PLSC 30600 Lab 5

Evaluating observational methods against a benchmark: LaLonde (1986)

This lab looks at a classic paper which studies the performance of different observational adjustment methods relative to an experimental benchmark. LaLonde (1986) carries out a re-analysis of a large-scale randomized experiment of a job training program. In the 1970s, the federal government instituted a fully randomized evaluation of the National Supported Work Demonstration, a subsidized work program. The paper compared the estimated effects obtained from the randomized trial to an artificial observational dataset that combined the treated units from the experiment with a non-experimental control group using respondent data from the Population Survey fo Income Dynamics (PSID). The original paper (and follow-up evaluations of other observational methods) compared estimates from regression on the "observational" dataset with the experimental benchmark to see how well the observational analysis could recover the "hidden experiment" in the data.

You will need two datasets. The experimental data is `nsw_exper.dta`. The observational data is `nsw_psid_withtreated.dta`. The variables of interest are:

- `re78` - Outcome: Real (inflation adjusted) earnings for 1978
- `nsw` - Treatment (1 for NSW participants, 0 otherwise)
- `age` - Age in years
- `educ` - Years of education
- `black` - Respondent is African American
- `hisp` - Respondent is Hispanic
- `married` - Respondent is married
- `re74` - Real (inflation adjusted) earnings for 1974
- `re75` - Real (inflation adjusted) earnings for 1975
- `u74` - Respondent was unmployed in 1974
- `u75` - Respondent was unmployed in 1975

The code below load these datasets

```
benchmark <- haven::read_dta("nsw_exper.dta")
lalonde <- haven::read_dta("nsw_psid_withtreated.dta")

head(benchmark)
```

```
## # A tibble: 6 x 12
##     nsw   age  educ black  hisp married  re74  re75   re78   u74   u75   u78
##   <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>
## 1     1    37    11     1     0       1     0     0  9930.     1     1     0
## 2     1    22     9     0     1       0     0     0  3596.     1     1     0
## 3     1    30    12     1     0       0     0     0 24910.     1     1     0
## 4     1    27    11     1     0       0     0     0  7506.     1     1     0
## 5     1    33     8     1     0       0     0     0   290.     1     1     0
## 6     1    22     9     1     0       0     0     0  4056.     1     1     0
```

```
head(lalonde)
```

```
## # A tibble: 6 x 12
```

```
##     nsw   age  educ black  hisp married  re74  re75  re78   u74   u75   u78
##   <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     0    47    12     0     0       0     0     0     0     1     1     1
## 2     0    50    12     1     0       1     0     0     0     1     1     1
## 3     0    44    12     0     0       0     0     0     0     1     1     1
## 4     0    28    12     1     0       1     0     0     0     1     1     1
## 5     0    54    12     0     0       1     0     0     0     1     1     1
## 6     0    55    12     0     1       1     0     0     0     1     1     1
```

Observational results vs. experimental benchmark

First let's estimate the benchmark ATE of assignment to the NSW program on real earnings in 1978 using the experiment

```
lm_robust(re78 ~ nsw, data=benchmark)
```

```
##             Estimate Std. Error   t value     Pr(>|t|)  CI Lower CI Upper  DF
## (Intercept) 4554.802   340.0931 13.392809 1.391003e-34 3886.4059 5223.199 443
## nsw         1794.343   670.9967  2.674146 7.769016e-03  475.6108 3113.075 443
```

On average, we estimate the program increased real earnings by about 1794 dollars. Our 95% confidence interval does not contain 0, so we'd reject the null of no ATE at $\alpha = .05$.

Let's compare this to the estimate we'd get if we took the simple difference in means in the observational data

```
lm_robust(re78 ~ nsw, data=lalonde)
```

```
##              Estimate Std. Error   t value      Pr(>|t|)  CI Lower  CI Upper
## (Intercept)  21553.92   311.7310  69.14269  0.000000e+00  20942.66  22165.18
## nsw         -15204.78   657.0765 -23.14004 3.933528e-108 -16493.21 -13916.35
##                DF
## (Intercept) 2673
## nsw         2673
```

In the observational data, the estimated "effect" of the program is around -15K dollars! Why is the estimate so far off? Because of selection-into-treatment bias. Let's diagnose using a balance test on some of the observed covariates.

```
# Experimental data
benchmark %>%
  group_by(nsw) %>%
  summarise(age = mean(age),
            educ = mean(educ),
            black = mean(black),
            hisp = mean(hisp),
            married = mean(married),
            re74 = mean(re74),
            re75 = mean(re75),
            u74 = mean(u74),
            u75 = mean(u75))
```

```
## # A tibble: 2 x 10
##     nsw   age  educ black   hisp married  re74  re75   u74   u75
##   <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     0  25.1  10.1 0.827 0.108    0.154 2107. 1267. 0.75  0.685
## 2     1  25.8  10.3 0.843 0.0595   0.189 2096. 1532. 0.708 0.6
```

```
cobalt::bal.tab(benchmark %>%
select(age, educ, black, hisp, married, re74, re75, u74, u75), treat=benchmark$nsw, binary="std", s.d.d
```

```
## Balance Measures
##            Type Diff.Un
## age     Contin.  0.1073
## educ    Contin.  0.1412
## black    Binary  0.0440
## hisp     Binary -0.1749
## married  Binary  0.0939
## re74    Contin. -0.0022
## re75    Contin.  0.0839
## u74      Binary -0.0944
## u75      Binary -0.1772
##
## Sample sizes
##     Control Treated
## All     260     185
```

```
# Observational data

lalonde %>%
  group_by(nsw) %>%
  summarise(age = mean(age),
            educ = mean(educ),
            black = mean(black),
            hisp = mean(hisp),
            married = mean(married),
            re74 = mean(re74),
            re75 = mean(re75),
            u74 = mean(u74),
            u75 = mean(u75))
```

```
## # A tibble: 2 x 10
##     nsw   age  educ black   hisp married   re74   re75    u74   u75
##   <dbl> <dbl> <dbl> <dbl>  <dbl>   <dbl>  <dbl>  <dbl>  <dbl> <dbl>
## 1     0  34.9  12.1 0.251 0.0325   0.866 19429. 19063. 0.0863   0.1
## 2     1  25.8  10.3 0.843 0.0595   0.189  2096.  1532. 0.708    0.6
```

```
cobalt::bal.tab(lalonde %>%
select(age, educ, black, hisp, married, re74, re75, u74, u75),
treat=lalonde$nsw, binary="std", s.d.denom="pooled")
```

```
## Balance Measures
##            Type Diff.Un
## age     Contin. -1.0094
## educ    Contin. -0.6805
## black    Binary  1.4816
## hisp     Binary  0.1288
## married  Binary -1.8453
## re74    Contin. -1.7178
## re75    Contin. -1.7744
## u74      Binary  1.6454
## u75      Binary  1.2309
##
```

```
## Sample sizes
##     Control Treated
## All    2490     185
```

While there might be some minor imbalance in the experimental data, the observational data is **clearly** imbalanced. Lower-income individuals are more likely to have enrolled in the program (which makes sense - the observational data was created for the LaLonde paper by merging the experiment with a sample from a general survey of the population). So past income is a major confounder!

Let's see how well the simple additive regression does at recovering the experimental target.

```
lm_robust(re78 ~ nsw + age + educ + black + hisp + married + re74 + re75 + u74 + u75, data=lalonde)
```

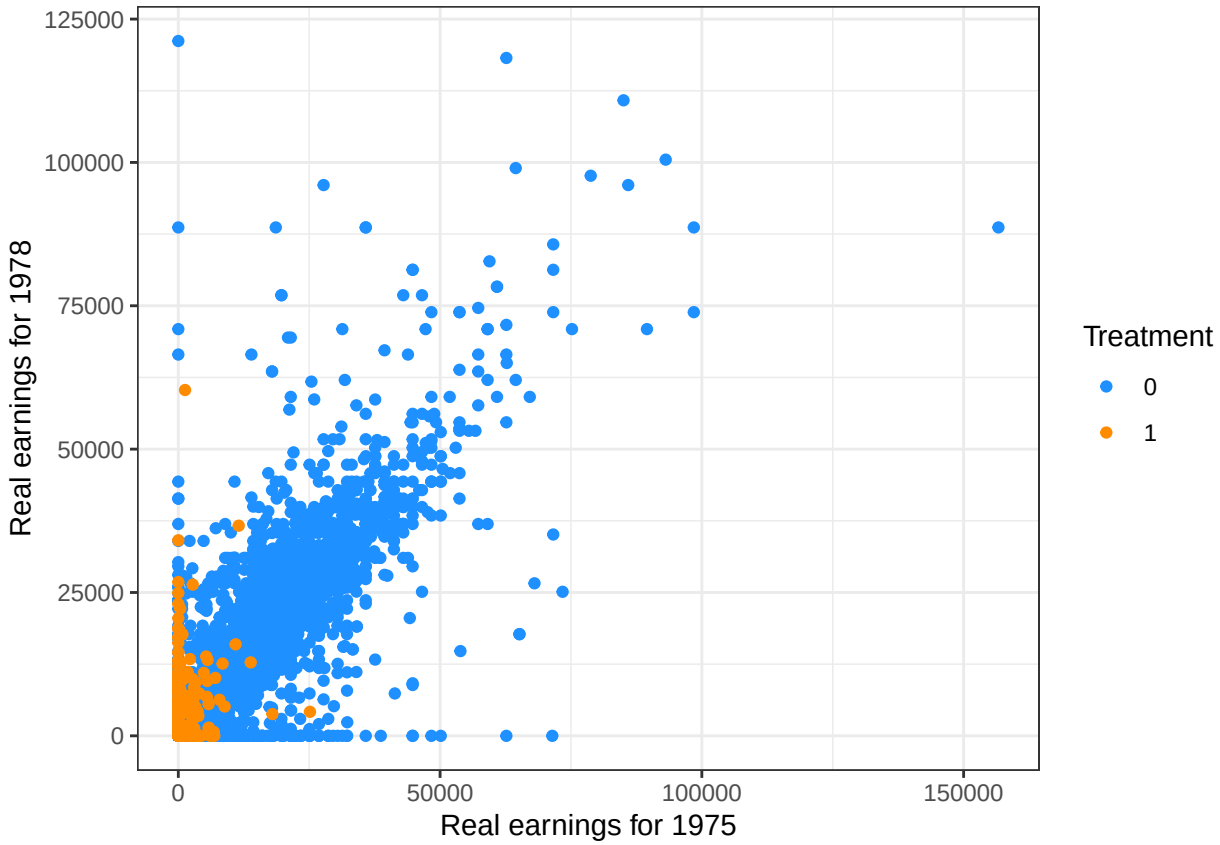
```
##                  Estimate   Std. Error    t value     Pr(>|t|)      CI Lower
## (Intercept)   953.6012324 1.505458e+03  0.6334293 5.265077e-01 -1998.3836642
## nsw           115.3825302 8.341658e+02  0.1383209 8.899973e-01 -1520.2955880
## age           -89.7654170 2.338262e+01 -3.8389806 1.264215e-04  -135.6153370
## educ          514.1239530 9.302165e+01  5.5269276 3.574808e-08   331.7219873
## black        -454.2159511 4.466251e+02 -1.0169960 3.092477e-01 -1329.9830044
## hisp         2197.3729428 1.238119e+03  1.7747672 7.605051e-02  -230.3987184
## married      1204.7846653 4.970610e+02  2.4238166 1.542454e-02   230.1202306
## re74            0.3126200 6.214554e-02  5.0304497 5.218660e-07     0.1907616
## re75            0.5436544 6.897191e-02  7.8822575 4.646173e-15     0.4084105
## u74          2389.5305091 1.366360e+03  1.7488291 8.043573e-02  -289.7035650
## u75         -1461.9650135 1.419215e+03 -1.0301222 3.030461e-01 -4244.8398682
##                 CI Upper   DF
## (Intercept) 3905.5861290 2664
## nsw         1751.0606484 2664
## age          -43.9154970 2664
## educ         696.5259186 2664
## black        421.5511023 2664
## hisp        4625.1446040 2664
## married     2179.4491000 2664
## re74           0.4344784 2664
## re75           0.6788983 2664
## u74         5068.7645832 2664
## u75         1320.9098411 2664
```

Not great - we estimate an average effect of 115 (an order of magnitude less than the benchmark). Let's see what's going on.

Illustrating regression

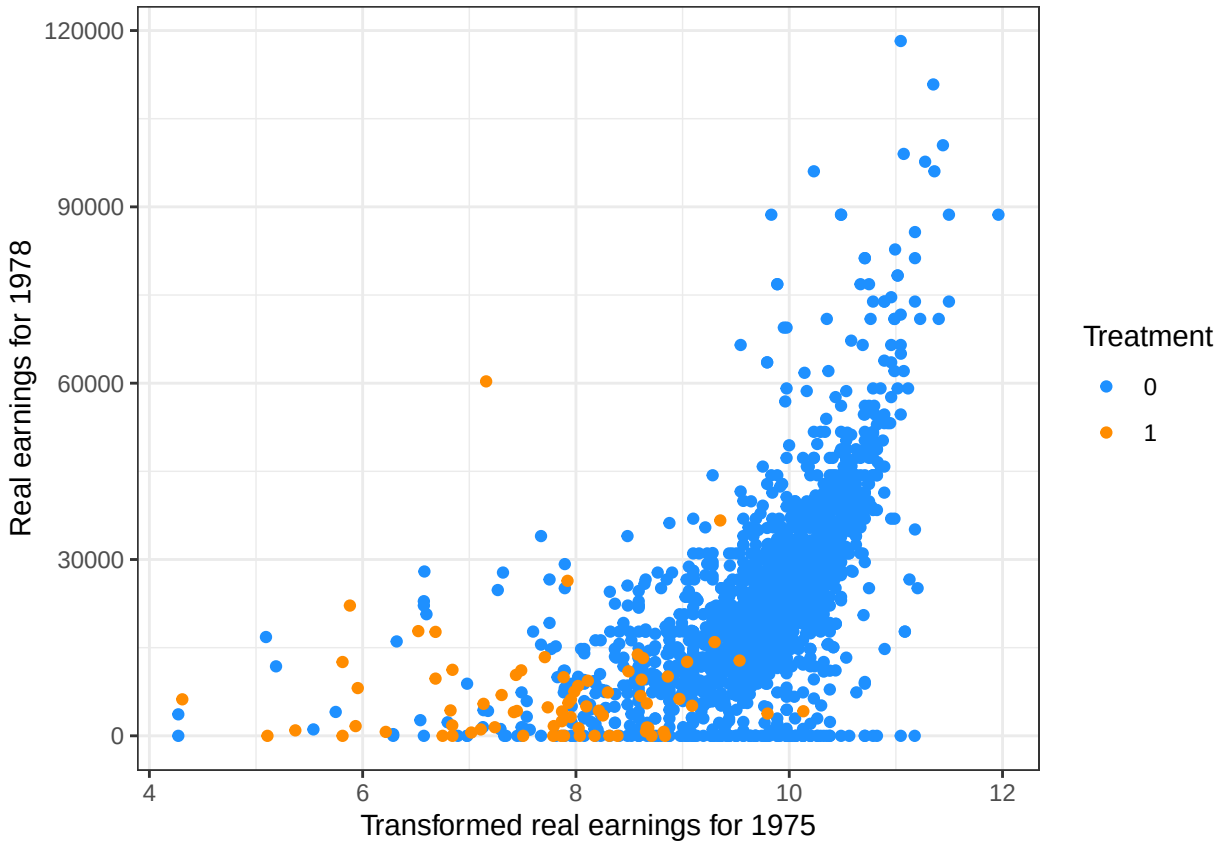
We'll start by looking at an adjustment for lagged income (re75). Start by plotting the relationship between re75 and re78 within the treated and control group. Start with a simple scatterplot:

```
lalonde %>%
ggplot(aes(x=re75, y=re78, colour = as.factor(nsw))) +
 geom_point() +
 xlab("Real earnings for 1975") +
 ylab("Real earnings for 1978") +
 scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
 theme_bw()
```



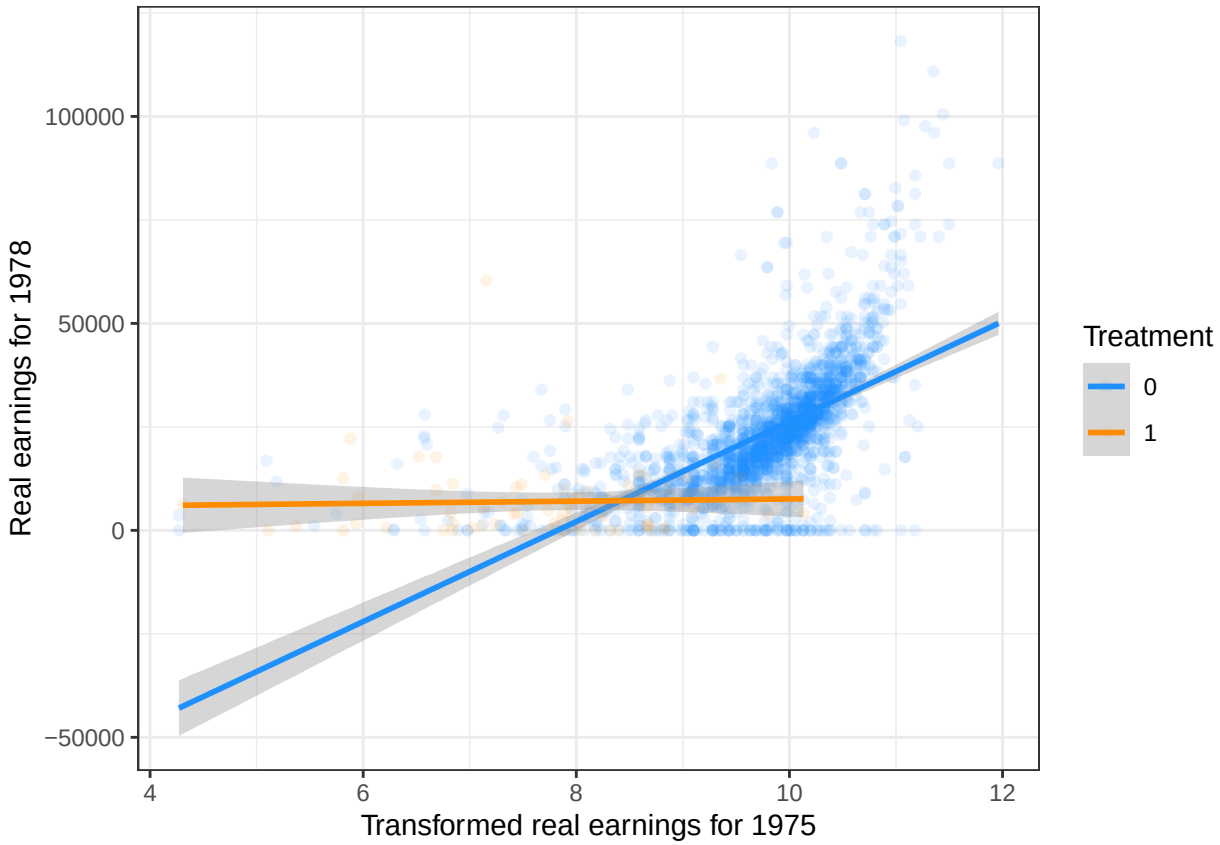
It looks like a lot of the difference is being driven by many of the participants in the NSW being unemployed in 1975 (re75 = 0). Let's condition on `u75 == 0` . We'll also do a log transform on the X to make visualization easier (we could do this on the outcome as well, but we'd have to deal with the zeroes somehow since log(0) is undefined).

```
lalonde %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw))) +
geom_point() +
xlab("Transformed real earnings for 1975") +
ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```

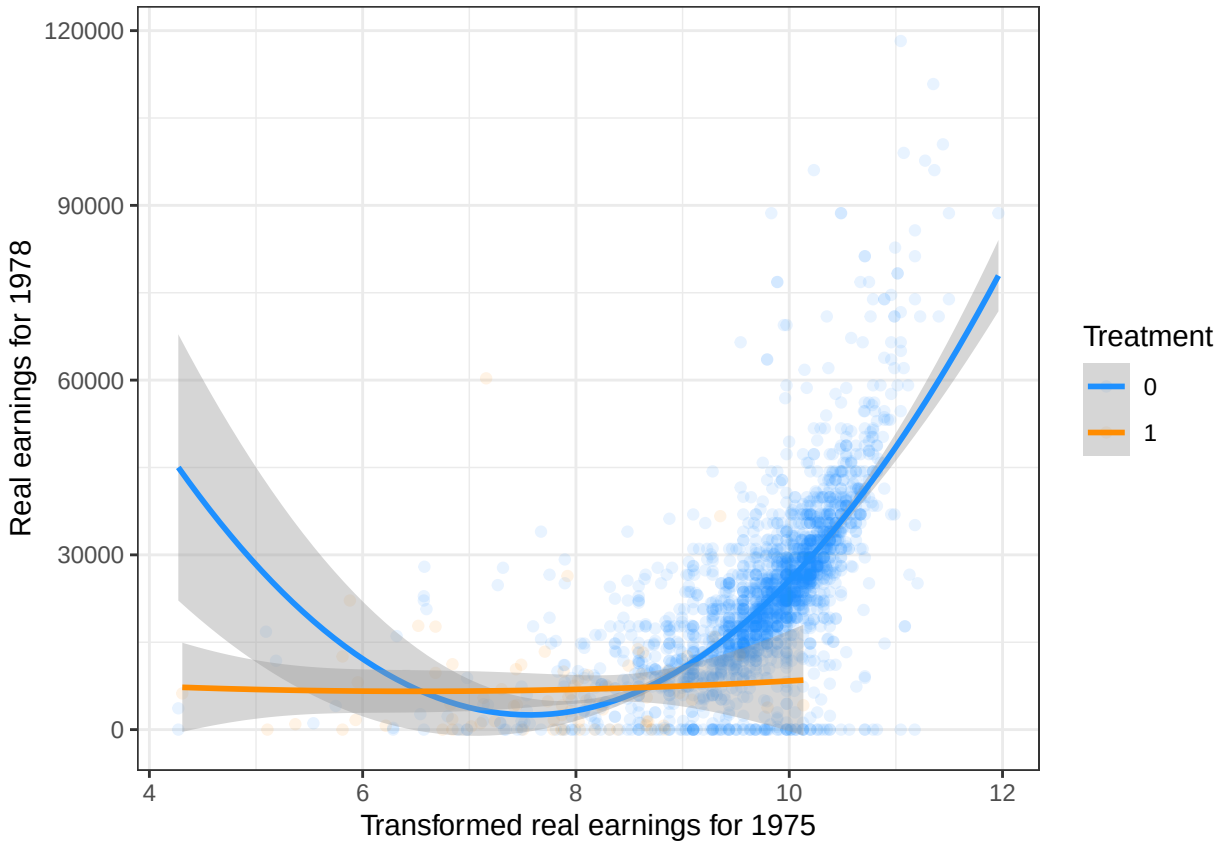


So we have some interesting modeling choices - let's compare a linear fit versus a quadratic versus a cubic fit.

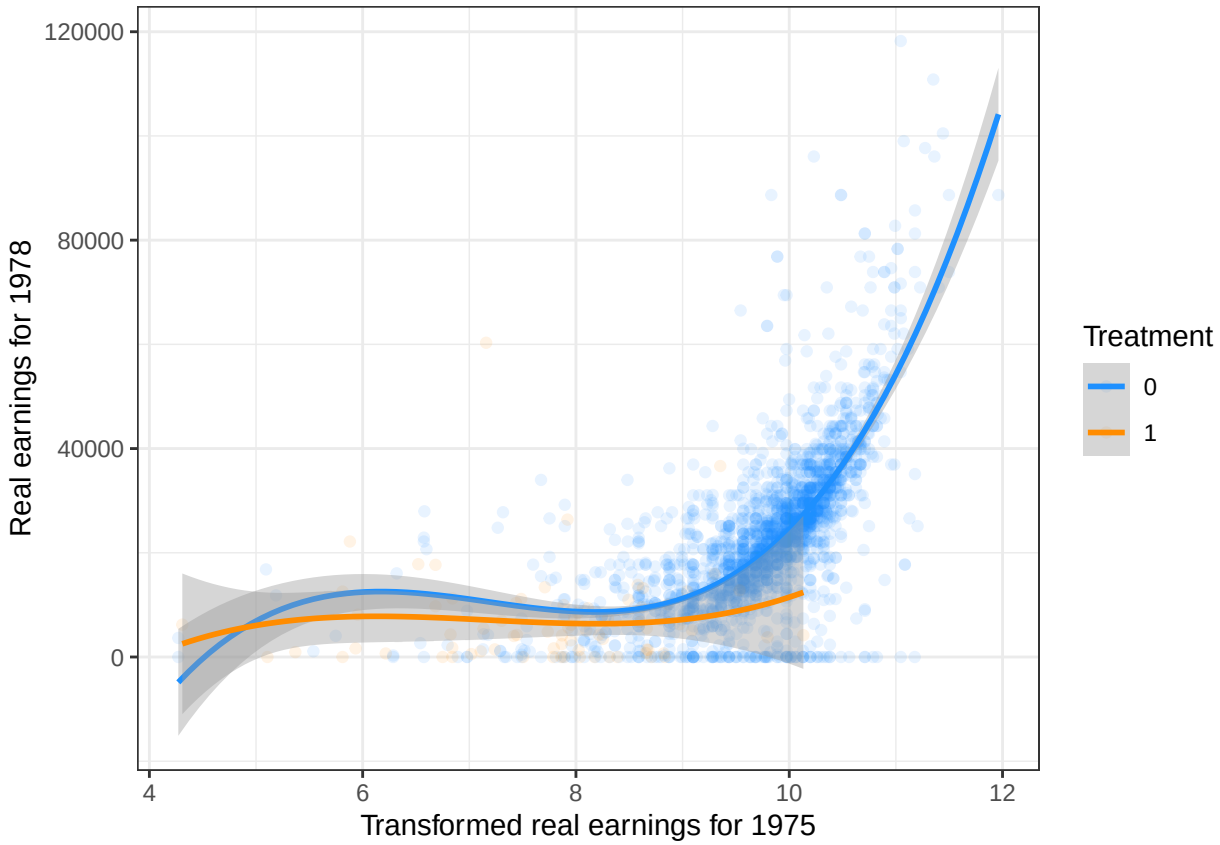
```
# Linear
lalonde %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw))) + geom_point(alpha = .1) +
geom_smooth(method="lm_robust", formula = y ~ x) +
xlab("Transformed real earnings for 1975") +
ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```




```
# Quadratic
lalonde %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw))) +
geom_point(alpha = .1) +
geom_smooth(method="lm_robust", formula = y ~ x + I(x^2)) + xlab("Transformed real earnings for 1975") +
ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```




```
#  Cubic
lalonde %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw))) +
geom_point(alpha = .1) +
geom_smooth(method="lm_robust", formula = y ~ x + I(x^2) + I(x^3)) +
xlab("Transformed real earnings for 1975") +
ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```

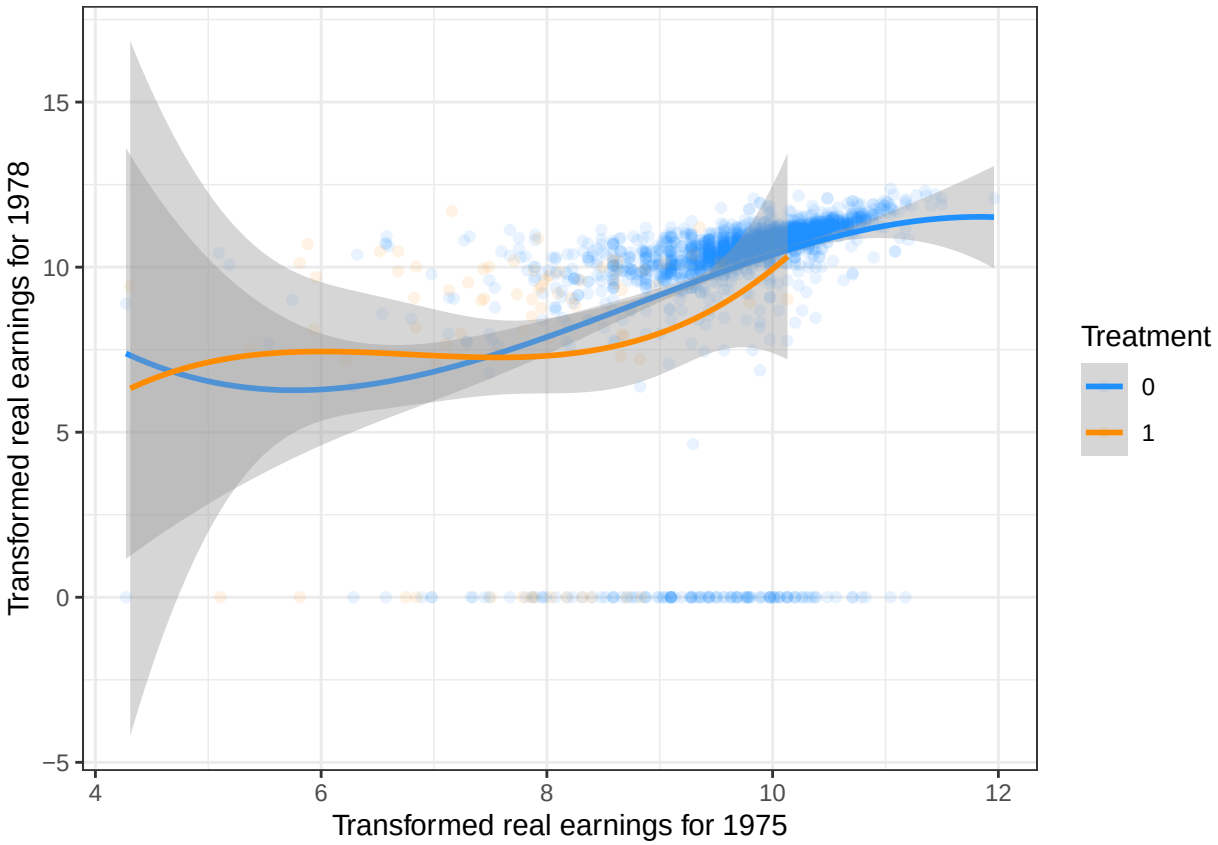


Why are these choices so consequential? Because there is very poor overlap between the distribution of the treated units and the control units. So a model fit to the entire control distribution. We might try transforming the outcome (and change the quantity of interest) to get a more plausible linear Conditional Expectation Function - CEF (common with log-log relationships). Note here we have zeroes in the outcome (unemployment), so a log-transform will need to address this - the inverse hyperbolic sine is popular, but arguably just as arbitrary as taking `log(x+1)`

```
ihs <- function(x) {
    y <- log(x + sqrt(x ^ 2 + 1))
    return(y)
}
# Linear
lalonde %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=ihs(re78), colour = as.factor(nsw))) +
geom_point(alpha=.1) +
geom_smooth(method="lm_robust", formula = y ~ x) +
xlab("Transformed real earnings for 1975") +
ylab("Transformed real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```


```
# Quadratic
lalonde %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=ihs(re78), colour = as.factor(nsw))) +
geom_point(alpha=.1)  +
geom_smooth(method="lm_robust", formula = y ~ x + I(x^2)) +
xlab("Transformed real earnings for 1975") +
ylab("Transformed real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```


```
#  Cubic
lalonde %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=ihs(re78), colour = as.factor(nsw))) +
geom_point(alpha=.1)  +
geom_smooth(method="lm_robust", formula = y ~ x + I(x^2) + I(x^3)) +
xlab("Transformed real earnings for 1975") +
ylab("Transformed real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```



The model choice is a bit less consequential here - and linearity is not a terrible approximation. But note that now our quantity of interest has changed - it's no longer the ATE on real earnings in 1978, it's the ATE on a *transformation* of real earnings. This may be fine if what we really care about is just the direction of the effect, but it does affect interpretability.

Matching as pre-processing.

Let's see what happens when we match on the pre-treatment covariates - let's ignore the bias adjustment for now.

```
# Set ties = F to randomly break ties in distances. ties = T will change M to accomodate ties and incre
# weight = 2: Mahalanobis distance
# M = 3: 3:1 matching, 3 control obs to 1 treated obs
set.seed(60637)
match_result <- Matching::Match(Y = lalonde$re78, Tr = lalonde$nsw, X = lalonde %>%
dplyr::select(age, educ, black, hisp, married, re74, re75, u74, u75),
              M = 3, Weight = 2, estimand = "ATT", ties = F)

summary(match_result)
```

```
##
## Estimate...  1480.3
## SE.........  733.64
## T-stat.....  2.0177
## p.val......  0.04362
##
## Original number of observations..............  2675
```

```
## Original number of treated obs...............  185
## Matched number of observations...............  185
## Matched number of observations  (unweighted).  555
```

We get a *lot* closer to the experimental benchmark! Let's calculate the "matching weights' for each observation and see what's going on.

```
match_weights = table(c(match_result$index.control, match_result$index.treated))/3
lalonde$matchweight <- 0
# match_weights is the weight for each row number -
# this is a quick trick to assign the weights to the right rows
lalonde$matchweight[as.numeric(names(match_weights))] <- match_weights


head(lalonde)
```

```
## # A tibble: 6 x 13
##     nsw   age  educ black  hisp married  re74  re75  re78   u74   u75   u78
##   <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     0    47    12     0     0       0     0     0     0     1     1     1
## 2     0    50    12     1     0       1     0     0     0     1     1     1
## 3     0    44    12     0     0       0     0     0     0     1     1     1
## 4     0    28    12     1     0       1     0     0     0     1     1     1
## 5     0    54    12     0     0       1     0     0     0     1     1     1
## 6     0    55    12     0     1       1     0     0     0     1     1     1
## # i 1 more variable: matchweight <dbl>
```

```
# What's the share of observations with 0 weight
mean(lalonde$matchweight == 0)
```

```
## [1] 0.875514
```

```
nrow(lalonde)
```

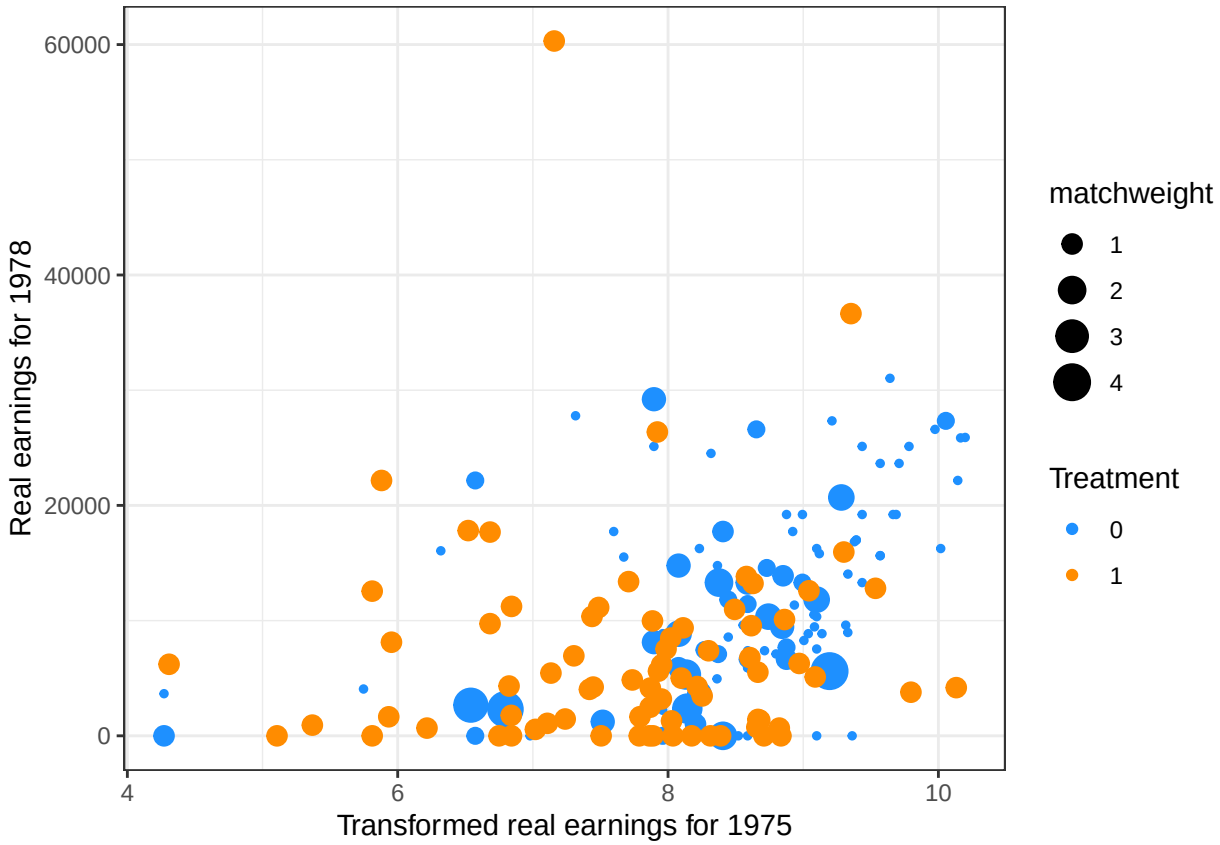
```
## [1] 2675
```

```
# Subset the data to post-matching observations
lalonde_postmatch <- lalonde %>% filter(matchweight != 0)
nrow(lalonde_postmatch)
```

```
## [1] 333
```

Matching throws away about 90% of the data! Let's see what this means for estimating the CEF of re78 given re75.

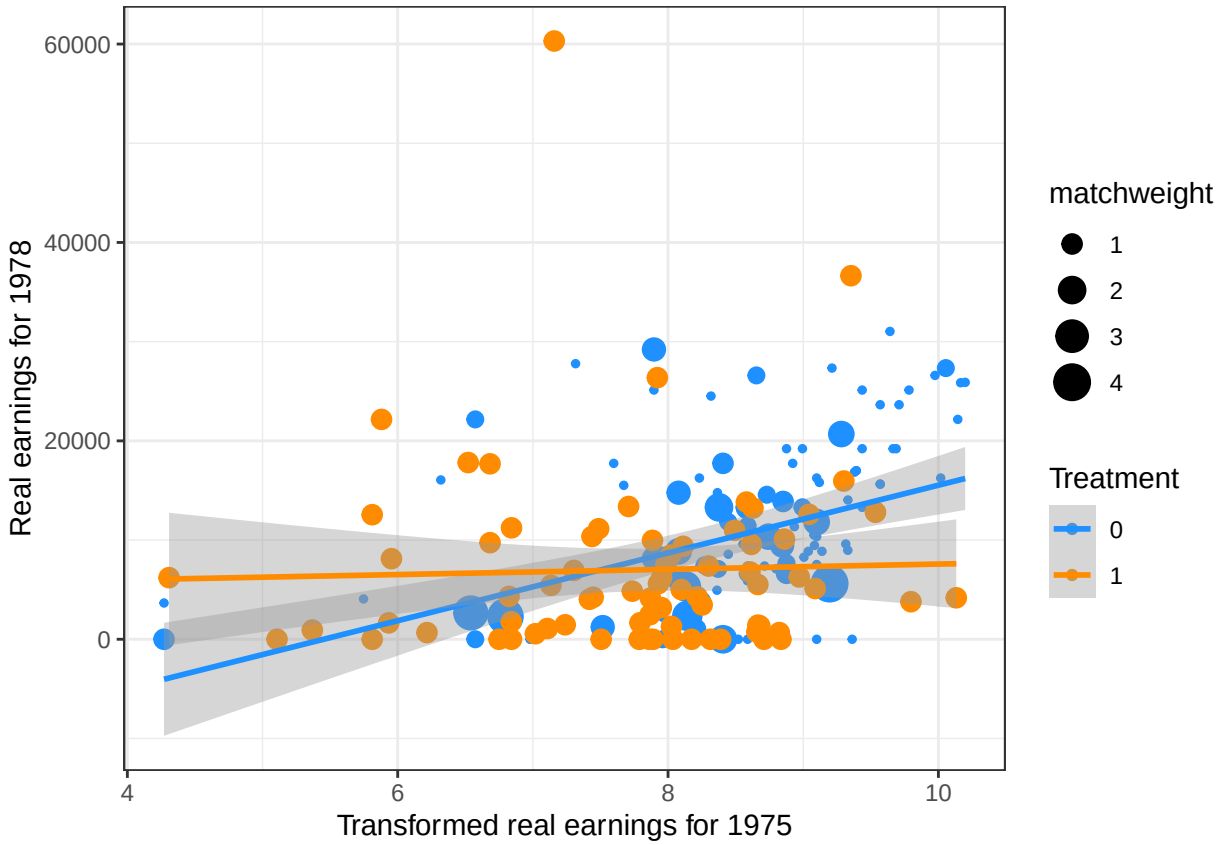
```
# Scatterplot
lalonde_postmatch %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw), size = matchweight)) +
geom_point() +
xlab("Transformed real earnings for 1975") +
ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```



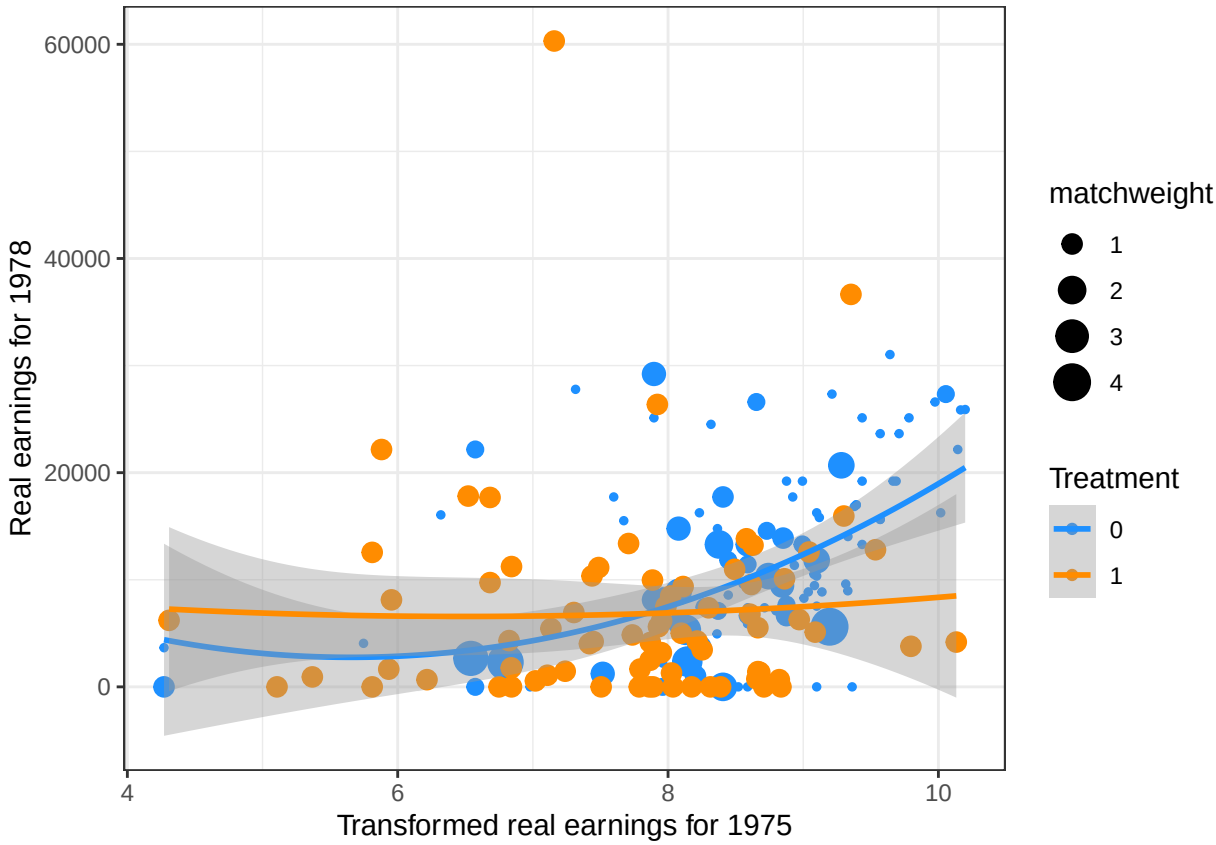
Post-matching, we have a lot better covariate overlap between treated and control groups. Let's see whether this affects the impact our modeling choices

```
# Linear
lalonde_postmatch  %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw), size = matchweight, weight=matchweight)) +
geom_point(aes(size=matchweight)) +
geom_smooth(method="lm_robust", formula = y ~ x, size=1) +
xlab("Transformed real earnings for 1975") +
ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# Quadratic
lalonde_postmatch  %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw), size = matchweight, weight=matchweight)) +
geom_point(aes(size=matchweight)) +
geom_smooth(method="lm_robust", formula = y ~ x + I(x^2), size=1) +
xlab("Transformed real earnings for 1975") +
ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```




```
# Cubic
lalonde_postmatch  %>% filter(u75 == 0) %>%
ggplot(aes(x=log(re75), y=re78, colour = as.factor(nsw), size = matchweight, weight=matchweight)) +
geom_point(aes(size=matchweight)) +
geom_smooth(method="lm_robust", formula = y ~ x + I(x^2) + I(x^3), size=1) +
xlab("Transformed real earnings for 1975") + ylab("Real earnings for 1978") +
scale_colour_manual("Treatment", values = c("dodgerblue", "darkorange")) +
theme_bw()
```

Each fit gives roughly similar results – interestingly what we find is that there's not a whole lot of an effect *among those employed* in the pre-treatment period.

Challenge Problem

Play around with different modeling choices in the regression and see how close you can get to the benchmark target even without matching. Try a regression imputation approach by fitting separate treatment and control models. Consider doing imputation only for the ATT rather than the ATE - how would the imputation estimator change?