Gelman & Hill Ch 10 Ex 1: NSW Lalonde

Exploratory analysis of the Lalonde / Dehejia data set.

```
#devtools::install_github("jjchern/lalonde")
#devtools::install_git("https://gitlab.nza.nl/GertjanV/lalonde")

library(lalonde)
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.2.
```

```
## v tibble  1.4.2     v purrr   0.2.5
## v tidyr   0.8.1     v dplyr   0.7.6
## v readr   1.1.1     v stringr 1.3.1
## v tibble  1.4.2     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts(
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(data.table)
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

Introduction

Constructed observational studies: the folder lalonde contains data from an observational study constructed by LaLonde (1986) based on a randomized ex- periment that evaluated the effect on earnings of a job training program called National Supported Work. The constructed observational study was formed by replacing the randomized control group with a comparison group formed using data from two national public-use surveys: the Current Population Survey (CPS) and the Panel Study in Income Dynamics.

The training program ran in 1976-1977. earnings 1978 is the outcome measure. Ppl that enrolled before jan 1976, or were still in the program in jan 1978 were excluded. treat is treatment, rest is pre-treatment.

Dehejia and Wahba (1999) used a subsample of these data to evaluate the po- tential efficacy of propensity score matching. The subsample they chose removes men for whom only one pre-treatment measure of earnings is observed. (There is substantial evidence in the economics literature that controlling for earnings from only one pre-treatment period is insufficient to satisfy ignorability.) This exercise replicates some of Dehejia and Wahbas findings based on the CPS comparison group.

Exercise

(a) Estimate the treatment effect from the experimental data in two ways: (i) a simple difference in means between treated and control units, and (ii) a regression-adjusted estimate (that is, a regression of outcomes on the treat- ment indicator as well as predictors corresponding to the pre-treatment characteristics measured in the study).

Lalonde RCT: male participants

```
# lalonde sample (RCT, male participants, no re74)
#nsw

table(nsw$treat)
```

```
##
##   0   1
## 425 297
```

```
nsw <-data.table(nsw)
nsw[, .(mean(re78)), .(treat)]
```

```
##    treat       V1
## 1:     1 5976.352
## 2:     0 5090.048
```

Difference of ~ $800.

Regression, unadjusted score

```
lmfit <- lm(re78 ~ treat, data = nsw)

summary(lmfit)
```

```
##
## Call:
## lm(formula = re78 ~ treat, data = nsw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5976  -5090  -1519   3361  54332
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5090.0      302.8  16.811   <2e-16 ***
## treat          886.3      472.1   1.877   0.0609 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6242 on 720 degrees of freedom
## Multiple R-squared:  0.004872,   Adjusted R-squared:  0.003489
## F-statistic: 3.525 on 1 and 720 DF,  p-value: 0.06086
```

Compare with Dehejia Table 2. Perfect match.

Regression: adjusted score

```r
lmfit <- lm(re78 ~ treat +  age + I(age^2) + education + black +
              hispanic + nodegree, data = nsw)

summary(lmfit)
```

```
##
## Call:
## lm(formula = re78 ~ treat + age + I(age^2) + education + black +
##     hispanic + nodegree, data = nsw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -8369  -4667  -1515   3225  54610
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4430.1626  3653.2238   1.213   0.2257
## treat         798.3512   472.1283   1.691   0.0913 .
## age            -3.8055   211.1663  -0.018   0.9856
## I(age^2)        0.5297     3.5562   0.149   0.8816
## education     219.7946   182.9296   1.202   0.2299
## black       -1762.8326   803.8800  -2.193   0.0286 *
## hispanic     -117.1480  1054.2282  -0.111   0.9116
## nodegree     -494.2816   749.2561  -0.660   0.5097
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6208 on 714 degrees of freedom
## Multiple R-squared:  0.02378,    Adjusted R-squared:  0.01421
## F-statistic: 2.484 on 7 and 714 DF,  p-value: 0.0159
```

We estimate treatment effect at +798 dollar. Significant at 10% level. Uncertainty is large! Low power.

Compare with Table 2. Exact match.

Dehejia-Wahba RCT + subset RE74

Added 1974 earnings, subset on obs with this var.

unadjusted score

```r
# Dehejia-Wahba Sample (male participants, with re74 --> reduces #obs)
table(nsw_dw$treat)
```

```
##
##   0   1
## 260 185
```

```r
lmfit <- lm(re78 ~ treat, data = nsw_dw)

summary(lmfit)
```

```
##
## Call:
## lm(formula = re78 ~ treat, data = nsw_dw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6349  -4555  -1829   2917  53959
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4554.8      408.0  11.162  < 2e-16 ***
## treat         1794.3      632.9   2.835  0.00479 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6580 on 443 degrees of freedom
## Multiple R-squared:  0.01782,    Adjusted R-squared:  0.01561
## F-statistic: 8.039 on 1 and 443 DF,  p-value: 0.004788
```

The so-called "Benchmark unbiased treatment effect" is $1794.

Dehejia & Wahba show that it is possible to get close to this number using observational data & propensity scores.

adjusted

```r
lmfit <- lm(re78 ~ treat +  age + I(age^2) + education + black +
              hispanic + nodegree +  re74, data = nsw_dw)

summary(lmfit)
```

```
##
## Call:
## lm(formula = re78 ~ treat + age + I(age^2) + education + black +
##     hispanic + nodegree + re74, data = nsw_dw)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10098  -4422  -1669   2926  54060
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.139e+02  4.668e+03  -0.067  0.94642
## treat        1.688e+03  6.360e+02   2.655  0.00823 **
## age          1.395e+02  2.675e+02   0.522  0.60220
## I(age^2)    -1.407e+00  4.412e+00  -0.319  0.75002
## education    3.884e+02  2.286e+02   1.699  0.09011 .
## black       -2.188e+03  1.165e+03  -1.879  0.06092 .
## hispanic     1.858e+02  1.545e+03   0.120  0.90434
## nodegree    -1.432e+01  9.939e+02  -0.014  0.98851
```

```
## re74         9.942e-02  5.802e-02   1.713  0.08735 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6506 on 436 degrees of freedom
## Multiple R-squared:  0.05471,    Adjusted R-squared:  0.03736
## F-statistic: 3.154 on 8 and 436 DF,  p-value: 0.001749
```

Adjusted RCT treatment effect on re74 subset: +1688 dollar. Apparently the adjusters do not add precision. The sample is already pretty well balanced.

Regression on constructed dataset

(b) Now use a regression analysis to estimate the causal effect from Dehejia and Wahba s subset of the constructed observational study. Examine the sensitiv- ity of the model to model specification (for instance, by excluding the em- ployed indicator variables or by including interactions). How close are these estimates to the experimental benchmark?

Create dataset using CPS controls.

```r
df_constr <- lalonde::nsw_dw %>%
    filter(treat == 1) %>%
    bind_rows(lalonde::cps_controls)
```

CHeck whats up with all these zero earnings

```r
df_constr <- df_constr %>% mutate(has_re74 = ifelse(re74 > 0, 1, 0)) %>%
  mutate(has_re75 = ifelse(re75 > 0, 1, 0)) %>%
  mutate(has_re78 = ifelse(re78 > 0, 1, 0))
```

```r
table(df_constr$has_re74, df_constr$has_re75)
```

```
##
##         0     1
##   0  1290   754
##   1   569 13564
```

```r
table(df_constr$has_re74, df_constr$has_re78)
```

```
##
##         0     1
##   0   913  1131
##   1  1304 12829
```

```r
table(df_constr$has_re75, df_constr$has_re78)
```

```
##
##         0     1
##   0  1014   845
##   1  1203 13115
```

Run regression to est treatment

```r
glm.fit <- glm (re78 ~ treat + age + I(age^2) + education + black + hispanic + married + nodegree + re7

summary(glm.fit)
```

```
##
## Call:
## glm(formula = re78 ~ treat + age + I(age^2) + education + black +
##     hispanic + married + nodegree + re74 + re75, data = df_constr)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -25130   -3601    1274    3668   55040
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7634.34415  736.67074  10.363  < 2e-16 ***
## treat        793.58704  548.25433   1.447  0.14778
## age         -233.67749   41.18067  -5.674 1.42e-08 ***
## I(age^2)       1.81437    0.56099   3.234  0.00122 **
## education    166.84923   28.65984   5.822 5.94e-09 ***
## black       -790.60856  213.24523  -3.708  0.00021 ***
## hispanic    -175.97512  218.99126  -0.804  0.42166
## married      224.26599  149.84542   1.497  0.13450
## nodegree     311.84453  178.51743   1.747  0.08068 .
## re74           0.29534    0.01222  24.175  < 2e-16 ***
## re75           0.47064    0.01216  38.700  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 49023575)
##
##     Null deviance: 1.5129e+12  on 16176  degrees of freedom
## Residual deviance: 7.9252e+11  on 16166  degrees of freedom
## AIC: 332381
##
## Number of Fisher Scoring iterations: 2
```

```r
glm.fit2 <- glm (re78 ~ treat + age + I(age^2) + education + black + hispanic + married + nodegree, data

summary(glm.fit2)
```

```
##
## Call:
## glm(formula = re78 ~ treat + age + I(age^2) + education + black +
##     hispanic + married + nodegree, data = df_constr)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -19493   -6916    2137    7637   53794
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -5875.1673   939.2882  -6.255 4.08e-10 ***
## treat       -3436.7947   710.2373  -4.839 1.32e-06 ***
## age           942.7597    51.1406  18.435  < 2e-16 ***
## I(age^2)      -12.2046     0.7045 -17.325  < 2e-16 ***
## education     219.9440    37.1831   5.915 3.38e-09 ***
## black       -2345.7758   276.3283  -8.489  < 2e-16 ***
## hispanic    -1070.4600   284.1377  -3.767 0.000166 ***
## married      3207.4879   191.1373  16.781  < 2e-16 ***
## nodegree    -1076.8590   231.2101  -4.657 3.23e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 82715640)
##
##     Null deviance: 1.5129e+12  on 16176  degrees of freedom
## Residual deviance: 1.3373e+12  on 16168  degrees of freedom
## AIC: 340841
##
## Number of Fisher Scoring iterations: 2
```

We also get a treatment effect of $800. Ouch! If we leave out the re74 and re75 we get a negative treatment effect of $-3400 So very sensitive.

Propensity scores on constructed dataset

(c) Now estimate the causal effect from the Dehejia and Wahba subset using propensity score matching. Do this by first trying several different specifica- tions for the propensity score model and choosing the one that you judge to yield the best balance on the most important covariates. Perform this propensity score modeling without looking at the estimated treat- ment effect that would arise from each of the resulting matching procedures. For the matched dataset you construct using your preferred model, report the estimated treatment effects using the differrence-in-means and regression- adjusted methods described in part (a) of this exercise. How close are these estimates to the experimental benchmark (about $1800)?

Fit Propensity score model.

```
ps.fit.1 <- glm (treat ~ age + I(age^2) + education + black + hispanic + married + nodegree + re74 + re

summary(ps.fit.1)
```

```
##
## Call:
## glm(formula = treat ~ age + I(age^2) + education + black + hispanic +
##     married + nodegree + re74 + re75, family = binomial(link = "logit"),
##     data = df_constr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7932  -0.0545  -0.0111  -0.0027   3.7912
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.498e+01  1.388e+00 -10.792  < 2e-16 ***
## age          7.257e-01  8.792e-02   8.254  < 2e-16 ***
```

```
## I(age^2)    -1.176e-02  1.490e-03  -7.896 2.89e-15 ***
## education   -1.679e-02  4.593e-02  -0.366    0.715
## black        3.936e+00  2.598e-01  15.152  < 2e-16 ***
## hispanic     1.590e+00  4.007e-01   3.967 7.26e-05 ***
## married     -1.438e+00  2.388e-01  -6.021 1.74e-09 ***
## nodegree     1.460e+00  2.775e-01   5.262 1.43e-07 ***
## re74        -6.394e-05  2.852e-05  -2.242    0.025 *
## re75        -2.177e-04  3.676e-05  -5.923 3.16e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2022.14  on 16176  degrees of freedom
## Residual deviance:  906.14  on 16167  degrees of freedom
## AIC: 926.14
##
## Number of Fisher Scoring iterations: 10
```

```r
df_constr$pscores <- predict (ps.fit.1, type="response")
```

Check if we can predict treatment for the RCT treated obs. This is of course a bit weird, because we have added a bunch of ctrl obs, that CAN be similar, but are not treated. So this lowers the probability of treatment.

```r
df_constr$treat_pred <- ifelse(df_constr$pscores >= 0.5, 1, 0)

table(df_constr[df_constr$data_id == "Dehejia-Wahba Sample",]$treat_pred,
      df_constr[df_constr$data_id == "Dehejia-Wahba Sample",]$treat)
```

```
##
##       1
##   0 142
##   1  43
```

So what matters is not the absolute prob of treatment (because this depends on the amount and type of control observations) but to match to make sure that a treat and obs data point have THE SAME relative prob of treatment.

Matching

Use propensity scores to create a matched control group. For each treatment obs, pick closest control obs based on pscore.

Compare matching algorithms. Simple function of Gelman vs Sekhon package.

```r
library(arm)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
## The following object is masked from 'package:tidyr':
##
##     expand
## Loading required package: lme4
##
## arm (Version 1.10-1, built: 2018-4-12)
## Working directory is Y:/Mijn Documenten/Gitlab/cursus_causal_learning/exercises/session5_gelman ch10
```

```
matches <- matching (z=df_constr$treat, score=df_constr$pscores)
matched <- df_constr[matches$matched,]
matched <- data.table(matched)
```

```
#do greedy matching on logit(PS) using Match with a caliper
library(Matching)
```

```
## ##
## ##  Matching (Version 4.9-3, Build Date: 2018-05-03)
## ##  See http://sekhon.berkeley.edu/matching for additional documentation.
## ##  Please cite software as:
## ##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
## ##   Software with Automated Balance Optimization: The Matching package for R.''
## ##   Journal of Statistical Software, 42(7): 1-52.
## ##
```

```
psmatch<-Match(Tr = df_constr$treat,M=1, X = df_constr$pscores,
               replace = FALSE, caliper = .2)

matched2 <- df_constr[unlist(psmatch[c("index.treated","index.control")]), ]

matched2 <- data.table(matched2)
```

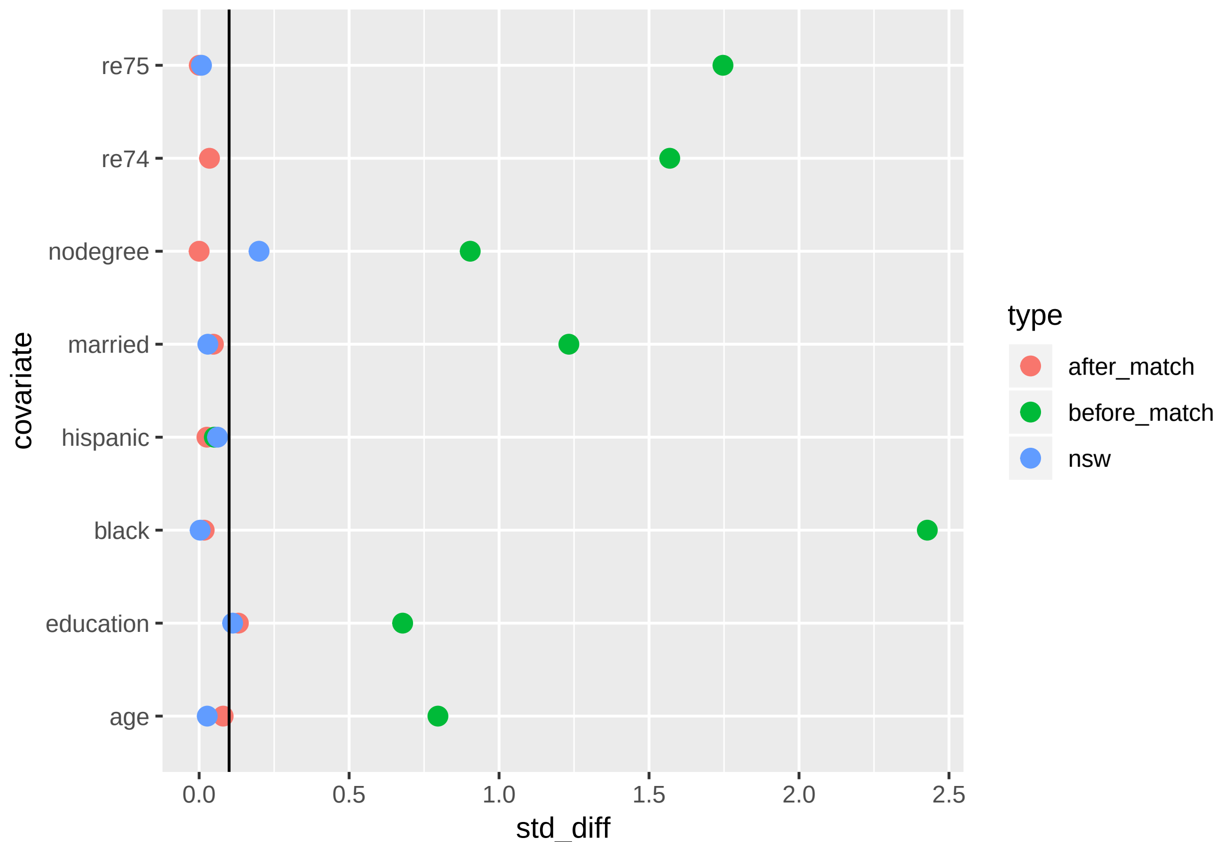
Sekhon package appears better.

Plot standardized differences in mean values between treat and control

From PC Austin (2009): Standardized differences are increasingly being used to compare balance in baseline covariates between treated and untreated subjects in the propensity-score matched sample. A limitation to their use is lack of consensus as to what value of a standardized difference denotes important residual imbalance between treated and untreated subjects in the matched sample. While there is no clear consensus on this issue, some researchers have proposed that a standardized difference of 0.1 (10 per cent) denotes meaningful imbalance in the baseline covariate

```
# this assumes a treatment variable called "treat", and performs the calculation on ALL variables not i
calc_abs_std_difference <- function(df, idvars){
  df <- data.table(df)
  mdf <- melt(df, id.vars = idvars)
  # for each variable, calc mean & sd by treat
  res <- mdf[, .(mu = mean(value), sd = sd(value)), .(treat, variable)]
```

```
  setnames(res, "variable", "covariate")
  mres <- melt(res, id.vars = c("treat", "covariate"))

  res <- dcast(mres, covariate ~ variable + treat)
  # calc abs std difference between treated / not treated for each variable
  res <- res[, std_diff := abs((mu_0 - mu_1)/sqrt(0.5*(sd_0^2 + sd_1^2)))]

  return(res[, .(covariate, std_diff, mu_0, mu_1)])
}

res <- calc_abs_std_difference(matched2, idvars = c("data_id", "treat"))
res$type <- "after_match"
res2 <- calc_abs_std_difference(df_constr, idvars = c("data_id", "treat"))
res2$type <- "before_match"
res3 <- calc_abs_std_difference(nsw, idvars = c("data_id", "treat"))
res3$type <- "nsw"

res <- rbind(res, res2, res3)

sel_vec <- c("age"    ,    "education",  "black"  ,    "hispanic"  , "married"   , "nodegree" ,       "
  
ggplot(res[covariate %in% sel_vec], aes(x = covariate, y = std_diff, col = type, group = type)) +
  geom_point(size = 3) + coord_flip() + geom_hline(yintercept = 0.1, col = "black")
```



Calculate sample means

```
matched2[, mean(re78), .(treat)]
```

```
##    treat       V1
## 1:     1 6367.703
## 2:     0 4985.908
```

```
matched[, mean(re78), .(treat)]
```

```
##    treat       V1
## 1:     0 4687.309
## 2:     1 6349.144
```

Fit model on matched data

```
lmfit <- lm(re78 ~ treat +  age + I(age^2) + education + black +
              hispanic + married + nodegree +  re74 + re75, data = matched)

summary(lmfit)
```

```
##
## Call:
## lm(formula = re78 ~ treat + age + I(age^2) + education + black +
##     hispanic + married + nodegree + re74 + re75, data = matched)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10907  -4786  -2105   3440  53688
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3856.9277  5635.0840  -0.684   0.4941
## treat        1603.0021   725.2223   2.210   0.0277 *
## age           231.6116   346.5260   0.668   0.5043
## I(age^2)       -3.6978     5.8463  -0.633   0.5275
## education     435.4154   198.0210   2.199   0.0285 *
## black        -756.8931  1182.3833  -0.640   0.5225
## hispanic      532.3462  1951.6774   0.273   0.7852
## married       186.3285  1001.8268   0.186   0.8526
## nodegree      921.3390  1102.1311   0.836   0.4037
## re74            0.1393     0.1024   1.360   0.1747
## re75            0.2603     0.1696   1.535   0.1257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6891 on 359 degrees of freedom
## Multiple R-squared:  0.06574,    Adjusted R-squared:  0.03972
## F-statistic: 2.526 on 10 and 359 DF,  p-value: 0.005979
```

(d) Assuming that the estimates from (b) and (c) can be interpreted causally, what causal effect does each estimate? (Hint: what populations are we making inferences about for each of these estimates?)

(e) Redo both the regression and the matching exercises, excluding the variable for earnings in 1974 (two time periods before the start of this study). How im- portant does the earnings-in-1974 variable appear to be in terms of satisfying the ignorability assumption?

PM Do this manually.