

EXAMPLE: ANALYSIS OF THE NHANES DATA

In this example, we will explore propensity score based analyses using the publicly available (U.S.) National Health and Nutrition Examination Survey (NHANES). For this, may use the libraries `NHANES`, `tableone`, and `Matching` in R. We will focus our analysis on the question of whether current smoking affects average systolic blood pressure. The variables we will need are:

- `BPSysAve` (systolic blood pressure, response),
- `SmokeNow` (smoking status: 0–No, 1–Yes),
- `Gender`,
- `Age`,
- `Race3`,
- `Education`,
- `MaritalStatus`, and
- `Poverty`

where the first two are the outcome and exposure of interest and the remaining are potential confounders. Additionally, we will restrict our attention to adults (> 17 years old) in the second wave of the survey.

```
library(NHANES);library(tableone);library(Matching)
NHANES$SmokeNow <- as.numeric(NHANES$SmokeNow)-1
small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2011_12" & NHANES$Age > 17,c(3,4,8:11,13,25,61)])
dim(small.nhanes) ## 1377

+ [1] 1377    9

vars <- c("Gender", "Age", "Race3", "Education", "MaritalStatus", "Poverty")

fit0<-lm(BPSysAve~SmokeNow,data=small.nhanes)
round(coef(summary(fit0)),5)

+           Estimate Std. Error   t value Pr(>|t|)
+ (Intercept) 125.61381    0.63365 198.23993  0.00000
+ SmokeNow    -3.67936    0.96395  -3.81696  0.00014

fit1<-lm(BPSysAve~SmokeNow+Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,data=small.nhanes)
round(drop1(fit1,test='F'),5)

+ Single term deletions
+
+ Model:
+ BPSysAve ~ SmokeNow + Gender + Age + Race3 + Education + MaritalStatus +
+   HHIncome + Poverty
+
+           Df Sum of Sq    RSS    AIC  F value           Pr(>F)
+ <none>             324962 7583.7
+ SmokeNow          1      336 325297 7583.1    1.3921         0.23826
+ Gender            1     3133 328095 7594.9   12.9870         0.00033 ***
+ Age               1     67002 391964 7839.8 277.7314 < 0.0000000000000002 ***
+ Race3             5      2495 327457 7584.2    2.0685         0.06680 .
+ Education         4      2982 327944 7588.2    3.0907         0.01515 *
+ MaritalStatus     5      4129 329090 7591.0    3.4228         0.00445 **
+ HHIncome          11     11696 336657 7610.3    4.4072 < 0.0000000000000002 ***
+ Poverty           1      8340 333301 7616.5   34.5697 < 0.0000000000000002 ***
+ ---
+ Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Checking Confounder balance

- In PS-based methods, the goal of the treatment model is to eliminate imbalance in the distribution of covariates between treated and untreated subjects.

- Achieving balance on other covariates (particularly strong predictors of treatment) is unhelpful.
- The goal is *not* to build an excellent predictive model for the treatment.

Common measures of balance include:

- Standardized mean difference (SMD) or proportion:

$$\frac{\bar{x}^{1,w} - \bar{x}^{0,w}}{\sqrt{0.5(v^{1,w} + v^{0,w})}} \quad \text{where} \quad \bar{x}^{z,w} = \frac{1}{n} \sum_{i=1}^n \frac{\mathbb{1}_z(z_i) x_i}{f_{Z|X}^{\mathcal{O}}(z_i|x_i)},$$

i.e. the weighted sample mean of variable X among those with treatment value z , and similarly $v^{z,w}$ is the weighted variance estimate.

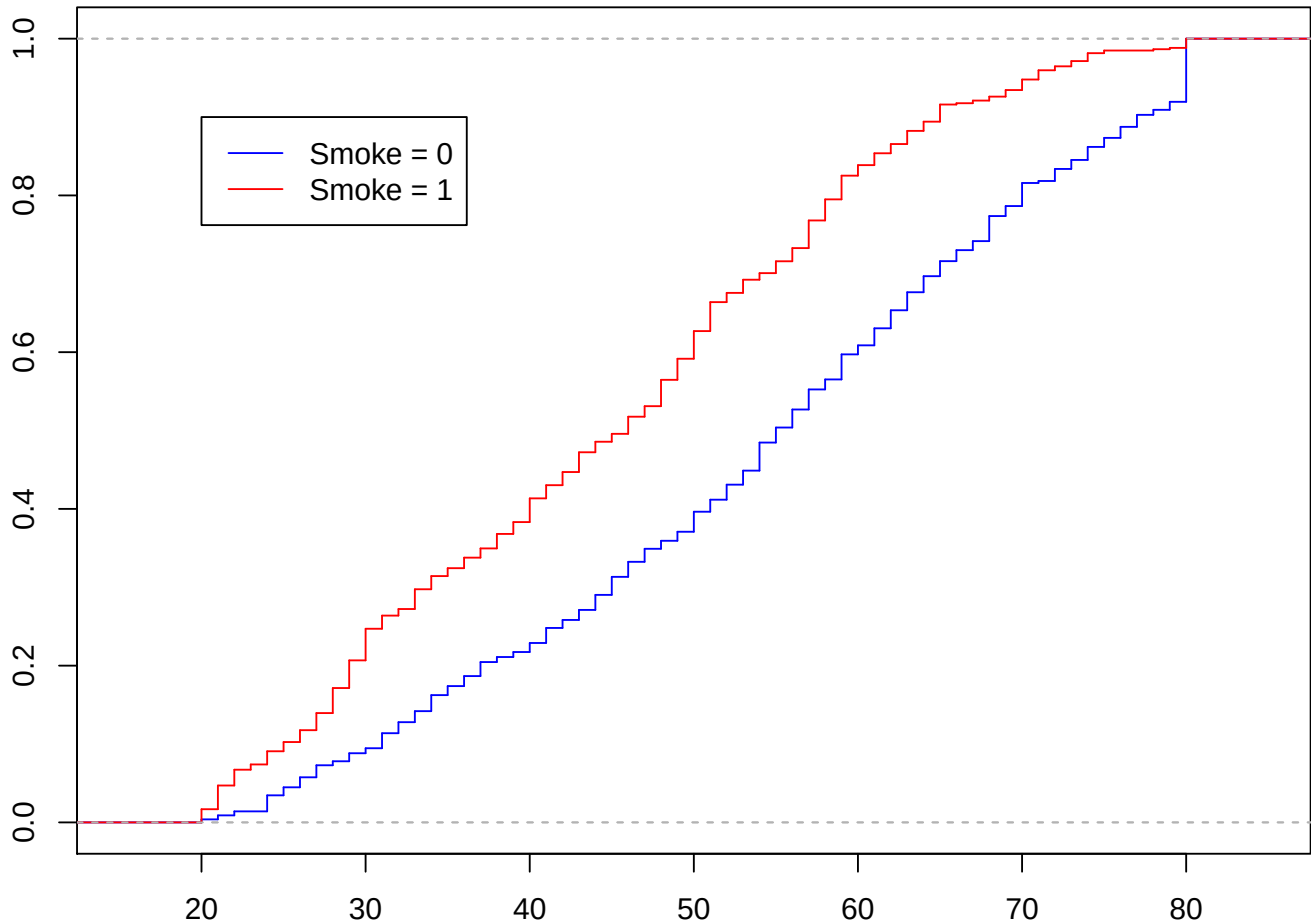
- For all methods of analysis other than IPW, the weights are taken to be 1 for all subjects.
- SMD of 0.1 or less typically considered reasonable.
- Visual examination of weighted empirical CDFs among the treated and untreated (for binary or categorical treatment).

```
tabUnmatched <- CreateTableOne(vars = vars, strata = "SmokeNow", data = small.nhanes, test = FALSE)
print(tabUnmatched, smd = TRUE)
```

```
+           Stratified by SmokeNow
+           0           1           SMD
+ n           782           595
+ Gender = male (%) 432 (55.2) 369 (62.0) 0.138
+ Age (mean (sd))  54.33 (16.52) 44.96 (15.11) 0.592
+ Race3 (%)
+   Asian           25 ( 3.2)   15 ( 2.5)   0.315
+   Black           43 ( 5.5)   64 (10.8)
+   Hispanic        26 ( 3.3)   38 ( 6.4)
+   Mexican         45 ( 5.8)   35 ( 5.9)
+   White          630 (80.6)  416 (69.9)
+   Other           13 ( 1.7)   27 ( 4.5)
+ Education (%)
+   8th Grade       59 ( 7.5)   33 ( 5.5)   0.512
+   9 - 11th Grade  71 ( 9.1)  120 (20.2)
+   High School     152 (19.4)  151 (25.4)
+   Some College    256 (32.7)  210 (35.3)
+   College Grad    244 (31.2)   81 (13.6)
+ MaritalStatus (%)
+   Divorced        85 (10.9)   77 (12.9)   0.488
+   LivePartner     61 ( 7.8)   96 (16.1)
+   Married         453 (57.9)  240 (40.3)
+   NeverMarried    108 (13.8)  142 (23.9)
+   Separated       6 ( 0.8)   14 ( 2.4)
+   Widowed         69 ( 8.8)   26 ( 4.4)
+ Poverty (mean (sd)) 3.11 (1.65) 2.38 (1.58) 0.453
```

```
Smoke <- small.nhanes$SmokeNow
age0 <- small.nhanes$Age[Smoke==0]
age1 <- small.nhanes$Age[Smoke==1]
ecdf0 <- ecdf(age0)
ecdf1 <- ecdf(age1)
par(mar=c(2,2,2,0))
plot(ecdf0, verticals=TRUE, do.points=FALSE, main='Empirical cdfs for Age', col='blue')
plot(ecdf1, verticals=TRUE, do.points=FALSE, add=TRUE, col='red')
legend(20,0.9,c('Smoke = 0', 'Smoke = 1'), col=c('blue','red'), lty=1)
```

Empirical cdfs for Age



There is clear age imbalance between the two smoking groups. This implies that if Age is also a possible predictor of the outcome, then it is a possible confounder.

Building the propensity score: We attempt to build and assess a propensity score model using the available covariates:

```
ps.mod <- glm(SmokeNow~Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
              data=small.nhanes,family="binomial")
ps.lr <- predict(ps.mod,type="response")
summary(ps.lr)

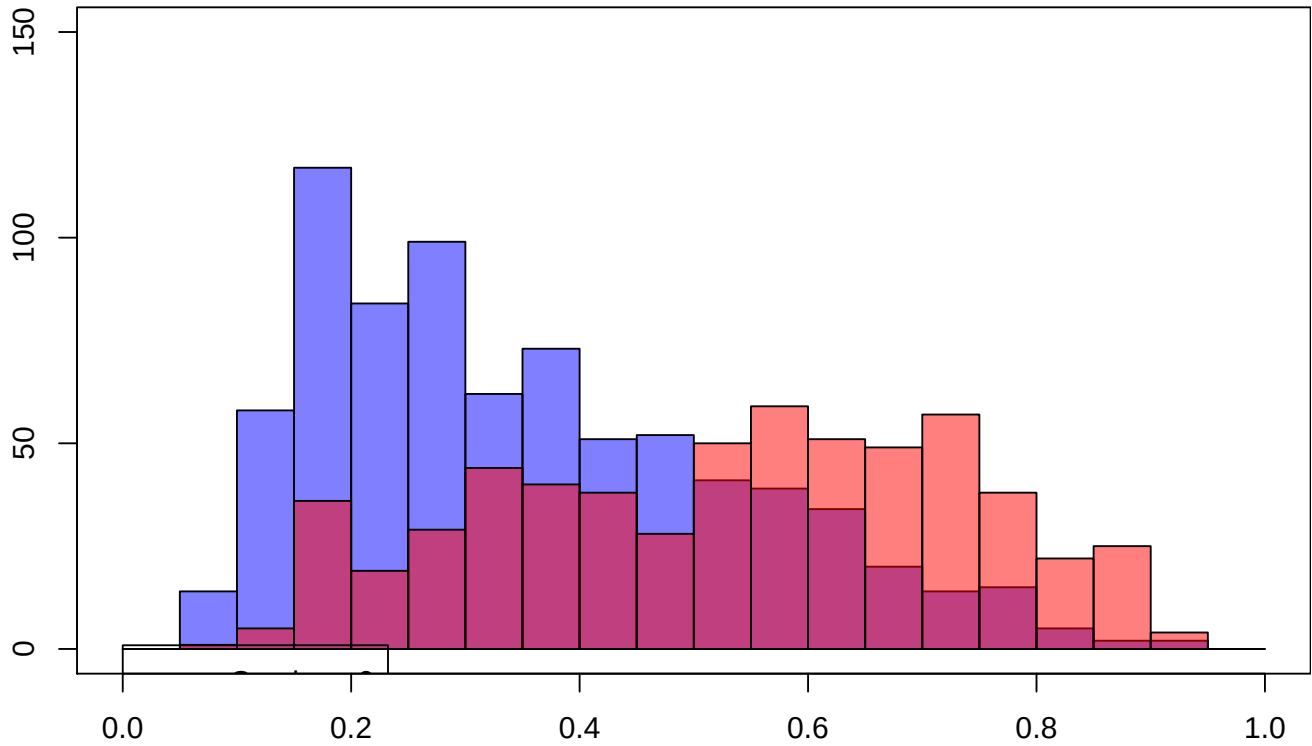
+   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
+ 0.06858 0.25228 0.40382 0.43210 0.59621 0.94074

ps0<-ps.lr[Smoke==0]
ps1<-ps.lr[Smoke==1]
quints <- c(0,quantile(ps.lr,seq(.2,1,.2)))
rbind(table(cut(ps.lr[Smoke==0],quints)),
      table(cut(ps.lr[Smoke==1],quints)))

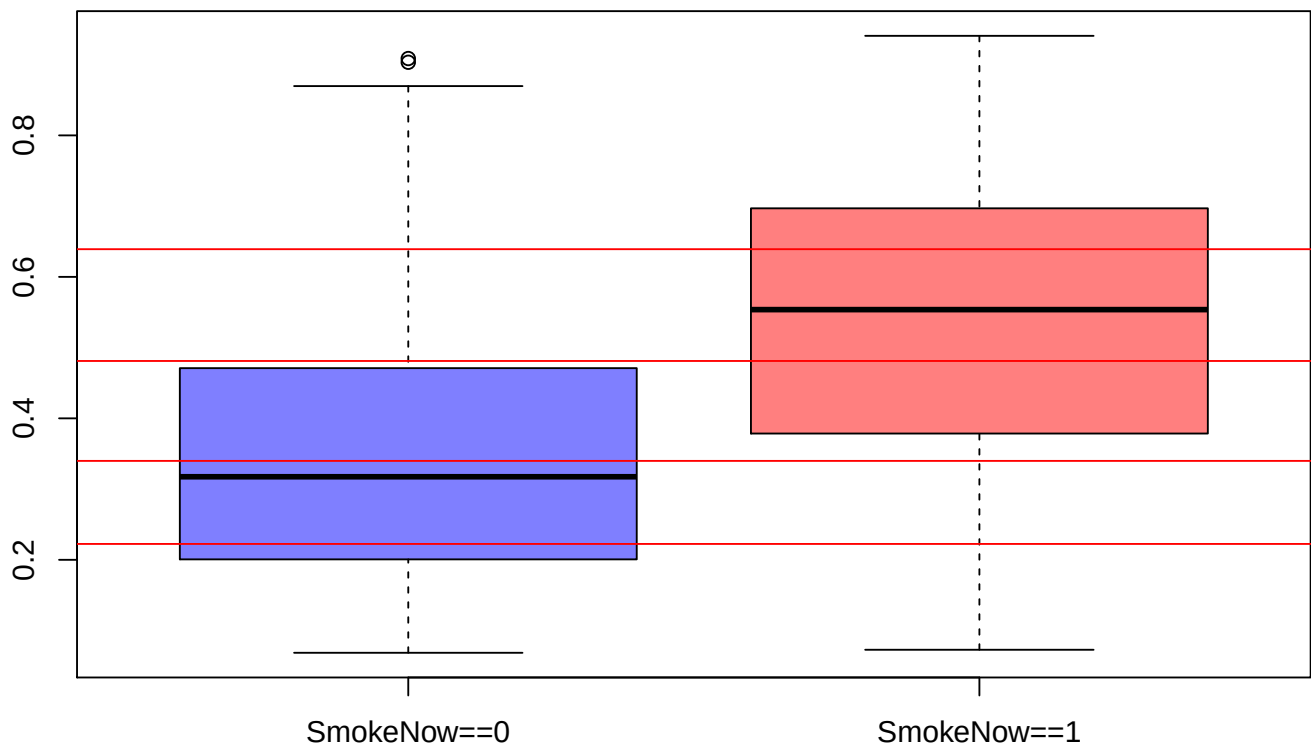
+   (0,0.222] (0.222,0.34] (0.34,0.481] (0.481,0.639] (0.639,0.941]
+ [1,]      231         194         167         121          69
+ [2,]      47          82         105         157         204

par(mar=c(2,2,2,0))
hist(ps0, col=rgb(0,0,1,0.5), breaks=seq(0,1,by=0.05), ylim=c(0,150),
     main="Propensity Score overlap", xlab="PS")
hist(ps1, col=rgb(1,0,0,0.5), breaks=seq(0,1,by=0.05), add=T);box()
legend(0,0.9,c('Smoke = 0', 'Smoke = 1'),col=c(rgb(0,0,1,0.5),rgb(1,0,0,0.5)),lty=1)
```

Propensity Score overlap



```
boxplot(ps0,ps1,ylab="PS",xlab="Treatment Group",names=c('SmokeNow==0','SmokeNow==1'),
        col=c(rgb(0,0,1,0.5),rgb(1,0,0,0.5)));abline(h=quints[2:5],col="red")
```



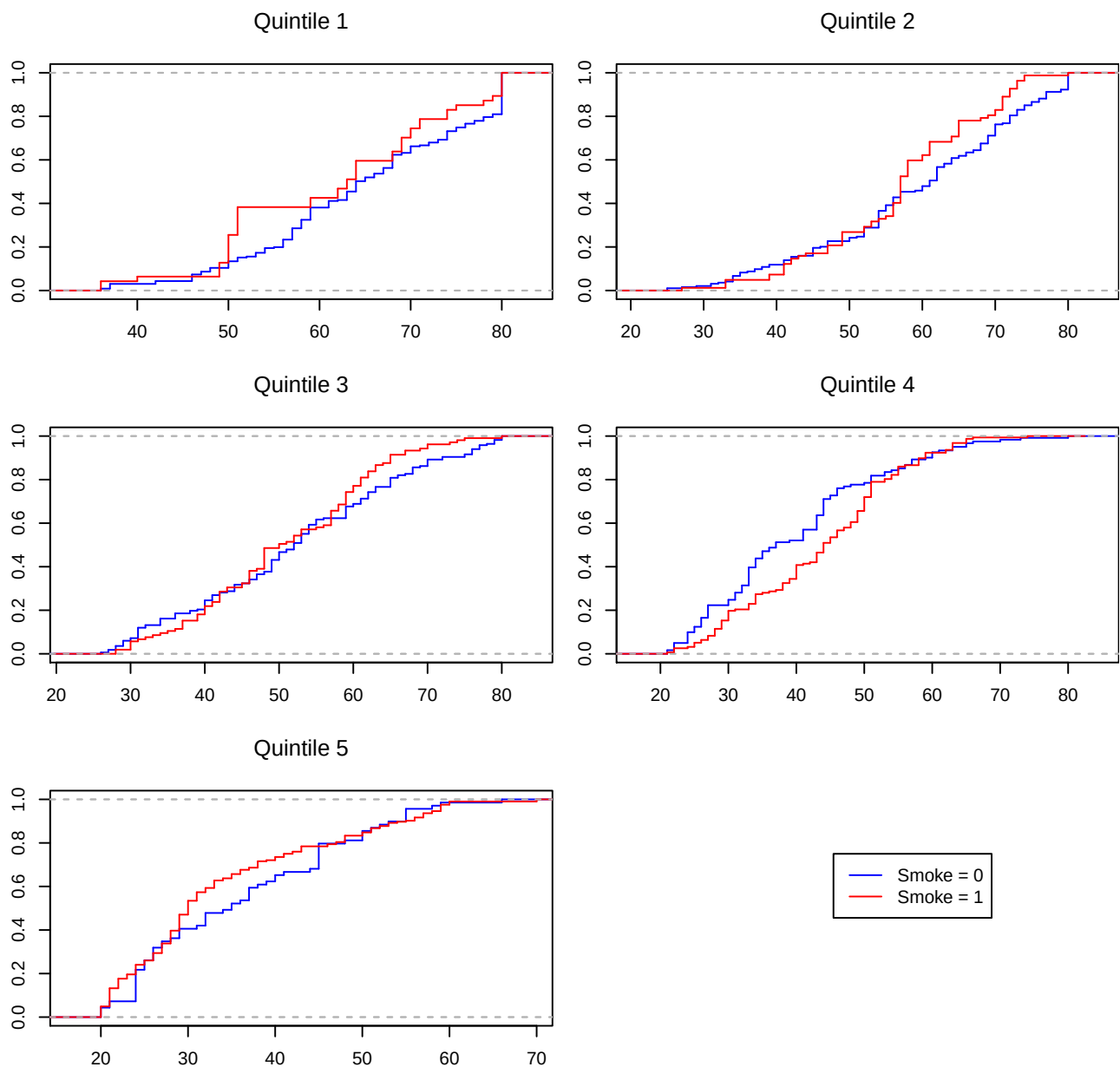
We can therefore proceed to check for balance knowing we have sufficient numbers of smokers and non-smokers in each quintile to ensure the stratum-specific estimates are not too unstable.

```

Pcat<-as.numeric(cut(ps.lr,quints,include.lowest=T))
par(mar=c(2,3,4,0),mfrow=c(3,2),oma=c(0,0,2,0))
for(k in 1:5){
  age0 <- small.nhanes$Age[Smoke==0 & Pcat==k]
  age1 <- small.nhanes$Age[Smoke==1 & Pcat==k]
  ecdf0 <- ecdf(age0)
  ecdf1 <- ecdf(age1)
  plot(ecdf0, verticals=TRUE, do.points=FALSE,
       main=substitute(paste('Quintile ',k),list(k=k)),col='blue')
  plot(ecdf1, verticals=TRUE, do.points=FALSE, add=TRUE, col='red')
}
plot(age0,type='n',ylim=range(0,1),axes=FALSE)
title("ECDFs for Age by PS quintile",outer = TRUE)
legend(30,0.75,c('Smoke = 0', 'Smoke = 1'),col=c('blue','red'),lty=1)

```

ECDFs for Age by PS quintile



Balance does not appear to have been achieved: we have SMDs > 0.1 for at least three quintiles for all variables,

and the empirical CDFs of age do not overlap in several quintiles.

```
W<-(Smoke==0)/(1-ps.lr)+(Smoke==1)/ps.lr
smd.mat<-ExtractSmd(tabUnmatched)
for(k in 1:5){
  nhanesQ<-small.nhanes[Pcat == k,]
  tabQs <- CreateTableOne(vars = vars, strata = "SmokeNow", data = nhanesQ, test = FALSE)
  smd.mat<-cbind(smd.mat,ExtractSmd(tabQs))
}
colnames(smd.mat)<-c('Original','Q1','Q2','Q3','Q4','Q5')
round(smd.mat,4)
```

	Original	Q1	Q2	Q3	Q4	Q5
+ Gender	0.1379	0.1017	0.1043	0.0286	0.2003	0.0308
+ Age	0.5918	0.2574	0.1715	0.0993	0.3106	0.1642
+ Race3	0.3148	0.3169	0.1117	0.3444	0.4147	0.2869
+ Education	0.5119	0.5378	0.4167	0.2800	0.2378	0.3018
+ MaritalStatus	0.4877	0.4320	0.2386	0.2725	0.2332	0.2608
+ Poverty	0.4530	0.0865	0.1256	0.1136	0.0041	0.1455

We may try to find evidence in smaller PS strata, such as those formed by deciles of the propensity score.

```
dec <- c(0,quantile(ps.lr,seq(0.1,1,.1)))
Pcat10 <- cut(ps.lr,dec,labels=1:10)
SMD.10.table <- ExtractSmd(tabUnmatched)
for(k in 1:10) {
  tabPSdec <- CreateTableOne(vars = vars, strata = "SmokeNow",
    data = small.nhanes[Pcat10==k,], test = FALSE)
  SMD.10.table <- cbind(SMD.10.table,ExtractSmd(tabPSdec))
}
colnames(SMD.10.table)<-c('Original','Q1','Q2','Q3','Q4','Q5','Q6','Q7','Q8','Q9','Q10')
round(SMD.10.table,4)
```

	Original	Q1	Q2	Q3	Q4	Q5	Q6	Q7
+ Gender	0.1379	0.0733	0.2857	0.2263	0.1167	0.2086	0.2267	0.2029
+ Age	0.5918	0.4091	0.1160	0.1359	0.1182	0.1356	0.0119	0.0223
+ Race3	0.3148	0.4563	0.5138	0.3124	0.1007	0.4134	0.4281	0.6159
+ Education	0.5119	0.4904	0.6253	0.7638	0.4040	0.5052	0.3837	0.3111
+ MaritalStatus	0.4877	0.6894	0.4963	0.7846	0.4276	0.4922	0.1702	0.2716
+ Poverty	0.4530	0.0340	0.0791	0.1044	0.1343	0.0866	0.1344	0.0347
		Q8	Q9	Q10				
+ Gender	0.2496	0.0551	0.0850					
+ Age	0.6445	0.1597	0.1157					
+ Race3	0.3408	0.3786	0.6859					
+ Education	0.6091	0.3403	0.5431					
+ MaritalStatus	0.4277	0.3800	0.3180					
+ Poverty	0.0248	0.0735	0.3522					

This does not assist with providing evidence of balance.

Matching: We now attempt matching using the propensity score. The function `MatchBalance` from the `Matching` library provides many more details than `CreateTableOne`, including:

- mean, median, and maximum difference in empirical CDF plots,
- mean, median, and maximum difference in empirical QQ plots,
- Kolmogorov-Smirnov statistics,
- ratio of variances,
- p-value for t-test.
- Note that SMDs reported are multiplied by 100.

```

small.nhanes$ps.lr<-ps.lr
ps.lr.match <- Match(Tr=small.nhanes$SmokeNow,X=small.nhanes$ps.lr,estimand="ATE",ties=FALSE)
matched.samp <- small.nhanes[c(ps.lr.match$index.control,ps.lr.match$index.treated),]
table(table(c(ps.lr.match$index.control, ps.lr.match$index.treated)))

+
+   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  23  29
+ 844 256 120  60  37  19   6   5   3   5   5   3   2   4   3   2   1   1
+  38
+   1

tabMatched <- CreateTableOne(vars = vars, strata = "SmokeNow",data = matched.samp, test = FALSE)
MatchBalance(SmokeNow~Gender+Age+Education+MaritalStatus+Poverty,data=small.nhanes,match.out=ps.lr.match)

+
+ ***** (V1) Gendermale *****
+
+               Before Matching      After Matching
+ mean treatment.....      0.62017      0.54539
+ mean control.....      0.55243      0.55628
+ std mean diff.....      13.945      -2.1869
+
+ mean raw eQQ diff.....      0.068908      0.010893
+ med  raw eQQ diff.....           0           0
+ max  raw eQQ diff.....           1           1
+
+ mean eCDF diff.....      0.033869      0.0054466
+ med  eCDF diff.....      0.033869      0.0054466
+ max  eCDF diff.....      0.067738      0.010893
+
+ var ratio (Tr/Co).....      0.9531      1.0045
+ T-test p-value.....      0.011311      0.55234
+
+
+ ***** (V2) Age *****
+
+               Before Matching      After Matching
+ mean treatment.....      44.958      50.385
+ mean control.....      54.327      50.318
+ std mean diff.....      -62.015      0.41148
+
+ mean raw eQQ diff.....      9.3294      0.97603
+ med  raw eQQ diff.....          10           1
+ max  raw eQQ diff.....          14           5
+
+ mean eCDF diff.....      0.1536      0.016001
+ med  eCDF diff.....      0.15554      0.014524
+ max  eCDF diff.....      0.2521      0.042847
+
+ var ratio (Tr/Co).....      0.83608      0.92646
+ T-test p-value..... < 0.000000000000000222      0.8909
+ KS Bootstrap p-value.. < 0.000000000000000222      0.118
+ KS Naive p-value..... < 0.000000000000000222      0.15956
+ KS Statistic.....      0.2521      0.042847
+
+
+ ***** (V3) Education9 - 11th Grade *****
+
+               Before Matching      After Matching
+ mean treatment.....      0.20168      0.14089
+ mean control.....      0.090793      0.14089
+ std mean diff.....      27.612           0
+
+ mean raw eQQ diff.....      0.11092           0
+ med  raw eQQ diff.....           0           0

```

```

+ max raw eQQ diff..... 1 0
+
+ mean eCDF diff..... 0.055444 0
+ med eCDF diff..... 0.055444 0
+ max eCDF diff..... 0.11089 0
+
+ var ratio (Tr/Co)..... 1.9512 1
+ T-test p-value..... 0.000000014534 1
+
+
+ ***** (V4) EducationHigh School *****
+ Before Matching After Matching
+ mean treatment..... 0.25378 0.22658
+ mean control..... 0.19437 0.21786
+ std mean diff..... 13.64 2.081
+
+ mean raw eQQ diff..... 0.060504 0.0087146
+ med raw eQQ diff..... 0 0
+ max raw eQQ diff..... 1 1
+
+ mean eCDF diff..... 0.029704 0.0043573
+ med eCDF diff..... 0.029704 0.0043573
+ max eCDF diff..... 0.059408 0.0087146
+
+ var ratio (Tr/Co)..... 1.2098 1.0284
+ T-test p-value..... 0.009248 0.57996
+
+
+ ***** (V5) EducationSome College *****
+ Before Matching After Matching
+ mean treatment..... 0.35294 0.34713
+ mean control..... 0.32737 0.33115
+ std mean diff..... 5.3473 3.3548
+
+ mean raw eQQ diff..... 0.026891 0.015977
+ med raw eQQ diff..... 0 0
+ max raw eQQ diff..... 1 1
+
+ mean eCDF diff..... 0.012788 0.0079884
+ med eCDF diff..... 0.012788 0.0079884
+ max eCDF diff..... 0.025575 0.015977
+
+ var ratio (Tr/Co)..... 1.0375 1.0232
+ T-test p-value..... 0.32201 0.38527
+
+
+ ***** (V6) EducationCollege Grad *****
+ Before Matching After Matching
+ mean treatment..... 0.13613 0.19245
+ mean control..... 0.31202 0.24328
+ std mean diff..... -51.246 -12.89
+
+ mean raw eQQ diff..... 0.17479 0.050835
+ med raw eQQ diff..... 0 0
+ max raw eQQ diff..... 1 1
+
+ mean eCDF diff..... 0.087943 0.025418
+ med eCDF diff..... 0.087943 0.025418
+ max eCDF diff..... 0.17589 0.050835
+
+ var ratio (Tr/Co)..... 0.54806 0.84419

```



```

+ T-test p-value..... 0.0000000000000013323      0.00028606
+
+
+ ***** (V7) MaritalStatusLivePartner *****
+               Before Matching      After Matching
+ mean treatment..... 0.16134      0.10167
+ mean control..... 0.078005      0.1024
+ std mean diff..... 22.637      -0.24021
+
+ mean raw eQQ diff.... 0.084034      0.00072622
+ med raw eQQ diff.... 0      0
+ max raw eQQ diff.... 1      1
+
+ mean eCDF diff..... 0.04167      0.00036311
+ med eCDF diff..... 0.04167      0.00036311
+ max eCDF diff..... 0.083339      0.00072622
+
+ var ratio (Tr/Co).... 1.8822      0.99371
+ T-test p-value..... 0.0000035776      0.94907
+
+
+ ***** (V8) MaritalStatusMarried *****
+               Before Matching      After Matching
+ mean treatment..... 0.40336      0.50327
+ mean control..... 0.57928      0.51198
+ std mean diff..... -35.831      -1.7423
+
+ mean raw eQQ diff.... 0.17479      0.0087146
+ med raw eQQ diff.... 0      0
+ max raw eQQ diff.... 1      1
+
+ mean eCDF diff..... 0.087961      0.0043573
+ med eCDF diff..... 0.087961      0.0043573
+ max eCDF diff..... 0.17592      0.0087146
+
+ var ratio (Tr/Co).... 0.98787      1.0005
+ T-test p-value..... 0.000000000073435      0.62133
+
+
+ ***** (V9) MaritalStatusNeverMarried *****
+               Before Matching      After Matching
+ mean treatment..... 0.23866      0.16558
+ mean control..... 0.13811      0.16558
+ std mean diff..... 23.569      0
+
+ mean raw eQQ diff.... 0.10084      0
+ med raw eQQ diff.... 0      0
+ max raw eQQ diff.... 1      0
+
+ mean eCDF diff..... 0.050274      0
+ med eCDF diff..... 0.050274      0
+ max eCDF diff..... 0.10055      0
+
+ var ratio (Tr/Co).... 1.5271      1
+ T-test p-value..... 0.0000029703      1
+
+
+ ***** (V10) MaritalStatusSeparated *****
+               Before Matching      After Matching
+ mean treatment..... 0.023529      0.013798
+ mean control..... 0.0076726      0.022513

```

```

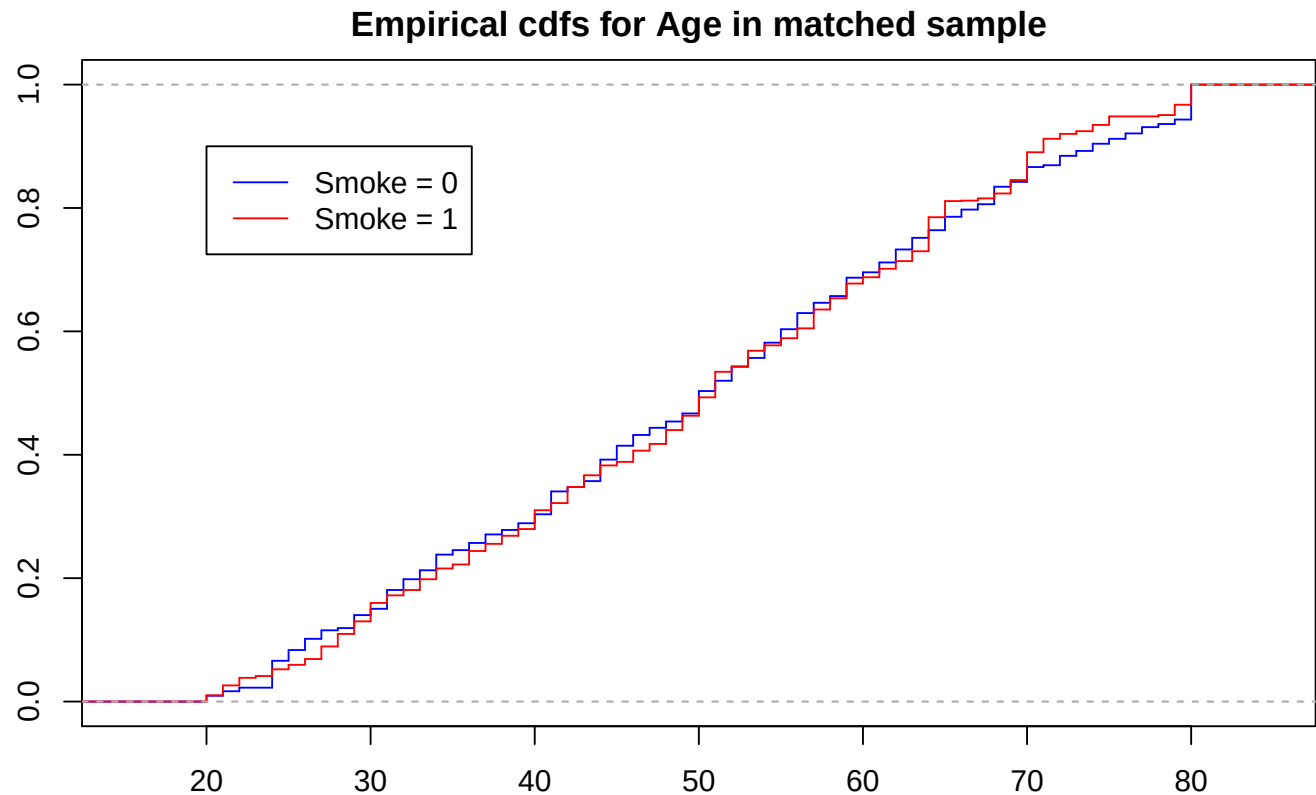
+ std mean diff..... 10.452 -7.4679
+
+ mean raw eQQ diff..... 0.016807 0.0087146
+ med raw eQQ diff..... 0 0
+ max raw eQQ diff..... 1 1
+
+ mean eCDF diff..... 0.0079284 0.0043573
+ med eCDF diff..... 0.0079284 0.0043573
+ max eCDF diff..... 0.015857 0.0087146
+
+ var ratio (Tr/Co)..... 3.0189 0.61837
+ T-test p-value..... 0.022929 0.089577
+
+
+ ***** (V11) MaritalStatusWidowed *****
+ Before Matching After Matching
+ mean treatment..... 0.043697 0.095134
+ mean control..... 0.088235 0.066812
+ std mean diff..... -21.769 9.6497
+
+ mean raw eQQ diff..... 0.043697 0.028322
+ med raw eQQ diff..... 0 0
+ max raw eQQ diff..... 1 1
+
+ mean eCDF diff..... 0.022269 0.014161
+ med eCDF diff..... 0.022269 0.014161
+ max eCDF diff..... 0.044538 0.028322
+
+ var ratio (Tr/Co)..... 0.51964 1.3807
+ T-test p-value..... 0.00073816 0.0047204
+
+
+ ***** (V12) Poverty *****
+ Before Matching After Matching
+ mean treatment..... 2.3776 2.6553
+ mean control..... 3.1091 2.7262
+ std mean diff..... -46.293 -4.3845
+
+ mean raw eQQ diff..... 0.728 0.10414
+ med raw eQQ diff..... 0.7 0.09
+ max raw eQQ diff..... 1.65 0.46
+
+ mean eCDF diff..... 0.13667 0.020709
+ med eCDF diff..... 0.15502 0.021786
+ max eCDF diff..... 0.20409 0.048656
+
+ var ratio (Tr/Co)..... 0.91853 0.98412
+ T-test p-value..... 0.0000000000000022204 0.20088
+ KS Bootstrap p-value.. < 0.000000000000000222 0.078
+ KS Naive p-value..... 0.0000000000011907 0.076774
+ KS Statistic..... 0.20409 0.048656
+
+
+ Before Matching Minimum p.value: < 0.000000000000000222
+ Variable Name(s): Age Poverty Number(s): 2 12
+
+ After Matching Minimum p.value: 0.00028606
+ Variable Name(s): EducationCollege Grad Number(s): 6

```

```

age0 <- matched.samp$Age[matched.samp$SmokeNow==0]
age1 <- matched.samp$Age[matched.samp$SmokeNow==1]
ecdf0 <- ecdf(age0)
ecdf1 <- ecdf(age1)
par(mar=c(2,2,2,0))
plot(ecdf0, verticals=TRUE, do.points=FALSE, main='Empirical cdfs for Age in matched sample', col='blue')
plot(ecdf1, verticals=TRUE, do.points=FALSE, add=TRUE, col='red')
legend(20,0.9,c('Smoke = 0', 'Smoke = 1'),col=c('blue','red'),lty=1)

```



Inverse Weighting: We can construct an inverse weighted (IPW) sample and examine it using the survey and Hmisc packages.

```

library(survey);library(Hmisc);library(htmlwidgets)
ps.lr.weight <- small.nhanes$SmokeNow/ps.lr + (1-small.nhanes$SmokeNow)/(1-ps.lr)
nhanes.IPW.lr <- svydesign(ids=~0, data=small.nhanes, weights=ps.lr.weight)
tabIPW <- svyCreateTableOne(vars = vars, strata = "SmokeNow", data = nhanes.IPW.lr, test = FALSE)
print(tabIPW, smd = TRUE)

```

Stratified by SmokeNow				
	0	1	SMD	
n	1379.0	1379.7		
Gender = male (%)	797.4 (57.8)	781.8 (56.7)	0.023	
Age (mean (sd))	50.06 (17.16)	49.83 (15.42)	0.014	
Race3 (%)			0.052	
Asian	40.4 (2.9)	37.9 (2.7)		
Black	119.0 (8.6)	103.3 (7.5)		
Hispanic	75.5 (5.5)	70.6 (5.1)		
Mexican	81.2 (5.9)	87.9 (6.4)		
White	1019.8 (74.0)	1038.8 (75.3)		
Other	43.0 (3.1)	41.2 (3.0)		
Education (%)			0.029	
8th Grade	89.4 (6.5)	96.2 (7.0)		
9 - 11th Grade	184.6 (13.4)	185.8 (13.5)		

```

+      High School      293.9 (21.3)    285.5 (20.7)
+      Some College    482.7 (35.0)    474.2 (34.4)
+      College Grad    328.4 (23.8)    338.1 (24.5)
+      MaritalStatus (%)                                0.023
+      Divorced        161.0 (11.7)    165.1 (12.0)
+      LivePartner     154.6 (11.2)    150.8 (10.9)
+      Married         695.0 (50.4)    690.1 (50.0)
+      NeverMarried    249.9 (18.1)    254.0 (18.4)
+      Separated        23.1 ( 1.7)     20.5 ( 1.5)
+      Widowed         95.4 ( 6.9)     99.2 ( 7.2)
+      Poverty (mean (sd)) 2.80 (1.67)  2.80 (1.63) <0.001

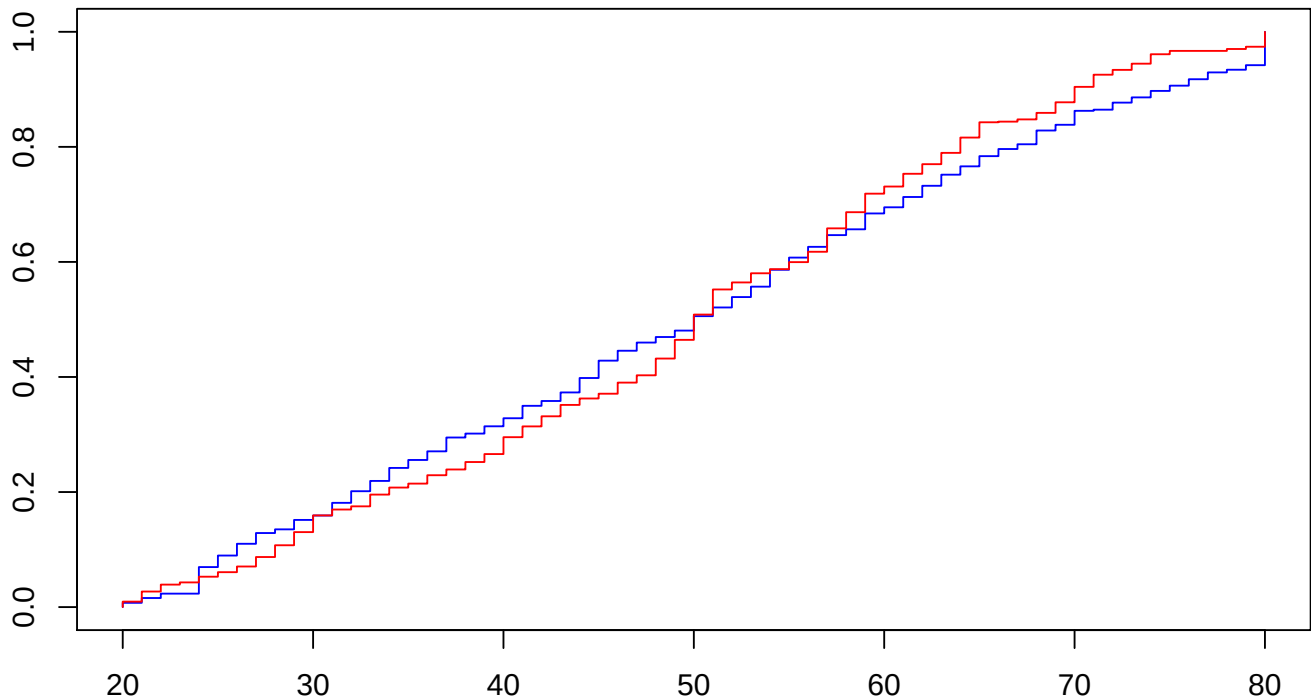
```

```

temp0 <- Ecdf(small.nhanes$Age[Smoke==0],weights=ps.lr.weight[Smoke==0],pl=FALSE)
temp1 <- Ecdf(small.nhanes$Age[Smoke==1],weights=ps.lr.weight[Smoke==1],pl=FALSE)
par(mar=c(2,2,2,0))
plot(temp0$x,temp0$y,ylab="ECDF(Age)",xlab="Age",
      main='Empirical cdfs for Age in weighted sample',col='blue',type="s",lwd=1)
lines(temp1$x,temp1$y,col="red",lwd=1,type='s')

```

Empirical cdfs for Age in weighted sample



```

smd.mat<-cbind(smd.mat,ExtractSmd(tabMatched))
smd.mat<-cbind(smd.mat,ExtractSmd(tabIPW))
colnames(smd.mat)<-c('Original','Q1','Q2','Q3','Q4','Q5','Match','IPW')
round(smd.mat,4)

```

	Original	Q1	Q2	Q3	Q4	Q5	Match	IPW
+ Gender	0.1379	0.1017	0.1043	0.0286	0.2003	0.0308	0.0219	0.0235
+ Age	0.5918	0.2574	0.1715	0.0993	0.3106	0.1642	0.0040	0.0139
+ Race3	0.3148	0.3169	0.1117	0.3444	0.4147	0.2869	0.1278	0.0516
+ Education	0.5119	0.5378	0.4167	0.2800	0.2378	0.3018	0.1471	0.0292
+ MaritalStatus	0.4877	0.4320	0.2386	0.2725	0.2332	0.2608	0.1230	0.0235
+ Poverty	0.4530	0.0865	0.1256	0.1136	0.0041	0.1455	0.0437	0.0001

Here the matched and inverse weighted samples exhibit good balance.

Summary:

- Creating or restoring confounder balance is essential to estimating a causal effect.
- It can be hard to assess overlap or achieve balance in high dimensions.
- The propensity score, a scalar summary of confounding variables, simplifies this task.
- However:
 - fitting a model for treatment does not guarantee balance,
 - fitting a model that predicts treatment with a high degree of precision can be unhelpful.

Estimating the ATE: We proceed now to estimating the average treatment effect (ATE), using:

- outcome regression,
- PS stratification,
- PS matching,
- PS regression,
- IPW.

We will use the PS estimated via logistic regression, as this provided the best balance.

- **Linear regression:**

```
coef(lm(BPSysAve~SmokeNow,data=small.nhanes))[2]
+ SmokeNow
+ -3.679357
coef(lm(BPSysAve~SmokeNow+Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,data=small.nhanes)
+ SmokeNow
+ -1.097768
```

The naive conditional effect estimate is more than 3 times greater than its confounder-adjusted counterpart.

- **Outcome regression no interaction:**

```
nhanes.allsmoke <- small.nhanes
nhanes.allsmoke$SmokeNow <- 1
nhanes.nosmoke <- small.nhanes
nhanes.nosmoke$SmokeNow <- 0
mod1.lm <- lm(BPSysAve~SmokeNow+Gender+Age+Race3+Education+MaritalStatus+HHIncome+Poverty,
data=small.nhanes)
APO.lm.1 <- mean(predict(mod1.lm,nhanes.allsmoke))
APO.lm.0 <- mean(predict(mod1.lm,nhanes.nosmoke))
APO.lm.1 - APO.lm.0
+ [1] -1.097768
```

Conditional and marginal effect are the same in a linear model with no interaction.

- **Outcome regression with interactions:**

```
mod1.lmX <- lm(BPSysAve~SmokeNow+Gender+Age+Race3+Education+
MaritalStatus+HHIncome+Poverty+SmokeNow:HHIncome+SmokeNow:Gender+SmokeNow:Age,data=small.nhanes)
APO.lmX.1 <- mean(predict(mod1.lmX,nhanes.allsmoke))
APO.lmX.0 <- mean(predict(mod1.lmX,nhanes.nosmoke))
APO.lmX.1 - APO.lmX.0
+ [1] -1.402538
```

- **PS stratification**

```
Y<-small.nhanes$BPSysAve
ps.lr.quints <- cut(ps.lr,quints,labels=1:5)
p.strat <- table(ps.lr.quints)/length(ps.lr.quints)
ATE.strat <- rep(NA,5)
for(j in 1:5) {
  ATE.strat[j] <- mean(Y[Smoke == 1 & ps.lr.quints==j])-mean(Y[Smoke == 0 & ps.lr.quints==j])
}
ATE.strat
```

```
+ [1] -8.1736207 -2.2701785 -0.2062732 -1.1820287 2.8633845
sum(ATE.strat*p.strat)
+ [1] -1.816879
```

• PS matching

```
ps.lr.match <- Match(Tr=small.nhanes$SmokeNow,
  X=small.nhanes$ps.lr, estimand="ATE", ties=FALSE)
matched.samp <- small.nhanes[c(ps.lr.match$index.control, ps.lr.match$index.treated),]

dim(matched.samp)
+ [1] 2754 10
mean(matched.samp$BPSysAve[matched.samp$SmokeNow == 1]) -
  mean(matched.samp$BPSysAve[matched.samp$SmokeNow == 0])
+ [1] -0.7000726
```

• PS regression

```
library(splines)
mod1.PS1m1 <- lm(BPSysAve~SmokeNow+ps.lr, data=small.nhanes)
APO.PS1m1.1 <- mean(predict(mod1.PS1m1, nhanes.allsmoke))
APO.PS1m1.0 <- mean(predict(mod1.PS1m1, nhanes.nosmoke))
APO.PS1m1.1 - APO.PS1m1.0
+ [1] -1.10791
mod1.PS1m2 <- lm(BPSysAve~SmokeNow+ps.lr+I(ps.lr^2),
  data=small.nhanes)
APO.PS1m2.1 <- mean(predict(mod1.PS1m2, nhanes.allsmoke))
APO.PS1m2.0 <- mean(predict(mod1.PS1m2, nhanes.nosmoke))
APO.PS1m2.1 - APO.PS1m2.0
+ [1] -1.110337
mod1.PS1m3 <- lm(BPSysAve~SmokeNow+bs(ps.lr, df=4),
  data=small.nhanes)
APO.PS1m3.1 <- mean(predict(mod1.PS1m3, nhanes.allsmoke))
APO.PS1m3.0 <- mean(predict(mod1.PS1m3, nhanes.nosmoke))
APO.PS1m3.1 - APO.PS1m3.0
+ [1] -1.133493
```

• IPW

```
ps.lr.weight <- Smoke/ps.lr + (1-Smoke)/(1-ps.lr)
mean(Smoke*Y*ps.lr.weight) - mean((1-Smoke)*Y*ps.lr.weight)
+ [1] -1.928655
coef(lm(Y ~ Smoke, weights = ps.lr.weight))[2]
+ Smoke
+ -1.991233
```

Additional considerations:

- All of the PS approaches considered rely on substitution estimators.
 - In PS regression, we plug in an estimated PS as a covariate.
 - In IPW, we plug in estimated weights.
- We need to account for this when estimating standard errors and/or confidence intervals.
- Analytically derived asymptotic variances can be used, but are not provided in many standard software packages.
- The easiest approach is to bootstrap.
- Note, however, that the bootstrap is *not* valid for matching.