Homework 3 Due Wednesday April 25th

Problem 1

In this problem and the next, we will analyze data from the National Supported Work Demonstration, a subsidized work program implemented in the mid-70's. The data set contains an experimental sample from a randomized evaluation of the NSW program (`nsw_exper.dta`), and also a non-experimental sample from the Population Survey of Income Dynamics (PSID) (`nsw_psid_withtreated.dta`). In both datasets, the variable `nsw` is the treatment, the variables `re78` and `u78` are outcomes, and the rest are covariates.

Variable Definitions:	
`nsw`	=1 for NSW participants, =0 otherwise
`age`	age in years
`educ`	years of education
`black`	=1 if African American, =0 otherwise
`hispanic`	=1 if Hispanic, =0 otherwise
`married`	=1 if married, =0 otherwise
`re74`	real (inflation adjusted) earnings for 1974
`re75`	real (inflation adjusted) earnings for 1975
`re78`	real (inflation adjusted) earnings for 1978
`u74`	=1 if unemployed in 1974, =0 otherwise
`u75`	=1 if unemployed in 1975, =0 otherwise
`u78`	=1 if unemployed in 1978, =0 otherwise

(a) Using the experimental data (`nsw_exper.dta`), obtain an unbiased estimate of the ATE of NSW on 1978 earnings, its SE and a 95% confidence interval.

```
> exper <- read.dta("nsw_exper.dta")
> nonexper <- read.dta("nsw_psid_withtreated.dta")
>
> #Unadjusted and adjusted experimental results
> library(sandwich)
> library(lmtest)
>
> exp.unadj <- lm(formula = re78~nsw,data=exper)
> coeftest(exp.unadj,vcov = vcovHC(exp.unadj,"HC2"))

t test of coefficients:
```

```
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  4554.80     340.09 13.3928 < 2.2e-16 ***
nsw          1794.34     671.00  2.6741  0.007769 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(b) Using the experimental data, use OLS to estimate the ATE controlling for age, education, race/ethnicity, marital status, and employment and earnings in 1974 and 1975. Report the estimate and its SE and compare it to the one obtained in (a), and explain how they compare with each other and why.

```
> exp.adj <- lm(re78~nsw+age+educ+black+hisp+married+re74+re75+u74+u75,data=exper)
> coeftest(exp.adj,vcov = vcovHC(exp.adj,"HC2"))

t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)
(Intercept)  2.2143e+02  2.8646e+03  0.0773  0.93842
nsw          1.6720e+03  6.6372e+02  2.5192  0.01212 *
age          5.3668e+01  4.0567e+01  1.3229  0.18655
educ         4.0295e+02  1.6308e+02  2.4709  0.01386 *
black       -2.0395e+03  1.0482e+03 -1.9456  0.05235 .
hisp         4.2465e+02  1.4432e+03  0.2942  0.76871
married     -1.4666e+02  8.7002e+02 -0.1686  0.86621
re74         1.2357e-01  1.3328e-01  0.9272  0.35436
re75         1.9458e-02  1.4426e-01  0.1349  0.89277
u74          1.3810e+03  1.5711e+03  0.8790  0.37988
u75         -1.0718e+03  1.4113e+03 -0.7594  0.44801
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The point estimate is very similar to the previous results, because the treatment was randomized and hence not correlated with the covariates in expectation. Of course, due to the finite sample, the correlations are non-zero, leading to a slightly different point estimate. Meanwhile, the standard error also appears very similar: it appears that the covariates are not explaining a huge amount of the variation in the DV (otherwise the standard error would have decreased more substantially), which may be due to the eligibility requirements for participation in the experiment.

(c) File `nsw_psid_withtreated.dta` contains treated units taken from the experiment, but control units replaced by the non-experimental sample from the PSID. We will refer to this file as the non-

experimental dataset. Check the covariate balance in the non-experimental dataset. Decide on a few sensible balance statistics and report them in a table. How do the treatment and control group differ? Among the observed covariates, what seem to be the most important factors that determine selection into the program (the "treatment")?

We can using the Matching code to produce before-matching balance statistics to report here, rather than calculating them "manually."

```
> source("baltestcollect.r")
> xvars <- c("age","educ","black","hisp","married","re74","re75","u74","u75")
>
> outmatch1 <- Match(Y=nonexper$re78, Tr=nonexper$nsw,X= nonexper[,xvars],M=1,
+                  BiasAdjust=T, Weight=1)
> mb <- MatchBalance(nonexper$nsw ~ . - re78 - u78,data=nonexper,match.out=outmatch1)
> out.bef <- baltest.collect(matchbal.out=mb,var.names=colnames(nonexper[,xvars]),after=F
Minimum P value from T-Tests is 0
Minimum P value from KS-Tests is 0
Max qq.max.diff 0.7736242
> round(out.bef,3)
          mean.Tr   mean.Co    sdiff sdiff.pooled var.ratio T pval KS pval qqmeandiff qqme
age        25.816    34.851 -126.266     -121.683     0.470  0.000       0      0.232
educ       10.346    12.117  -88.077      -82.018     0.425  0.000       0      0.109
black       0.843     0.251  162.564      184.373     0.707  0.000      NA      0.296
hisp        0.059     0.033   11.357       20.894     1.786  0.132      NA      0.013
married     0.189     0.866 -172.406     -248.886     1.331  0.000      NA      0.339
re74     2095.574 19428.746 -354.707     -178.635     0.133  0.000       0      0.468
re75     1532.056 19063.338 -544.576     -178.652     0.056  0.000       0      0.469
u74         0.708     0.086  136.391      261.975     2.633  0.000      NA      0.311
u75         0.600     0.100  101.786      207.158     2.680  0.000      NA      0.250
```

Relative to the control group, the randomly assigned treatment group is comprised of people who are:
Younger
Less educated
Higher proportion black and hispanic
Lower proportion married
Lower income
Higher proportion unemployed
It looks like all of these covariates (age, education, race, earnings, and employment) are important

in determining selection into the program.

(d) Using the non-experimental data, estimate the (naive) ATE of the program on 1978 earnings without adjusting for any of the covariates. Report the estimate and a standard error.

```
> nonexp.unadj <- lm(formula = re78~nsw,data=nonexper)
> coeftest(nonexp.unadj,vcov = vcovHC(nonexp.unadj,"HC2"))

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  21553.92     311.73  69.143 < 2.2e-16 ***
nsw         -15204.78     657.08 -23.140 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(e) Still using the non-experimental data, use OLS to estimate the ATE controlling for age, education, race/ethnicity, marital status, and employment and earnings in 1974 and 1975. Report the estimate and its SE and compare these results to those in part (d). Did your point estimate change? Why?

```
> nonexp.adj <- lm(re78~nsw+age+educ+black+hisp+married+re74+re75+u74+u75,data=nonexper)
> coeftest(nonexp.adj,vcov = vcovHC(nonexp.adj,"HC2"))

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  9.5360e+02  1.5055e+03  0.6334 0.5265077
nsw          1.1538e+02  8.3417e+02  0.1383 0.8899973
age         -8.9765e+01  2.3383e+01 -3.8390 0.0001264 ***
educ         5.1412e+02  9.3022e+01  5.5269 3.575e-08 ***
black       -4.5422e+02  4.4663e+02 -1.0170 0.3092477
hisp         2.1974e+03  1.2381e+03  1.7748 0.0760505 .
married      1.2048e+03  4.9706e+02  2.4238 0.0154245 *
re74         3.1262e-01  6.2146e-02  5.0304 5.219e-07 ***
re75         5.4365e-01  6.8972e-02  7.8823 4.646e-15 ***
u74          2.3895e+03  1.3664e+03  1.7488 0.0804357 .
u75         -1.4620e+03  1.4192e+03 -1.0301 0.3030461
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The results changed substantially because in this dataset, there is a strong correlation between

the treatment and all of the covariates, as well as a relationship between those covariates and the response variable.

(f) Using the non-experimental data, condition (only) on the marital status of individuals, and manually compute the subclassification estimator of the ATT.

```
> table(nonexper$nsw,nonexper$married)

      0    1
  0  333 2157
  1  150   35
> mean(nonexper$re78[nonexper$married==1 & nonexper$nsw==1]) -
+     mean(nonexper$re78[nonexper$married==1 & nonexper$nsw==0])
[1] -14600.1
> mean(nonexper$re78[nonexper$married==0 & nonexper$nsw==1]) -
+     mean(nonexper$re78[nonexper$married==0 & nonexper$nsw==0])
[1] -10313.41
> (-14600.1*(35/185))+(-10313.41*(150/185))
[1] -11124.41
```

(g) Now write your own function in R that takes a dataset, a dependent variable, a treatment, and a vector of discrete covariates, and then computes and returns the subclassification point estimate of the ATT. What is the estimated ATT when conditioning on marriage status, race, unemployment status in 1975, and reported earnings in 1975? For this question, you should make earnings discrete by binning it by tercile (i.e. make it a discrete variable with 3 levels, which indicate the tercile).

```
subclass <- function(data, dv, treatment, covars, Estimand){
  data$group <- with(data, interaction(data[,covars]))
  k <- length(unique(data$group))
  lstrata <- unique(data$group)
  strata.tau.t <- NA; strata.tau.c <- NA; strata.tau <- NA; strata.weights <- NA
  for (i in 1:k){
    strata.tau.t[i] <- mean(data[data[,treatment] == 1 & data[,"group"] == lstrata[i],dv]
    strata.tau.c[i] <- mean(data[data[,treatment] == 0 & data[,"group"] == lstrata[i],dv]
    strata.tau[i] <- strata.tau.t[i] - strata.tau.c[i]
  }

  if (Estimand == "ATE"){
    n <- nrow(data)
    for (i in 1:k){
```

```
      strata.weights[i] <- sum(data$group == lstrata[i])/n
    }

    if (sum(is.na(strata.tau.t) != is.na(strata.tau.c)) > 0){
      stop("Insufficient Overlap!")
    }

  }

  if (Estimand == "ATT"){
    n <- nrow(data[data[,treatment] == 1,])
    for (i in 1:k){
      strata.weights[i] <- sum(data[data[,treatment] == 1,"group"] == lstrata[i])/n
    }

    if (sum(is.na(strata.tau.t) == FALSE & is.na(strata.tau.c) == TRUE) > 0){
      stop("Insufficient Overlap!")
    }
  }

  strata.tau[is.na(strata.tau)] <- 0
  tau <- strata.tau%*%strata.weights
  return(tau)
}
```

```
> data <- nonexper
>
> cuts<-quantile(data$re75, probs=c(.3333333,.6666666))
> data$earn75<-NA
> data$earn75[data$re75<cuts[1]]<-1
> data$earn75[data$re75>=cuts[1] & data$re75<cuts[2] ]<-2
> data$earn75[data$re75>=cuts[2] ]<-3
>
> subclass(data=data, covars=c("black","hisp","married","u75","earn75"),
 dv="re78", treatment="nsw", Estimand = "ATT")
```

```
          [,1]
[1,] 757.5707
```

(h) When would it be the case that the ATT is not identified (and hence cannot be computed)? If you have not already done so, build into the function you just wrote a safeguard whereby the function will break and return a warning message in such a case. Test this by conditioning on education (including all levels of education in the data) in addition to the variables you conditioned on in part (g).

This will be the case if there are any strata that contain treated units but no control units. If there are strata that contain control units but no treated units, this does not pose a problem for ATT identification (but it would for ATE identification).

```
> subclass(data=data, covars=c("black","hisp","married","u75","earn75","educ"), dv="re78"
Error in subclass(data = data, covars = c("black", "hisp", "married",  :
  Insufficient Overlap!
```

Problem 2

In this problem, we will continue to use the non-experimental data (`nsw_psid_withtreated.dta`) and implement matching estimators.

(a) With the non-experimental data (`nsw_psid_withtreated.dta`), use the following covariates to match on: "age", "educ", "black", "hisp", "married", "re74", "re75", "u74", and "u75". Use the matching approach, with one match per treated unit and using normalized Euclidean distance (default in the `Match()` function) as your distance metric, to estimate the ATT—the average effect of the program (`nsw`) on earnings (`re78`) for those who are treated. Include bias adjustment.

```
> xvars <- c("age","educ","black","hisp","married","re74","re75","u74","u75")
>
> outmatch1 <- Match(Y=nonexper$re78, Tr=nonexper$nsw,X= nonexper[,xvars],M=1,
+                  BiasAdjust=T, Weight=1)
> summary(outmatch1)

Estimate...  2415.5
AI SE......  1684.4
T-stat.....  1.434
p.val......  0.15156

Original number of observations..............  2675
```

```
Original number of treated obs.....  185
Matched number of observations.....  185
Matched number of observations  (unweighted).  218
```

(b) Using the matched data, report the balance statistics for the covariates that you matched on. How does the balance look in the matched data compared to the balance in the full non-experimental data set?

```
> mb <- MatchBalance(nonexper$nsw ~ . - re78 - u78,data=nonexper,match.out=outmatch1)
> out.bef <- baltest.collect(matchbal.out=mb,var.names=colnames(nonexper[,xvars]),after=FA
Minimum P value from T-Tests is 0
Minimum P value from KS-Tests is 0
Max qq.max.diff 0.7736242
> out.aft <- baltest.collect(matchbal.out=mb,var.names=colnames(nonexper[,xvars]),after=TR
Minimum P value from T-Tests is 0.001022885
Minimum P value from KS-Tests is 0
Max qq.max.diff 0.2201835
>
> round(out.bef,3)
          mean.Tr   mean.Co    sdiff sdiff.pooled var.ratio T pval KS pval qqmeandiff qqmed
age        25.816    34.851 -126.266     -121.683     0.470  0.000       0      0.232     0
educ       10.346    12.117  -88.077      -82.018     0.425  0.000       0      0.109     0
black       0.843     0.251  162.564      184.373     0.707  0.000      NA      0.296     0
hisp        0.059     0.033   11.357       20.894     1.786  0.132      NA      0.013     0
married     0.189     0.866 -172.406     -248.886     1.331  0.000      NA      0.339     0
re74     2095.574 19428.746 -354.707     -178.635     0.133  0.000       0      0.468     0
re75     1532.056 19063.338 -544.576     -178.652     0.056  0.000       0      0.469     0
u74         0.708     0.086  136.391      261.975     2.633  0.000      NA      0.311     0
u75         0.600     0.100  101.786      207.158     2.680  0.000      NA      0.250     0
> round(out.aft,3)
          mean.Tr  mean.Co   sdiff sdiff.pooled var.ratio T pval KS pval qqmeandiff qqmeddi
age        25.816   26.077  -3.651       -3.651     0.924  0.601   0.000      0.059     0.0
educ       10.346   10.652 -15.234      -15.234     1.312  0.002   0.000      0.034     0.0
black       0.843    0.843   0.000        0.000     1.000  1.000      NA      0.000     0.0
hisp        0.059    0.059   0.000        0.000     1.000  1.000      NA      0.000     0.0
married     0.189    0.189   0.000        0.000     1.000  1.000      NA      0.000     0.0
re74     2095.574 2173.748  -1.600       -1.600     0.973  0.433   0.502      0.013     0.0
re75     1532.056 2095.316 -17.497      -17.497     0.745  0.001   0.092      0.043     0.0
u74         0.708    0.708   0.000        0.000     1.000  1.000      NA      0.000     0.0
```

```
u75          0.600     0.600    0.000          0.000      1.000  1.000       NA       0.000      0.
```

(c) Re-estimate the ATT using the same matching approach as in part (a) except do not include bias adjustment, and compare the results. What is the importance of the bias adjustment? When is it most important to use the bias adjustment?

```
> outmatch1.unadj <- Match(Y=nonexper$re78, Tr=nonexper$nsw,X= nonexper[,xvars],M=1,
+ BiasAdjust=F, Weight=1)
> summary(outmatch1.unadj)

Estimate...  2073.5
AI SE......  1678.6
T-stat.....  1.2353
p.val......  0.21673

Original number of observations..............  2675
Original number of treated obs...............  185
Matched number of observations...............  185
Matched number of observations  (unweighted).  218
```

For each treated unit, you need to find a similar untreated unit. If there is not an exact match, then the untreated unit you pick will have slightly different x's than the treated unit. Some of the difference in y may then be due not to the treatment, but to differences in the x's. So you need to account for this as well as possible. If you don't, there can be a bias that does not go away as N grows large. This is increasingly important as the number of dimensions on which you (non-exact-match on) grows. However the problem is alleviated if you have many times more controls than treated units, since it will improve your chances of finding a suitable control unit.

(d) Compare the number of treated to untreated units. Comment on the result, and whether it is good or bad in your view.

```
> sum(nonexper$nsw==0)
[1] 2490
> sum(nonexper$nsw==1)
[1] 185
```

Given our focus on the ATT, it is good that we have many more control units than treated units (conditional on the number of treated units), because that increases our chances of finding good matches for each treated unit.

(e) Now, match again on the same covariates, but this time use genetic matching. Report your estimate of the ATT and its standard error. You may want to read up on genetic matching at: http:

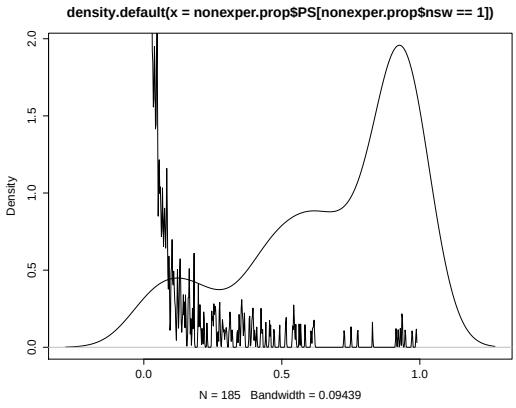
//sekhon.berkeley.edu/papers/GenMatch.pdf. Use GenMatch to obtain the weight matrix, which you then pass to Match.

```
> library(rgenoud)
> genout <- GenMatch(Tr=nonexper$nsw,X=nonexper[,xvars],
+ BalanceMatrix=nonexper[,xvars],estimand="ATT",pop.size=200)
> match.gen <- Match(Y=nonexper$re78,Tr=nonexper$nsw, X=nonexper[,xvars],M=1,estimand="AT
> summary(match.gen)

Estimate...  2024.8
AI SE......  1606.5
T-stat.....  1.2604
p.val......  0.20753

Original number of observations..............  2675
Original number of treated obs...............  185
Matched number of observations...............  185
Matched number of observations  (unweighted).  220
```

(f) Estimate propensity scores using a logistic regression with the non-experimental data (the full set, not the matched set). Plot the distributions of propensity scores for treated and control groups and comment on the overlap.

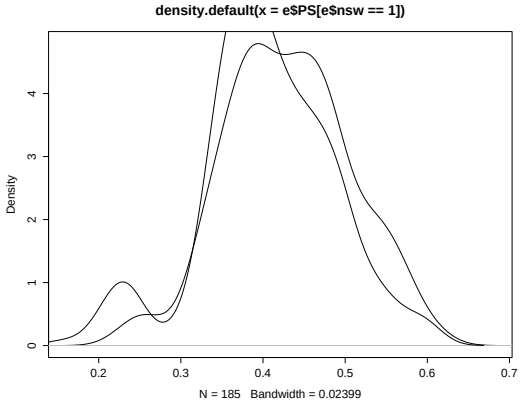
```
> #Pscore with nonexp
> nonexper.prop <- nonexper[,c("re78","nsw",xvars)]
> #d=na.omit(d)
> psout <- glm(nsw~age+educ+black+hisp+married+re74+re75+u74+u75,family=binomial,data=none
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(psout)

Call:
glm(formula = nsw ~ age + educ + black + hisp + married + re74 +
    re75 + u74 + u75, family = binomial, data = nonexper)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9490  -0.0935  -0.0178  -0.0032   4.1200
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.795e+00  9.793e-01   1.833 0.066870 .
age         -1.109e-01  1.771e-02  -6.263 3.78e-10 ***
educ        -1.009e-01  5.611e-02  -1.798 0.072208 .
black        2.650e+00  3.606e-01   7.350 1.98e-13 ***
hisp         2.248e+00  5.909e-01   3.804 0.000142 ***
married     -1.561e+00  2.818e-01  -5.538 3.06e-08 ***
re74         2.018e-05  3.131e-05   0.644 0.519307
re75        -2.743e-04  4.771e-05  -5.750 8.92e-09 ***
u74          3.272e+00  4.888e-01   6.695 2.15e-11 ***
u75         -1.371e+00  4.546e-01  -3.017 0.002554 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1345.30  on 2674  degrees of freedom
Residual deviance:  418.78  on 2665  degrees of freedom
AIC: 438.78

Number of Fisher Scoring iterations: 10

> nonexper.prop$PS <- psout$fitted
```



(g) Now, use the experimental data and estimate propensity scores using the same model. Again, plot and compare the distributions of the propensity scores for treated and control groups here. What do you observe? Compare your results with part (f).

```
> #Pscore with exp
> e <- exper[,c("nsw",xvars)]
> #d=na.omit(d)
> psout <- glm(nsw~age+educ+black+hisp+married+re74+re75+u74+u75,family=binomial,data=exp
> summary(psout)

Call:
glm(formula = nsw ~ age + educ + black + hisp + married + re74 +
    re75 + u74 + u75, family = binomial, data = exper)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3614  -1.0373  -0.9156   1.2623   1.7097

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.736e-01  8.244e-01  -0.575   0.5656
age          1.426e-02  1.421e-02   1.004   0.3156
educ         4.998e-02  5.641e-02   0.886   0.3756
black       -3.477e-01  3.607e-01  -0.964   0.3351
hisp        -9.285e-01  5.066e-01  -1.833   0.0668 .
married      1.760e-01  2.749e-01   0.640   0.5219
re74        -3.393e-05  2.926e-05  -1.160   0.2462
re75         1.221e-05  4.714e-05   0.259   0.7956
u74         -1.516e-01  3.716e-01  -0.408   0.6833
u75         -3.719e-01  3.177e-01  -1.171   0.2417
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 604.2  on 444  degrees of freedom
Residual deviance: 592.5  on 435  degrees of freedom
AIC: 612.5

Number of Fisher Scoring iterations: 4
> e$PS <- psout$fitted
```



(h) Back to the non-experimental dataset: trim off observations that have propensity scores lower than 0.1. Then report balance statistics for the trimmed data. How does the balance look in this trimmed dataset compared to the balance in the full dataset? Do results differ? Why or why not?

```
> # treated vs control with trimming
> nonexper.prop.trim <- subset(nonexper.prop, nonexper.prop$PS >= 0.1)
> sum(nonexper.prop.trim$nsw==0)
[1] 136
> sum(nonexper.prop.trim$nsw==1)
[1] 173
> xvars <- c("age","educ","black","hisp","married","re74","re75","u74","u75")
>
> outmatch.trim <- Match(Y=nonexper.prop.trim$re78, Tr=nonexper.prop.trim$nsw,
+                        X= nonexper.prop.trim[,xvars],M=1,
+                    BiasAdjust=T, Weight=1)
> mb.trim <- MatchBalance(nsw ~ . - re78 - PS,data=nonexper.prop.trim,match.out=outmatch.
> out.bef.trim <- baltest.collect(matchbal.out=mb.trim,var.names=colnames(nonexper.prop.t
Minimum P value from T-Tests is 2.049063e-09
Minimum P value from KS-Tests is 0
Max qq.max.diff 0.3612292
>
> round(out.bef.trim,3)
         mean.Tr  mean.Co    sdiff sdiff.pooled var.ratio T pval KS pval qqmeandiff qqmed
age       25.538   29.382  -53.688      -59.132     0.430  0.000   0.030      0.098     0
educ      10.260   10.331   -3.540       -4.049     0.452  0.812   0.008      0.054     0
black      0.861    0.787   21.492       27.907     0.711  0.092      NA      0.037     0
hisp       0.058    0.081   -9.861      -12.947     0.731  0.434      NA      0.012     0
```

```
married    0.145    0.441  -84.131      -93.801     0.501  0.000      NA      0.148     0
re74    1521.938 4319.117  -75.370      -82.370     0.449  0.000   0.000      0.203     0
re75    1051.119 2925.637 -100.782      -99.577     0.354  0.000   0.000      0.231     0
u74        0.746    0.412   76.451       96.181     0.782  0.000      NA      0.167     0
u75        0.624    0.375   51.322       70.422     0.999  0.000      NA      0.125     0
> round(out.bef,3)
         mean.Tr   mean.Co    sdiff sdiff.pooled var.ratio T pval KS pval qqmeandiff qqme
age       25.816    34.851 -126.266     -121.683     0.470  0.000       0      0.232     (
educ      10.346    12.117  -88.077      -82.018     0.425  0.000       0      0.109     (
black      0.843     0.251  162.564       184.373     0.707  0.000      NA      0.296     (
hisp       0.059     0.033   11.357        20.894     1.786  0.132      NA      0.013     (
married    0.189     0.866 -172.406     -248.886     1.331  0.000      NA      0.339     (
re74    2095.574 19428.746 -354.707     -178.635     0.133  0.000       0      0.468     (
re75    1532.056 19063.338 -544.576     -178.652     0.056  0.000       0      0.469     (
u74        0.708     0.086  136.391       261.975     2.633  0.000      NA      0.311     (
u75        0.600     0.100  101.786       207.158     2.680  0.000      NA      0.250     (
```

The balance is much better in the trimmed dataset, because eliminating observations with low propensity scores likely got rid of many control observations with covariate values not representative of the treated set, while preserving most of the distribution of the treated set. The result is better balanced treated and control sets.

(i) Now, match on the estimated propensity scores from part (f) to estimate the ATT. Report your point estimate and standard error.

```
> head(nonexper.prop)
  re78 nsw age educ black hisp married re74 re75 u74 u75          PS
1    0   0  47   12     0    0       0    0    0   1   1 0.061327297
2    0   0  50   12     1    0       1    0    0   1   1 0.122224232
3    0   0  44   12     0    0       0    0    0   1   1 0.083518015
4    0   0  28   12     1    0       1    0    0   1   1 0.615080528
5    0   0  54   12     0    0       1    0    0   1   1 0.006272343
6    0   0  55   12     0    1       1    0    0   1   1 0.050763174
> outmatch.ps <- Match(Y=nonexper.prop$re78, Tr=nonexper.prop$nsw,X= nonexper.prop$PS,M=1
+                 BiasAdjust=T, Weight=1)
> summary(outmatch.ps)

Estimate...  2145.8
AI SE......  1554
```

```
T-stat.....  1.3808
p.val......  0.16733

Original number of observations..............  2675
Original number of treated obs...............  185
Matched number of observations...............  185
Matched number of observations  (unweighted).  1997
```

(j) How closely were you able to replicate the experimental results using matching estimators with the non-experimental data? Which version of your matching estimator seemed to perform the best?

Problem 3

The curse of dimensionality makes it difficult to work in high-dimensional covariate spaces. Here we will see for ourselves what happens as the number of covariates grows large. Suppose you have covariates $X_1, X_2, \ldots, X_P$ where P is the number of dimensions of your data (i.e. the number of covariates). Each observation's x_i will be a (column) vector describing the covariate values for observation i. That is, $x_i = [x_{i,1}, x_{i,2}, \ldots x_{i,P}]^T$.

a) Write a mathematical expression that gives the Euclidean distance between observations i and j, with covariates x_i and x_j respectively.

If $x_i = [x_{1i}x_{2i}...x_{ni}]$ and $x_j = [x_{1j}x_{2j}...x_{nj}]$, then the Euclidean distance is defined as $\sqrt{\sum_{k=1}^{n}(x_{ki} - x_{kj})^2}$

b) Suppose you have many covariates, each one uniformly distributed on the interval $[0, 1]$. You want to draw a square (or cube, or hypercube) in the covariate space that in expectation would capture 5% of the observations. If you have a 2-dimensional covariate space, how large would the sides of the square have to be to include 5% of the observations? Now consider a 3-dimensional covariate space. Again, if you want to capture 5% of the observations, how long would the side of that cube have to be?

In the 2D case, let the side-length be r. We want $r^2 = 0.05$ (note the observations are uniformly distributed in the covariate space). Thus $r = (0.05)^{1/2}$. Similarly, in the 3D case, we want $r^3 = 0.05$. Thus $r = (0.05)^{1/3}$

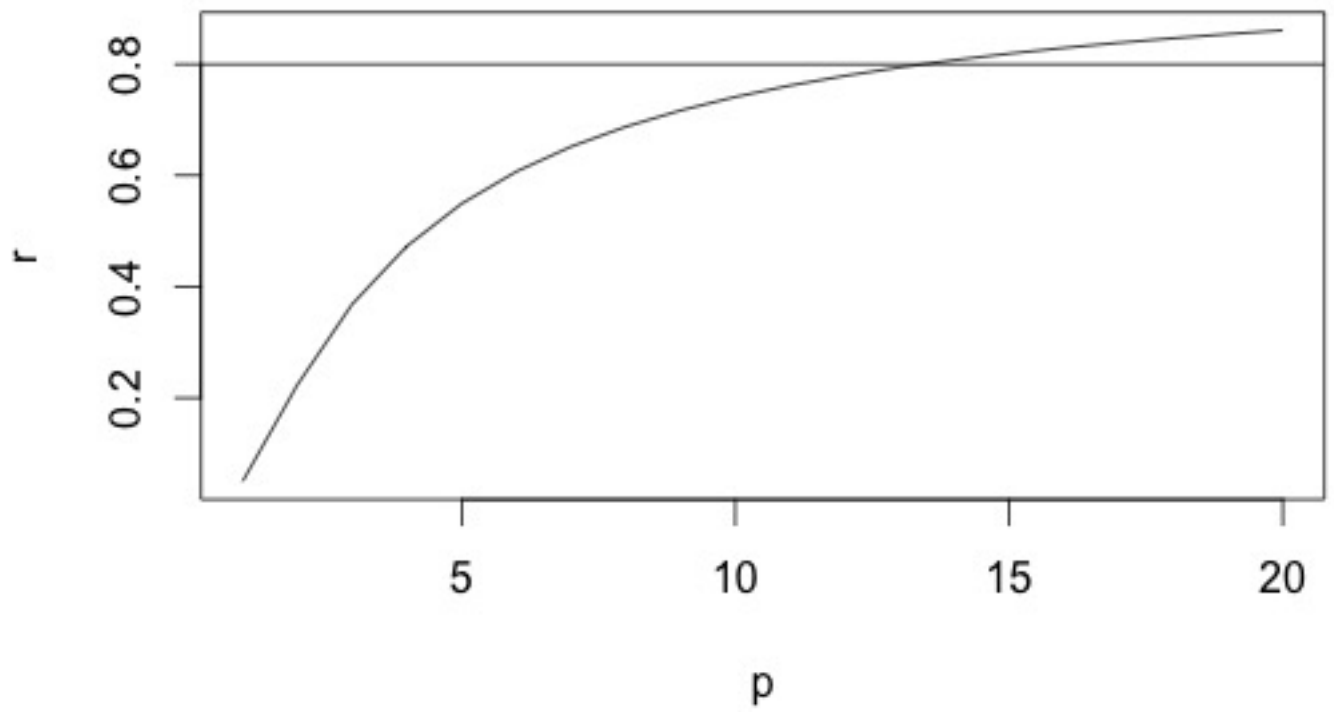
c) Derive a general expression that gives the side-length, r, that you would need for a (hyper)cube to capture a percentage of the data, v, when the number of dimensions is p. Create a graph that shows the r needed to capture $v = 0.05$ of the total observations as the number of dimensions goes from 1 to 20. What do you notice? What does this tell you about matching in high-dimensional

spaces?

$r^p = v$, thus $r = v^{1/p}$

The side-length needed to capture a fixed volume of the data as the number of dimensions increases grows very quickly, though it is bounded by 1 since that is the maximum length. This shows that in high dimensional spaces, you have to go a long distance across the covariate space to capture even a small volume of the nearest observations.

```
> p<-1:20; v<-0.5; r<-v^(1/p)
> plot(p,r)
```



d) Create a simulated dataset, X. It should have 100 observations, and 20 dimensions. Each dimension of X should be drawn from $unif(-1, 1)$. We will use this as the base dataset, but will not always use all of the dimensions. Consider a single point, $x^\star$, whose covariates are all equal to zero. You want to find this point's closest neighbor in the dataset.

(i) Start with a one dimensional version of X, using just the first dimension of the simulated dataset that you made. Find the point in the dataset i whose covariate value x_i is closest to $x^\star$ in a Euclidean sense. Record the (Euclidean) distance from this point x_i to $x^\star$,

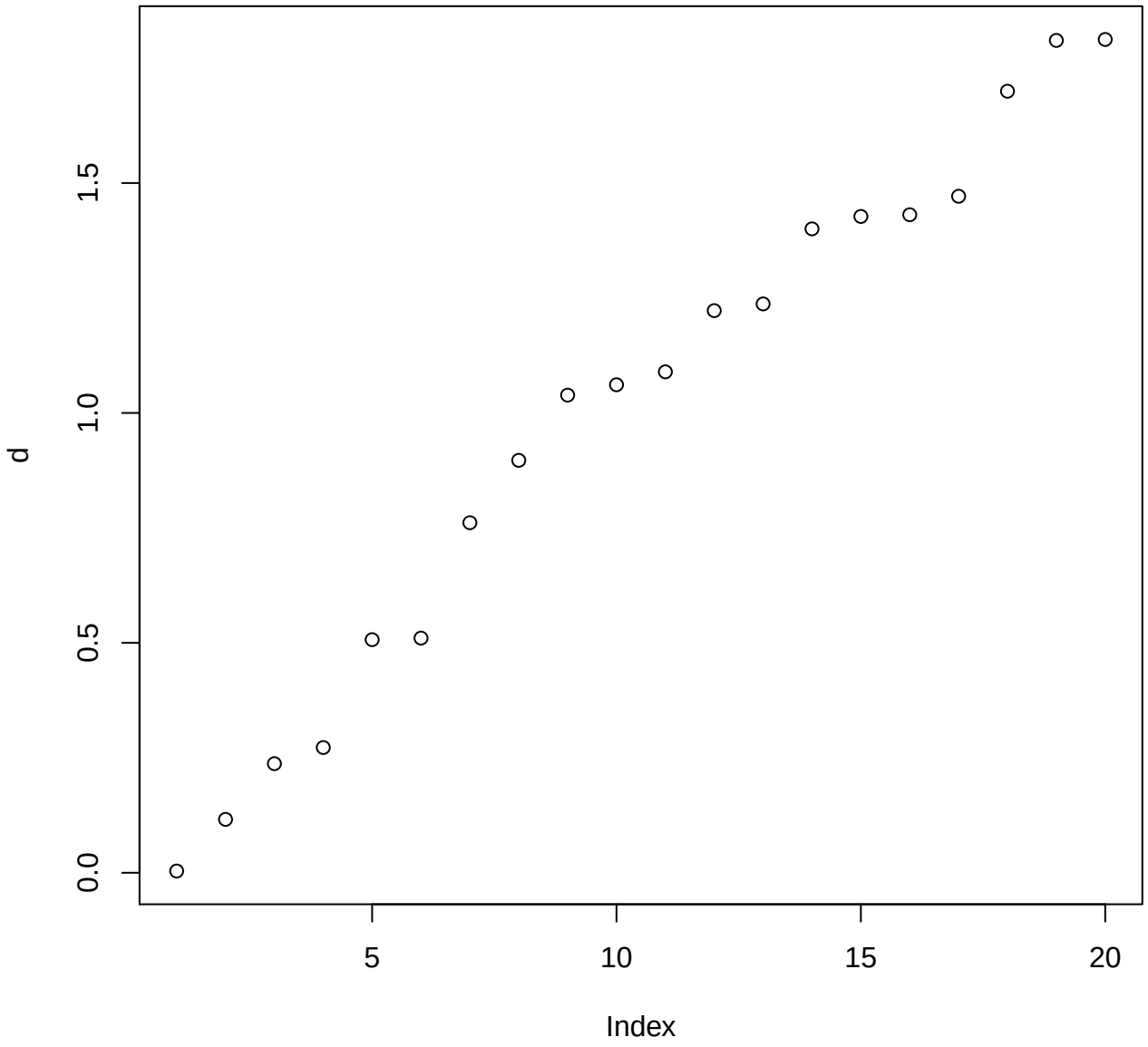
(ii) Repeat the step above, but using 2 dimensions, then 3, and so on up to all 20 dimensions. In each case, get the Euclidean distance between the x_i that was the nearest neighbor and $x^\star$.

(iii) Plot the Euclidean distance as a function of the number of dimensions.

(iv) What do you conclude from this? What does this teach us about matching?

The distance to the nearest neighbor increases linearly with the number of dimensions (It can be proven that with variables of equal variance, as long as the x_i's are chosen i.i.d., the Euclidean distance grows linearly with the dimensions regardless of covariance of the different variables). This suggests that matching may be increasingly ineffective as the dimensionality increases.

```
> library(fields)
> d<-c()
> x<-matrix(NA,nrow=100,ncol=20)
> for(i in 1:20){
+   x[,i]<-runif(100,-1,1)
+   euc<-c()
+   for(j in 1:100){  euc[j]<-sqrt(sum((x[j,1:i] - 0) ^ 2)) }
+   d[i]<-min(euc)
+ }
> plot(d)
```


d
0.0
0.5
1.0
1.5
5
10
15
20
Index