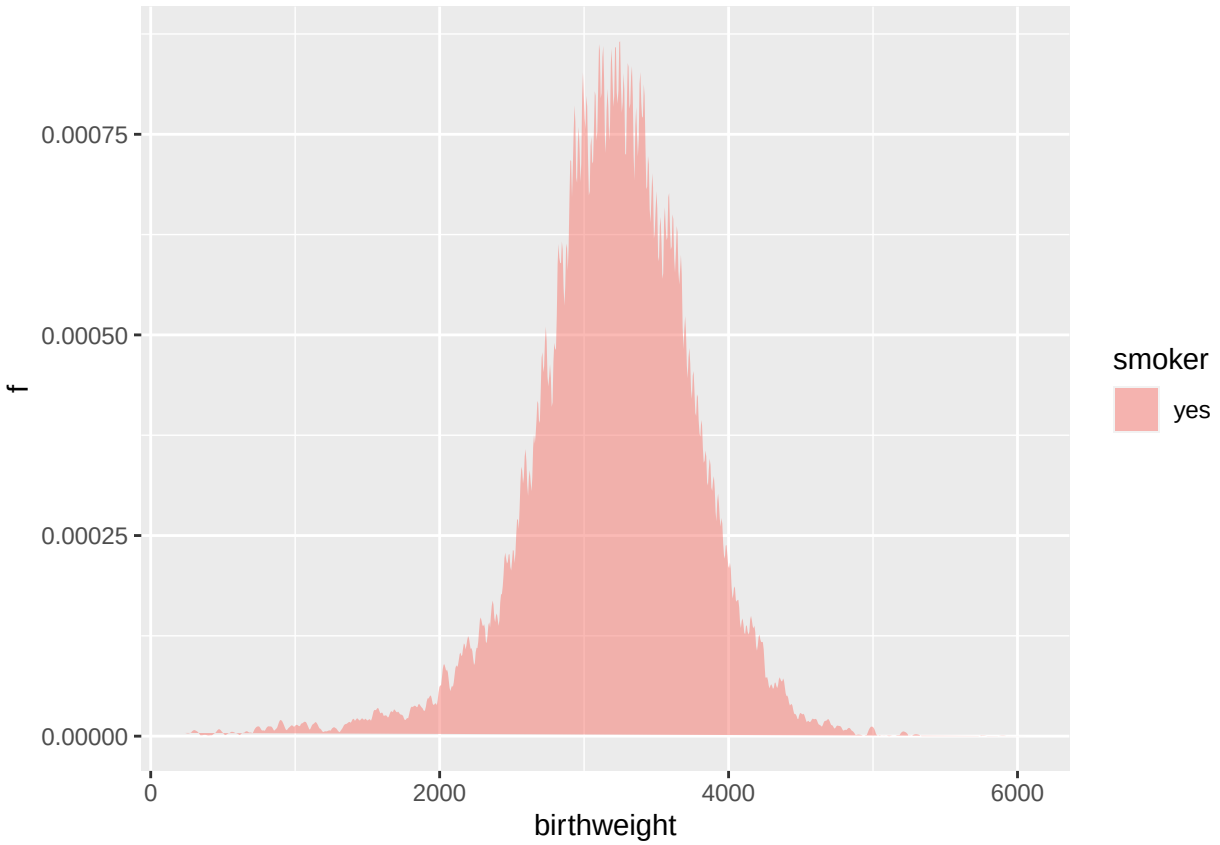
```
# for nonsmokers
density_fun(3000, filter(df2d, tobacco==2), 2*h2) / norm_cons1
```

```
## [1] 0.0001008476
```

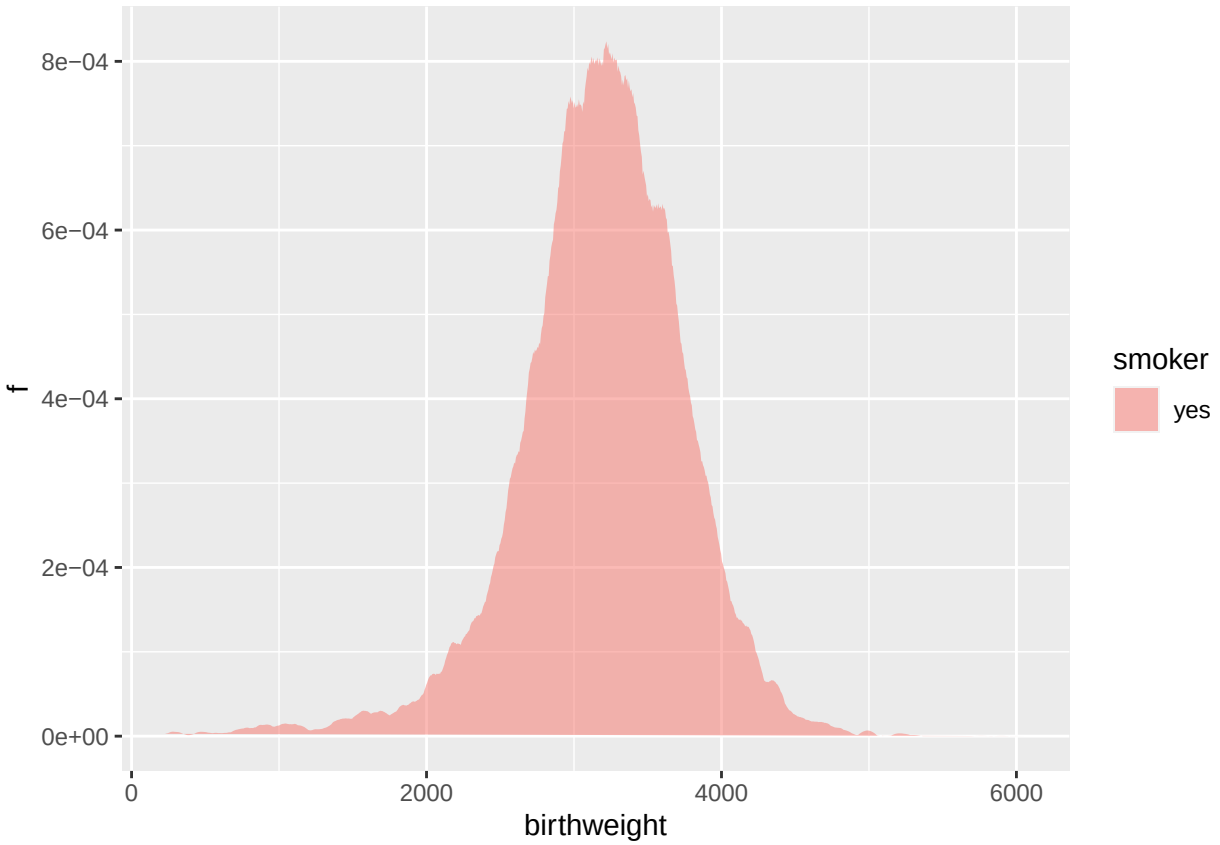
Part (e)

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?

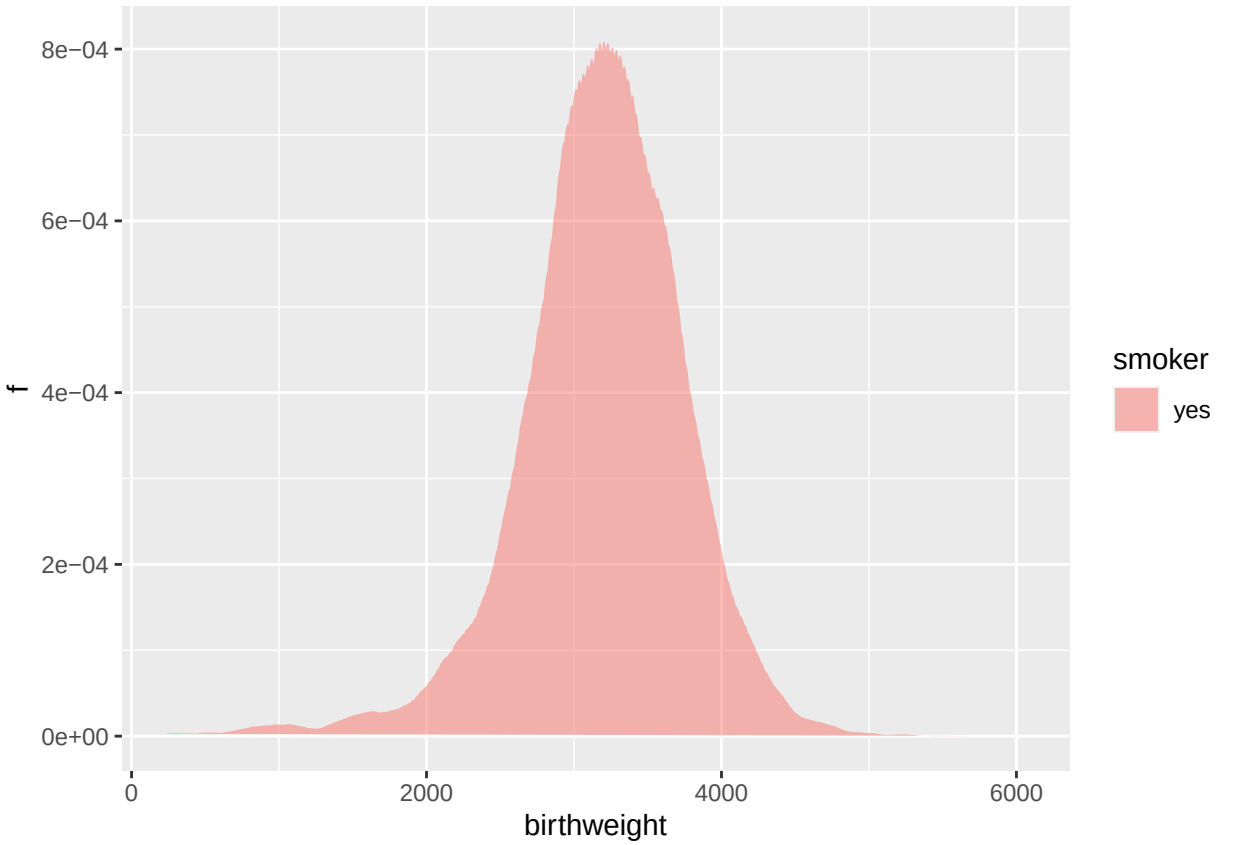
Smokers' density with bandwidth = 1/2 * Stata bandwidth



Smokers' density with bandwidth = 1 Stata bandwidth



Smokers' density with bandwidth = 2 Stata bandwidth



As we increase the bandwidth, the density esimtate gets smoother and smoother.

Part (f)

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.

Weighting on the propensity score, we can reduce the dimensionality of the covariate space to compare observations that have similar probabilities of selection into the treatment group. If we have good covariate overlap, this means we can estimate treatment effects based on comparisons of control and treatment observations that "look alike."

If we don't have good covariate overlap, however, we may have very few observations that are matched to the opposite treatment group. We expect the majority of the control observations to have a low propensity score (since it's estimating the probability of treatment based on covariates) and the majority of treatment observations to have high propensity scores. Without good covariate overlap, we would expect very few treatment observations to have very low p-scores (near 0) and very few control observations to have high p-scores (near 1).

Since our weights are $1/p(X)$ for treated observations, the weights for the rare treatment observations with p-scores near 0 can be extremely large and thus we might be relying on relatively few observations to estimate the treatment effect. Similarly, the weights are $1/(1-p(X))$ for control observations and we could be heavily weighting few observations with p-scores near 1. If we trust the outlier untreated observations, this may be reasonable, however the fact that they are outliers among the untreated is also a worrying sign such outliers may be unrepresentative as counterfactual controls.

Part (g)

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:

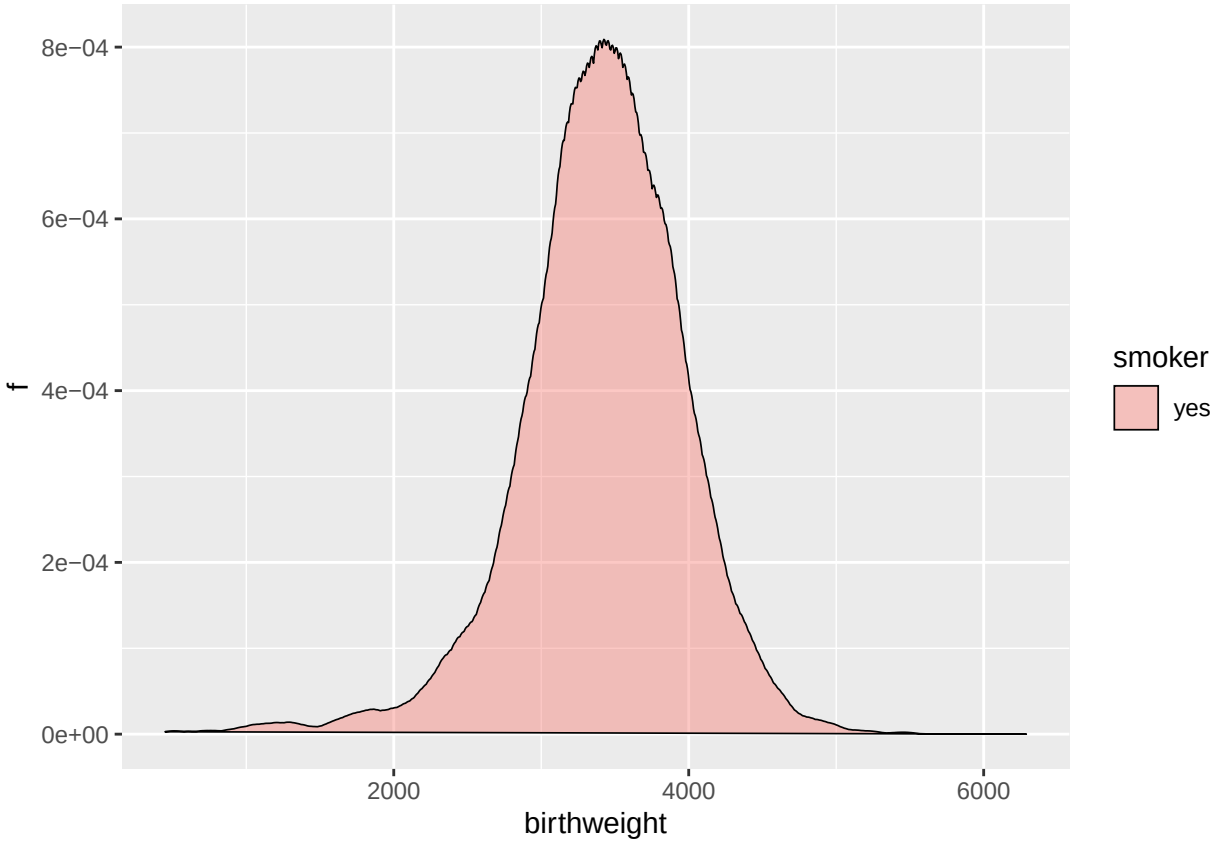
(i) The treatment effect heterogeneity is linear in the propensity score.
(ii) The treatment effect heterogeneity is not linear in the propensity score.
(iii) The decision to smoke is completely randomly assigned.
(iv) Conditional on the exogenous variables the decision to smoke is randomly assigned.

For part (b) where we estimate the ATE using a regression on the treatment and the pscore, we estimate an effect of -223.1 – that is, on average, we expect that smoking during pregnancy decreases the birthweight of the child by 223.1 grams for the general population (assuming our sample is representative of the general population). This requires that condition (iv) holds, and additionally, that the treatment effect is homogeneous since we are regressing on just the treatment indicator and the pscore.

For part (c), we use the pscore to weight the observations to create balanced set of covariates and estimate the average treatment effect to be -222.1 – on average, we expect that smoking during pregnancy decreases the birthweight of the child by 222.1 grams for the general population. We also estimate the average effect on the treated to be -225.3 – on average for women that smoke, we expect that smoking during pregnancy decreases the birthweight of the child by 222.1 grams. These results both require conditions (i) and (iv) to be valid.

We can assess the homogeneous treatment effect assumption by comparing the shape of the counterfactual distributions. Below, we have shifted the smokers' distribution up by the ATE estimated in part (c). We can see that the distributions are extremely close, giving evidence that the treatment effect is fairly homogenous in across the pscore values.

```
dens_df2 = dens_df %>%
  mutate(birthweight = ifelse(smoker=='yes', birthweight - ATE2c, birthweight))

ggplot(dens_df2, aes(x = birthweight, y = f)) +
  geom_polygon(aes(fill=smoker), alpha = 0.4, color='black', size=0.3)
```



Problem 3

A potentially more informative way to describe how birth weight affects smoking is to estimate the "non-parametric" conditional mean of birth weight as a function of the estimated probability of smoking, separately for smokers and non-smokers on the same graph. To do so, divide the data from smokers into 100 approximately equally spaced bins based on the estimated propensity score. Do the same for nonsmokers. Use the blocking estimator we discussed in class. Interpret your findings and relate them to the results in (2b).

```
#Prepping dataset arranged by pscore
df3 <- df %>%
  select(dbrwt, tobacco) %>%
  mutate(pscore = p1, smoke = ifelse(tobacco == 1, 1,0)) %>%
  select(-tobacco) %>%
  arrange(pscore)

# Check Overlap Assumption
minmax <- df3 %>%
            group_by(smoke) %>%
            summarise(max = max(pscore), min = min(pscore))

#Creating bins separately for smokers and non-smokers
df3_smoke <- df3 %>%
  filter(smoke == 1) %>%
  mutate(pscore_bin = cut(pscore, breaks = 100, labels = seq(1,100)))
```

```
df3_nosmoke <- df3 %>%
  filter(smoke == 0) %>%
  mutate(pscore_bin = cut(pscore, breaks = 100, labels = seq(1,100)))

#Generating t_k (treatment effects for bins)
df3_smoke_mean <- df3_smoke %>%
  group_by(pscore_bin) %>%
  summarise(bin_mean_smoke = mean(dbrwt), N_smoke = n())

df3_nosmoke_mean <- df3_nosmoke %>%
  group_by(pscore_bin) %>%
  summarise(bin_mean_nosmoke = mean(dbrwt), N_nosmoke = n())

#Generating t (weighted total treatment effect)
df3_merged <- full_join(df3_nosmoke_mean, df3_smoke_mean, by = "pscore_bin") %>%
  filter(is.na(bin_mean_nosmoke) == 0 & is.na(bin_mean_smoke) == 0) %>%
  mutate(dbrwt_diff = bin_mean_smoke - bin_mean_nosmoke, N = N_smoke + N_nosmoke)

N_total_3 <- sum(df3_merged$N)

df3_merged <- df3_merged %>%
  mutate(t_k_weighted = dbrwt_diff * (N/N_total_3))

t_3 <- sum(df3_merged$t_k_weighted)
```

First, for each group of smokers and nonsmokers, we see that the range of estimated propesntiy score is (0.001, 0.9405) for nonsmokers and (0.0055, 0.9348) for smokers. The range is very similar, meaning overlap assumption is satisfied, and trimming for blocking is not needed. This also ensures that what we have done above, which is creating 100 equal bins between the minimum and maximum propensity score for each group, will be very similar to creating the same 100 bins for the two groups such that first bin is [0, 0.01] and so on.

Using the blocking estimator, we find a smoking treatment effect of -220.021

Problem 4

Low birth weight births (less than 2500 grams) are considered particularly undesirable since they comprise a large share of infant deaths. Redo question 3 using an indicator for low birth weight birth as the outcome of interest. Interpret your findings.

```
# Prepping dataset with indicator variable
df4 <- df3 %>%
  mutate(low_dbrwt = ifelse(dbrwt < 2500, 1, 0))

# Creating bins separately for smokers and non-smokers
df4_smoke <- df4 %>%
  filter(smoke == 1) %>%
  mutate(pscore_bin = cut(pscore, breaks = 100, labels = seq(1,100)))

df4_nosmoke <- df4 %>%
  filter(smoke == 0) %>%
  mutate(pscore_bin = cut(pscore, breaks = 100, labels = seq(1,100)))

# Generating t_k (treatment effects for bins)
df4_smoke_mean <- df4_smoke %>%
```

```
  group_by(pscore_bin) %>%
  summarise(bin_mean_smoke = mean(low_dbrwt), N_smoke = n())

df4_nosmoke_mean <- df4_nosmoke %>%
  group_by(pscore_bin) %>%
  summarise(bin_mean_nosmoke = mean(low_dbrwt), N_nosmoke = n())

# Generating t (weighted total treatment effect)
df4_merged <- full_join(df4_nosmoke_mean, df4_smoke_mean, by = "pscore_bin") %>%
  filter(is.na(bin_mean_nosmoke) == 0 & is.na(bin_mean_smoke) == 0) %>%
  mutate(low_dbrwt_diff = bin_mean_smoke - bin_mean_nosmoke, N = N_smoke + N_nosmoke)

N_total_4 <- sum(df4_merged$N)

df4_merged <- df4_merged %>%
  mutate(t_k_weighted = low_dbrwt_diff * (N/N_total_4))

t_4 <- sum(df4_merged$t_k_weighted)
```

Using the blocking estimator, we find a smoking treatment effect of 0.037. That is smokers are 3.7 percentage points more likely to have a baby with low birth weight than non-smokers.

Problem 5

Let's link matching back to regression. Consider the conditional expectation function $\mathbb{E}[birthweight \mid X]$, where X contains the following variables: `rectype pldel3 cntocpop stresfip dmage mrace3 dmar adequacy csex dplural`.

Part (a)

Develop a regression that you are confident estimates $\mathbb{E}[birthweight \mid X]$ as $N \to \infty$? Why are you confident that your regression gets the CEF right?

```
# Select variables
df5a = df %>%
          select(dbrwt, rectype, pldel3, cntocpop, stresfip,
                 dmage, mrace3, dmar, adequacy, csex, dplural) %>%
          mutate(stresfip = as.factor(stresfip), dmage = as.factor(dmage))

# Check to make sure all are dummies
dummies <- c("rectype", "pldel3", "cntocpop", "stresfip",
             "dmage", "mrace3", "dmar", "adequacy", "csex", "dplural")

for (var in dummies) {
  print(class(df5a[, var]))
}


# Run saturated regression
reg_5a <- lm(dbrwt ~ rectype + pldel3 + cntocpop + stresfip +
               dmage + mrace3 + dmar + adequacy + csex + dplural +
               rectype*pldel3 + rectype*cntocpop + rectype+stresfip +
               rectype*dmage + rectype*mrace3 + rectype*dmar +
               rectype*adequacy + rectype*csex + rectype*dplural +
```

```
               pldel3*cntocpop + pldel3*stresfip + pldel3*dmage +
               pldel3*mrace3 + pldel3*dmar + pldel3*adequacy +
               pldel3*csex + pldel3*dplural +
               cntocpop*stresfip + cntocpop*dmage + cntocpop*mrace3 +
               cntocpop*dmar + cntocpop*adequacy + cntocpop*csex + cntocpop*dplural +
               stresfip*dmage +  stresfip*mrace3 +  stresfip*dmar +
               stresfip*adequacy +  stresfip*csex +  stresfip*dplural +
               dmage*mrace3 + dmage*dmar + dmage*adequacy +
               dmage*csex + dmage*dplural +
               mrace3*dmar + mrace3*adequacy + mrace3*csex + mrace3*dplural +
               dmar*adequacy + dmar*csex + dmar*dplural +
               adequacy*csex + adequacy*dplural +
               csex*dplural,
             data = df5a)

summary(reg_5a)
```

```
# The below code is just a helpful reference for using local variable names in for loop
make_dummy <- c("rectype")
for (var in make_dummy) {
  name <- paste0(var, "_1")
  df5a <- df5a %>%
    mutate(!!name := ifelse(df5a[,var] == 1, 1 ,0))
}
```

```
# Select variables (Error: cannot allocate vector of size 2616.7 Gb)
#df5a = df %>%
#          select(dbrwt, rectype, pldel3, cntocpop, stresfip,
#                 dmage, mrace3, dmar, adequacy, csex, dplural) %>%
#          mutate(stresfip = as.factor(stresfip), dmage = as.factor(dmage))

# Saturated Regression
#reg_5a_2nd_attempt <- lm(dbrwt ~ (.)^10, data = df5a)
#summary(reg_5a_2nd_attempt)



# Select variables
df5a = df %>%
          select(dbrwt, tobacco, rectype, pldel3, cntocpop, stresfip,
                 dmage, mrace3, dmar, adequacy, csex, dplural) %>%
          mutate(stresfip = as.factor(stresfip), dmage = as.factor(dmage))

# Create factor variable (for dummies needed for regression) for unique combination of covariate values
df5a_unique = df5a %>%
                select(-c(dbrwt, tobacco)) %>%
                distinct() %>%
                arrange(rectype, pldel3, cntocpop, stresfip,
                  dmage, mrace3, dmar, adequacy, csex, dplural) %>%
                mutate(uniq = row_number())
df5a <- full_join(df5a, df5a_unique, by = c("rectype", "pldel3", "cntocpop",
                  "stresfip", "dmage", "mrace3", "dmar", "adequacy", "csex", "dplural"))
df5a <- df5a %>% mutate(uniq = as.factor(uniq))
```

```r
# Run regression on dbrwt with saturated model
reg_5a <- lm(dbrwt ~ uniq, data = df5a)
summary(reg_5a)
```

Saturated model for discrete regressors is a sufficient condition for a linear CEF. Hence, I would create a dummy variable for each unique combination of covariate values and run linear regression on these newly created dummy variables to estimate CEF.

Part (b)

Now run the regression you propose above, but add the treatment (your binary smoking variable) as the righthand side variable of interest. Prove that if the treatment effect of smoking on birthweight is independent of the covariates in X, then exact matching and your regression estimate the same thing. You may assume the conditional independence assumption holds given the variables in X listed above.

```r
# Select vars and smoking indicator
df5b = df %>%
  select(dbrwt, tobacco, rectype, pldel3, cntocpop, stresfip,
                      dmage, mrace3, dmar, adequacy, csex, dplural) %>%
  mutate(stresfip = as.factor(stresfip), dmage = as.factor(dmage))

# Check to make sure all are dummies
dummies <- c("tobacco", "rectype", "pldel3", "cntocpop", "stresfip",
                "dmage", "mrace3", "dmar", "adequacy", "csex", "dplural")
for (var in dummies) {
  print(class(df5b[, var]))
}

# Run regression
reg_5b <- lm(dbrwt ~ tobacco + rectype + pldel3 + cntocpop + stresfip +
               dmage + mrace3 + dmar + adequacy + csex + dplural +
               rectype*pldel3 + rectype*cntocpop + rectype+stresfip +
               rectype*dmage + rectype*mrace3 + rectype*dmar +
               rectype*adequacy + rectype*csex + rectype*dplural +
               pldel3*cntocpop + pldel3*stresfip + pldel3*dmage +
               pldel3*mrace3 + pldel3*dmar + pldel3*adequacy +
               pldel3*csex + pldel3*dplural +
               cntocpop*stresfip + cntocpop*dmage + cntocpop*mrace3 +
               cntocpop*dmar + cntocpop*adequacy + cntocpop*csex + cntocpop*dplural +
               stresfip*dmage +  stresfip*mrace3 +  stresfip*dmar +
               stresfip*adequacy +  stresfip*csex +  stresfip*dplural +
               dmage*mrace3 + dmage*dmar + dmage*adequacy +
               dmage*csex + dmage*dplural +
               mrace3*dmar + mrace3*adequacy + mrace3*csex + mrace3*dplural +
               dmar*adequacy + dmar*csex + dmar*dplural +
               adequacy*csex + adequacy*dplural +
               csex*dplural
               , data = df5b)
summary(reg_5b)
```

```r
# Select variables (Error: cannot allocate vector of size 2616.7 Gb)
#df5a = df %>%
#          select(dbrwt, rectype, pldel3, cntocpop, stresfip,
#                 dmage, mrace3, dmar, adequacy, csex, dplural) %>%
```

```
#          mutate(stresfip = as.factor(stresfip), dmage = as.factor(dmage))

# Saturated Regression
#reg_5a_2nd_attempt <- lm(dbrwt ~ (.)^10, data = df5a)
#summary(reg_5a_2nd_attempt)



# Select variables
df5a = df %>%
          select(dbrwt, tobacco, rectype, pldel3, cntocpop, stresfip,
                 dmage, mrace3, dmar, adequacy, csex, dplural) %>%
          mutate(stresfip = as.factor(stresfip), dmage = as.factor(dmage))

# Create factor variable (for dummies needed for regression) for unique combination of covariate values
df5a_unique = df5a %>%
                select(-c(dbrwt, tobacco)) %>%
                distinct() %>%
                arrange(rectype, pldel3, cntocpop, stresfip,
                  dmage, mrace3, dmar, adequacy, csex, dplural) %>%
                mutate(uniq = row_number())
df5a <- full_join(df5a, df5a_unique, by = c("rectype", "pldel3", "cntocpop",
                  "stresfip", "dmage", "mrace3", "dmar", "adequacy", "csex", "dplural"))
df5a <- df5a %>% mutate(uniq = as.factor(uniq))

# Run regression on dbrwt with saturated model (incl tobacco)
start_t = proc.time()
reg_5a <- lm(dbrwt ~ tobacco + uniq, data = df5a)
end_t = proc.time()
print('time to regress:')
print(end_t[3] - start_t[3])
summary(reg_5a)
```

Let Y = dbrwt, D = tobacco, and X = covariates listed for quesiton 5a. Assume $Y_i = \alpha + \beta D_i + \delta h(x_i) + v_i$ and thereby $Y_i = \alpha + \beta \tilde{D}_i + \delta h(x_i) + u_i$, where $\tilde{D}_i = D_i - \mathbb{E}[D_i|X_i]$ and $u_i = \epsilon_i + \beta \mathbb{E}[D_i|X_i]$. We showed, in class, that $\mathbb{E}[u_i \tilde{D}_i] = 0$ under un-confounded-ness and homogeneous treatment, and can get a consistent estimate of β if the CEF is known. While our CEF is not known, given we have linear CEF with a saturated model, we can simply include X's as additional control.

Let matrix $Z = [D\ E[D|X_i]]$.

$$
\begin{aligned}
\tau_{(OLS\ 5b)} &= \frac{\mathbb{E}[Z_i Y_i]}{Var[Z_i]} \\
&= \frac{\mathbb{E}[\tilde{D}_i Y_i]}{Var[\tilde{D}_i]} && \text{given consistent estimate} \\
&= \frac{\mathbb{E}[\mathbb{E}[\tilde{D} Y_i | D_i, X_i]]}{Var[\tilde{D}_i]} \\
&= \frac{\mathbb{E}[\tilde{D} \mathbb{E}[Y_i | D_i, X_i]]}{Var[\tilde{D}_i]} \\
&= \frac{\mathbb{E}[\tilde{D}(\mathbb{E}[Y_i | D_i = 0, X_i] + \mathbb{E}[\tau_i | D_i, X_i] D_i)]}{Var[\tilde{D}_i]} \\
&= \frac{\mathbb{E}[\tilde{D}(\mathbb{E}[Y_i | D_i = 0, X_i] + \mathbb{E}[\tau_i | X_i] D_i)]}{Var[\tilde{D}_i]} && \text{CIA} \\
&= \frac{\mathbb{E}[\tilde{D} \mathbb{E}[Y_i | D_i = 0, X_i] + \tilde{D} D_i \mathbb{E}[\tau_i]]}{Var[\tilde{D}_i]} && \text{Trt independent of X} \\
&= \frac{\mathbb{E}[D_i \mathbb{E}[Y_i | D_i = 0, X_i] - \mathbb{E}[D_i | X_i] \mathbb{E}[Y_i | D_i = 0, X_i] + \tilde{D} D_i \mathbb{E}[\tau_i]]}{Var[\tilde{D}_i]} \\
&= \frac{\mathbb{E}[D_i \mathbb{E}[Y_i | D_i = 0, X_i]] - \mathbb{E}[\mathbb{E}[D_i | X_i] \mathbb{E}[Y_i | D_i = 0, X_i]] + \mathbb{E}[\tilde{D} D_i \mathbb{E}[\tau_i]]}{Var[\tilde{D}_i]} \\
&= \frac{\mathbb{E}[D_i] \mathbb{E}[Y_i | D_i = 0, X_i] - \mathbb{E}[D_i] \mathbb{E}[Y_i | D_i = 0, X_i] + \mathbb{E}[\tilde{D} D_i] \mathbb{E}[\tau_i]}{Var[\tilde{D}_i]} \\
&= \frac{\mathbb{E}[\tilde{D} D_i] \mathbb{E}[\tau_i]}{Var[\tilde{D}_i]} \\
&= \frac{[\mathbb{E}[D_i^2] - \mathbb{E}[D_i]^2] \mathbb{E}[\tau_i]}{Var[\tilde{D}_i]} \\
&= \frac{Var[D_i] \mathbb{E}[\tau_i]}{Var[D_i]} \\
&= \mathbb{E}[\tau_i]
\end{aligned}
$$

Hence, we know our regression from 5b estimates Average Treatment Effect.

Furthermore, we have assumed CIA holds, that conditioning on X's, materal tobacco is as good as randomly assigned. Hence, matching exactly on X's and comparing outcomes within the match across treatment and control groups should also derive Average Treatment Effect. Hence, our regression from 5b would estimate the same thing as exact matching.

Part (c)

Develop a weighted version of the exact matching estimator that estimates the same thing as the regression above (regardless of whether the treatment effect is independent of covariates).

```
#Match on covariates
match_on <- c("rectype", "pldel3", "cntocpop",
              "stresfip", "dmage", "mrace3", "dmar", "adequacy", "csex", "dplural")
```

Part (d)

Estimate the weighted matching estimator you propose. Compare it to the regression estimate from part (b). Are they similar?

```
count <- df5a %>%
            count(tobacco, uniq) %>%
            arrange(uniq, tobacco) %>%
            reshape(idvar = "uniq", timevar = "tobacco", direction = "wide") %>%
            mutate(bin_w_trt_cont = ifelse(is.na(n.1) |is.na(n.2), NA, 1)) %>%
            rename(tobacco_2 = n.2, tobacco_1 = n.1) %>%
            gather(tobacco_val, count, -c("uniq", "bin_w_trt_cont")) %>%
            mutate(tobacco = as.factor(ifelse(tobacco_val == "tobacco_1", 1, 2))) %>%
            select(uniq, tobacco, count, bin_w_trt_cont, -tobacco_val) %>%
            arrange(uniq, tobacco)

head(count, n=30)


df5c <- full_join(df5a, count, by = c("tobacco", "uniq")) %>%
          mutate(weight_control_cell = NA, weight_control_cell = ifelse(tobacco = 1, 1/n, NA),
                 wijyj = weight_control_cell*dbrwt)
```

Part (e)

Is the sample size of your regression the same as the sample size of your matching estimator, or does the regression have more observations? If the regression has more observations, why don't these extra observations influence the treatment effect estimate?

The regression has more observations. However, these do not influence the treatment effect estimate because treated and untreated observations without any untreated or treated match (respectively) are given a weight of zero.

Part (f)

Compute a standard error for your matching estimator using the formula from Imbens (2015). Specifically, note that your matching estimator should have a form

$$\frac{1}{N_t}\sum_{d_i=1} w_i y_i - \frac{1}{N_c}\sum_{d_i=0} w_i y_i$$

where $\sum_{d_i=1} w_i = N_t$ and $\sum_{d_i=0} w_i = N_c$. Then the conditional variance is approximately

$$\sum_i \left(\frac{d_i}{N_t^2} + \frac{1-d_i}{N_c^2} w_i^2 \hat{\sigma}^2_{d_i}(x_i) \right),$$

where $\hat{\sigma}^2_{d_i}(x_i) = \frac{1}{2}(y_i - y_{nn(i)})$, and $y_{nn(i)}$ is the nearest neighbor to observation i with the *same* treatment status. Figure out the implicit weights w_i in your estimator from part (d), and compute the conditional variance. Is it close to your regression coefficient variance?

```
# Compute a standard error for your matching estimator using the formula from Imbens (2015).


# compute the conditional variance of estimator from (d)
```

Problem 6

Concisely and coherently summarize your overall results, providing some intuition. Write it like you would the conclusion of a paper. In this summary, describe whether you think your best estimate of the effects of smoking is credibly identified. State why or why not.