Problem Set 3

Tessie Dong, Derek Li, Andi Liu

Due Jan 18th, 2024

Part 1: Describe the Data
Combine the NSW data in `nswre74_control.csv` and `nswre74_treated.csv` and complete Table 1. Note that variables `1-10` are *predetermined*, i.e., capture characteristics determined at or before treatment assignment; some of these variables are background characteristics (e.g., `edu`), others capture a subject's pre-RCT labor market experience (e.g., `u75`). `re78` is the observed outcome variable. `treat` is the indicator of treatment status.

```
# load and combine the data sets into one dataframe
df1 <- utils::read.csv(file = "nswre74_control.csv")
df2 <- utils::read.csv(file = "nswre74_treated.csv")
df <- rbind(df1, df2)
```

```
# count units in each sample
dplyr::tally(dplyr::group_by(df, treat))
```

```
## # A tibble: 2 x 2
##   treat     n
##   <int> <int>
## 1     0   260
## 2     1   185
```

```
# generate mean summary statistics for each variable and treatment
dplyr::summarise_all((dplyr::group_by(df, treat)), list(mean))
```

```
## # A tibble: 2 x 12
##   treat   age   edu black   hisp married nodegree  re74  re75  re78   u74   u75
##   <int> <dbl> <dbl> <dbl>  <dbl>   <dbl>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     0  25.1  10.1 0.827 0.108    0.154    0.835 2107. 1267. 4555. 0.75  0.685
## 2     1  25.8  10.3 0.843 0.0595   0.189    0.708 2096. 1532. 6349. 0.708 0.6
```

A completed version of Table 1 is provided on the following page.

Variable Counter	Variable Name	Variable Definition	Sample Average	
			Treated	Control
1	age	Age in years	25.8	25.1
2	edu	Education in years	10.3	10.1
3	nodegree	1 if education < 12	0.708	0.835
4	black	1 if Black	0.843	0.827
5	hisp	1 if Hispanic	0.0595	0.108
6	married	1 if married	0.189	0.154
7	u74	1 if unemployed in '74	0.708	0.75
8	u75	1 if unemployed in '75	0.6	0.685
9	re74	Real earnings in '74 (in '82 $)	2096	2107
10	re75	Real earnings in '75 (in '82 $)	1532	1267
11	re78	Real earnings in '78 (in '82 $)	6349	4555
12	treat	1 if received offer of training	1	0
Sample Size			185	260

Table 1: Descriptive statistics for the NSW data by group.

Part 2: Test Balance

1. Test balance for each of the 10 OPVs in Table (1), i.e., test that each variable's mean is the same in the control and treated groups. Do so by running 10 simple linear regressions (SLR) specifications. Use a 5% significance level and look at the relevant 10 t-tests, comment on your findings. **Hint**: Each OPV is the dependent variable in its regression equation. All 10 SLR models have the same covariates.

```
ols_p_values <- list()
ols_coefficients <- list()
ols_se <- list()
vars <- names(df)[2:12]
for (var in vars) {
    formula <- stats::formula(paste(var, "~treat"))
    lm_model <- lm(formula = formula, data = df)
    ols_p_values[[var]] <- summary(lm_model)$coefficients[2, 4]
    ols_coefficients[[var]] <- summary(lm_model)$coefficients[2, 1]
    ols_se[[var]] <- summary(lm_model)$coefficients[2, 2]
}
# print(ols_p_values)
# print(ols_coefficients)
# print(ols_se)
```

Table 2 shows the p-values for the t-tests of balance for each OPV.
We can see that for most of our variables, the p-value is greater than 0.05, so we fail to reject the null hypothesis that the means are the same in the control and treated groups. However, for `nodegree` we reject the null hypothesis that the means are the same in the control and treated groups, and we cannot claim that these groups are balanced for this variable.

2. Testing balance as done in Q1 suffers from the so called "multiple comparisons" or "multiple testing" problem which occurs when one considers a set of statistical inferences simultaneously. The problem emerges because as more variables are compared, it becomes more likely that the treatment and control groups appear to differ on at least one attribute *by random chance alone*. To deal with this problem we use an estimation methodology called SUR estimation and then test just one hypothesis, the *joint* hypothesis that all OPVs are balanced, i.e., their means are the same in the two groups. SUR stands for seemingly unrelated regression and it is a special case of feasible Generalized Least Squares (GLS)

	Variable	Coefficient	se	p-value
1	age	0.7623701	0.6827511	0.264764
2	edu	0.2574844	0.1721353	0.135411
3	nodegree	-0.1265073	0.0393452	0.001398
4	black	0.01632017	0.03588617	0.649493
5	hisp	-0.04823285	0.0271632	0.076474
6	married	0.03534304	0.03604844	0.327408
7	u74	-0.04189189	0.0426221	0.326209
8	u75	-0.08461538	0.04582176	0.065469
9	re74	-11.45296	516.478	0.982318
10	re75	265.1463	303.1555	0.382254

Table 2: Coefficients, standard errors, and p-values for the t-tests of balance for each OPV.

estimation. Instead of estimating the coefficients *equation-by-equation* by OLS (and done in Q1 you combine the 10 equations in a system of equations and estimate the coefficients present in all the equations jointly, accounting for the fact that the unobservables may be correlated across equations within an individual (we continue to assume that they are uncorrelated across units). After estimation, you use standard testing procedures to test the *joint* hypothesis that the slope coefficients in all the equations of the SUR system are zero. This joint test is a test of covariate balance that does not suffer from the "multiple testing'' problem.

a. Estimate the SUR system. Are the estimated coefficients and their SEs different from those obtained in Q1? Comment. **Hint**: In 2 situations there is no efficiency payoff to GLS versus OLS: 1) the unobservables are uncorrelated across equations within an individual; and 2) the equations have identical covariates.

```
# create a list of formulas for each equation
formulas <- list()
vars <- names(df)[2:12]
for (var in vars) {
  formulas[[var]] <- formula(paste(var, "~treat"))
}
```

```
# estimate the SUR model
sur_fit <- systemfit::systemfit(formula = formulas, data = df, method = "SUR")
```

```
# print the coefficients and SE of the SUR model
i <- 1
while (i <= 11){
  print(sprintf("%s: %f, %f",
                vars[i],
                summary(sur_fit)$coefficients[i * 2, 1],
                summary(sur_fit)$coefficients[i * 2, 2]))
  i <- (i + 1)
}
```

```
## [1] "age: 0.762370, 0.682751"
## [1] "edu: 0.257484, 0.172135"
## [1] "black: 0.016320, 0.035886"
## [1] "hisp: -0.048233, 0.027163"
## [1] "married: 0.035343, 0.036048"
```

```
## [1] "nodegree: -0.126507, 0.039345"
## [1] "re74: -11.452958, 516.477971"
## [1] "re75: 265.146299, 303.155497"
## [1] "re78: 1794.342382, 632.853392"
## [1] "u74: -0.041892, 0.042622"
## [1] "u75: -0.084615, 0.045822"
```

```
# print the coefficients and SE of the OLS model
for (var in vars)
{
  print(sprintf("%s: %f, %f ", var, ols_coefficients[var], ols_se[var]))
}
```

```
## [1] "age: 0.762370, 0.682751 "
## [1] "edu: 0.257484, 0.172135 "
## [1] "black: 0.016320, 0.035886 "
## [1] "hisp: -0.048233, 0.027163 "
## [1] "married: 0.035343, 0.036048 "
## [1] "nodegree: -0.126507, 0.039345 "
## [1] "re74: -11.452958, 516.477971 "
## [1] "re75: 265.146299, 303.155497 "
## [1] "re78: 1794.342382, 632.853392 "
## [1] "u74: -0.041892, 0.042622 "
## [1] "u75: -0.084615, 0.045822 "
```

The estimated coefficients and SEs are the same across OLS and GLS estimations for all the OPVs. This result may be due to the fact that the equations have the same covariate, i.e. `treat`.

b. Test *joint* balance by testing the joint hypothesis that the coefficients of `treat` are zero in all the equations of the system. Spell out null and alternative hypotheses, comment on findings, and verify the test's value and p-value "manually." **Hint**: Use the Likelihood Ratio (LR) test where the unrestricted model is the system of equations with `treat` as covariate, and the restricted model is the system with no regression covariates, i.e., with only the constant term. To verify the value of the test use its algebraic expression. Recall that the test has a χ^2_{df} distribution with df = number of restrictions.

```
# create a list of formulas null_system with only the constant
null_formulas <- list()
for (var in vars) {
  null_formulas[[var]] <- formula(paste(var, "~1"))
}

# pass the null_system to systemfit
null_fit <- systemfit::systemfit(formula = null_formulas, data = df, method = "SUR")

# calculate the LR test statistic
lrtest_obj <- lmtest::lrtest(null_fit, sur_fit)
lr_statistic <- lrtest_obj$Chisq[2]
lr_df <- lrtest_obj$Df[2]

p_value <- stats::pchisq(lr_statistic, df = lr_df, lower.tail = FALSE)
print(sprintf("LR test statistic: %f, p-value: %f", lr_statistic, p_value))
```

```
## [1] "LR test statistic: 27.021898, p-value: 0.004560"
```

The null hypothesis is that the coefficients of `treat` are zero in all the equations of the system. The alternative hypothesis is that there is at least one equation with a nonzero coefficient, implying that the control and treatment groups have different means for an OPV.

The p-value of this test is significant (p-value < 0.05). This result shows that we can reject null hypothesis that the coefficients of `treat` are zero in all the equations of the system, implying that the assignment of treatment did not balance all the subjects' characteristics.

3. Test that the OPVs do not predict treatment assignment. Spell out null and alternative hypotheses, comment on findings, and verify the test's value and p-value "manually.'' Why would scientists carry out this test? **Hint:** Use a MLRM and the F-test for the overall significance of the regression. **Programming guidance**: Use `summary()` after estimation, e.g., `summary(lm_fit)$fstatistic` returns the test's value and degrees of freedom. Use `stats::pf()` to verify the test's p-value.

   We want to test that, given the covariates, the treatment assignment is random. We can do this by testing that the coefficients of the covariates are zero in the regression of treatment assignment on the covariates.

   Given (Y, X^{OPV}) where Y represents the treatment assignment, and X is a vector of our OPV covariates, we can test the following model: $Y = \beta' X^{OPV} + \epsilon$ with these hypotheses:

$$\begin{cases} H_0 : \beta = 0 \\ H_1 : \beta \neq 0 \end{cases}$$

   In which β is a vector of coefficients for the OPV covariates. Thus, we want $\beta = 0$, i.e the OPV covariates all have no correlation to the treatment Y.

```
# create a list of formulas for each equation

lm_fit <- lm(formula = formula(paste("treat", "~", paste(vars, collapse = "+"))), data = df)
# summary(lm_fit)
# summary(lm_fit)$fstatistic
f_stat <- summary(lm_fit)$fstatistic[1]
numdf <- summary(lm_fit)$fstatistic[2]
dendf <- summary(lm_fit)$fstatistic[3]
p_val <- stats::pf(f_stat,numdf,dendf)

sprintf("The p-value of the ftest is: %f", p_val)
```

```
## [1] "The p-value of the ftest is: 0.994656"
```

With a p-value of 0.994656, we conclude that there is not enough evidence to reject the null hypothesis that OPVs do not predict treatment assignment. This result implies that proper random assignment was carried out such that treatment was given randomly irrespective of the subjects' characteristics. Scientists carry out this test to ensure that the treatment and control groups are similar in all respects except for the treatment itself.

Part 3: Estimate the Effect of the Offer of Training

1. Implement specs 0 (3 p) and 1 (2 p). Describe your findings.

We find in the implementation of spec 0 that with the relevant t-test between sample means the result is significant for $\alpha = 0.01$ with a p-value of 0.007893. This means we reject the null hypothesis that the means of the two populations are the same.

```
stats::t.test(df1$re78, df2$re78)
```

```
##
##  Welch Two Sample t-test
##
## data:  df1$re78 and df2$re78
## t = -2.6741, df = 307.13, p-value = 0.007893
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3114.6743  -474.0105
## sample estimates:
## mean of x mean of y
##  4554.801  6349.144
```

We find in the implementation of spec 1 an estimated slope coefficient of 1794.3 for the treatment variable in the Simple Linear Regression model. Both the intercept and slope are significant at $\alpha = 0.01$, which suggests a significant impact of the predictor variable (treatment) from the response variable (earnings).

```
spec1 = stats::lm(data=df, re78 ~ treat)
summary(spec1)
```

```
##
## Call:
## stats::lm(formula = re78 ~ treat, data = df)
##
## 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
```

2. Implement specs 2 through 4.

a. Report the estimates of the ATE and test H_0 that ATE is zero (3 p each spec). **Hint:** You consider spec 2 in light of the imbalance in educational attainment documented in Part 2. This approach to account for the presence of observable confounders (i.e., imbalance in OPVs) is called the regression adjustment approach. Spec 3 has 11 regression covariates. Spec 4 has 12 regression covariates.

In spec 2 we estimate the ATE to be \$1645.9. We reject H_0 at $\alpha = 0.05$.

```
spec2 = stats::lm(data=df, re78 ~ treat + factor(nodegree) + edu)
summary(spec2)
```

```
##
## Call:
## stats::lm(formula = re78 ~ treat + factor(nodegree) + edu, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7570  -4762  -1708   3129  53900
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1741.4     2889.8   0.603   0.5471
## treat               1645.9      638.0   2.580   0.0102 *
## factor(nodegree)1   -518.3      984.0  -0.527   0.5987
## edu                  321.8      224.9   1.431   0.1533
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6554 on 441 degrees of freedom
## Multiple R-squared:  0.02988,    Adjusted R-squared:  0.02328
## F-statistic: 4.527 on 3 and 441 DF,  p-value: 0.003862
```

In spec 3 we estimate the ATE to be \$1671. We reject H_0 at $\alpha = 0.01$.

```
spec3 = stats::lm(data=df, re78 ~ treat + factor(nodegree) + edu + age
                  + factor(black) + factor(hisp) + factor(married)
                  + re74 + re75 + factor(u74) + factor(u75))
summary(spec3)
```

```
##
## Call:
## stats::lm(formula = re78 ~ treat + factor(nodegree) + edu + age +
##     factor(black) + factor(hisp) + factor(married) + re74 + re75 +
##     factor(u74) + factor(u75), data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9612  -4355  -1572   3054  53119
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.567e+02  3.522e+03   0.073  0.94193
## treat              1.671e+03  6.411e+02   2.606  0.00948 **
## factor(nodegree)1 -1.518e+01  1.006e+03  -0.015  0.98797
## edu                4.008e+02  2.288e+02   1.751  0.08058 .
## age                5.357e+01  4.581e+01   1.170  0.24284
## factor(black)1    -2.037e+03  1.174e+03  -1.736  0.08331 .
## factor(hisp)1      4.258e+02  1.565e+03   0.272  0.78562
## factor(married)1  -1.463e+02  8.823e+02  -0.166  0.86834
## re74               1.234e-01  8.784e-02   1.405  0.16080
## re75               1.974e-02  1.503e-01   0.131  0.89554
## factor(u74)1       1.380e+03  1.188e+03   1.162  0.24590
## factor(u75)1      -1.071e+03  1.025e+03  -1.045  0.29651
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6517 on 433 degrees of freedom
## Multiple R-squared:  0.05822,    Adjusted R-squared:  0.0343
## F-statistic: 2.433 on 11 and 433 DF,  p-value: 0.005974
```

In spec 4 we estimate the ATE to be \$1660. We reject H_0 at $\alpha = 0.01$.

```
df$age_dev = (df$age - mean(df$age))/sd(df$age)
```

```
spec4 = stats::lm(data=df, re78 ~ treat + factor(nodegree) + edu + age
                  + factor(black) + factor(hisp) + factor(married) + re74
                  + re75 + factor(u74) + factor(u75) + age_dev : treat)

summary(spec4)
```

```
##
## Call:
## stats::lm(formula = re78 ~ treat + factor(nodegree) + edu + age +
##     factor(black) + factor(hisp) + factor(married) + re74 + re75 +
##     factor(u74) + factor(u75) + age_dev:treat, data = df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9655  -4340  -1549   2979  52961
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        1.079e+03  3.624e+03   0.298  0.76612
## treat              1.660e+03  6.413e+02   2.588  0.00998 **
## factor(nodegree)1  4.570e+00  1.006e+03   0.005  0.99638
## edu                4.070e+02  2.289e+02   1.778  0.07616 .
## age                1.837e+01  5.858e+01   0.314  0.75394
## factor(black)1    -1.988e+03  1.175e+03  -1.692  0.09138 .
## factor(hisp)1      4.807e+02  1.566e+03   0.307  0.75900
## factor(married)1  -1.740e+02  8.828e+02  -0.197  0.84386
## re74               1.230e-01  8.785e-02   1.400  0.16218
## re75               1.281e-02  1.505e-01   0.085  0.93222
## factor(u74)1       1.359e+03  1.188e+03   1.144  0.25321
## factor(u75)1      -1.124e+03  1.026e+03  -1.095  0.27402
## treat:age_dev      6.079e+02  6.307e+02   0.964  0.33563
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6517 on 432 degrees of freedom
## Multiple R-squared:  0.06024,    Adjusted R-squared:  0.03414
## F-statistic: 2.308 on 12 and 432 DF,  p-value: 0.007352
```

b. Are there reasons to add OPVs as covariates when they are balanced across treatment and control groups? If so, what are they? Comment on the estimation results from specs 2 and 3.

We can add additional OPVs even if they're balanced across the treatment and control groups because even with no systematic differences between the groups, those variables may still possibly contribute some causal

effect on the outcome. By regressing on those variables, we can remove some variance in the treatment effect that may be attributed to those OPVs. Additionally, even in randomized experiments, some imbalance can occur purely by chance. We notice that in spec 2, the p-value is significant at a lower α level when compared to the p-value in spec 3. Though the standard error for the estimated ATE is slightly higher for spec 3 compared to spec 2.

c. With reference to spec 3, do you think that it is problematic to use lagged / past values of the dependent variable as regression covariates? Explain.

No. It is common practice in time series analysis to regress on lagged values of the dependent variable. Though some assumptions may be required. In this case specifically, it does not seem problematic as the past earnings show no significant effect on earnings in 1978 as a result of the t-test.

d. Are there reasons to add interactions between OPVs and the treatment indicator? If so, what are they? With reference to spec 4, test the following two hypothesis: i) the ATE is zero; ii) the effect of the treatment does not vary by the age of the subject. **Note:** If additional assumptions are needed to carry out i) and ii), state them. **Programming Guidance**: Use `car::linearHypothesis()`.

Yes there are reasons to add interactions into the regression specification. Interactions with the treatment effect allow us to see how the treatment effect changes among individuals with different covariates. It may allow us to capture heterogenous effects of the treatment.

An additional assumption needed for ii) is homoskedasticity. The variance of the unobserved should be constant across different levels of age. If this assumption is violated, it can lead to incorrect conclusions about the interaction effect.

Here is the test for (i), we find significance at level of $\alpha = 0.01$ and we can claim to reject the null hypothesis that the ATE is 0.

```
linearHypothesis(spec4, c("treat=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## treat = 0
##
## Model 1: restricted model
## Model 2: re78 ~ treat + factor(nodegree) + edu + age + factor(black) +
##     factor(hisp) + factor(married) + re74 + re75 + factor(u74) +
##     factor(u75) + age_dev:treat
##
##   Res.Df        RSS Df Sum of Sq      F   Pr(>F)
## 1    433 1.8634e+10
## 2    432 1.8349e+10  1 284471777 6.6973 0.009981 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here is the test for (ii), we find that the test result is not significant thus we fail to to reject the null hypothesis that the effect of the treatment does not vary by the age of the subject.

```
linearHypothesis(spec4, c("treat:age_dev=0"))
```

```
## Linear hypothesis test
##
## Hypothesis:
## treat:age_dev = 0
##
## Model 1: restricted model
## Model 2: re78 ~ treat + factor(nodegree) + edu + age + factor(black) +
##     factor(hisp) + factor(married) + re74 + re75 + factor(u74) +
##     factor(u75) + age_dev:treat
##
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1    433 1.8389e+10
## 2    432 1.8349e+10  1  39465057 0.9291 0.3356
```

Both test results can also be witnessed in the summary of spec 4 that we originally ran.

3. Brainstorm on the possible mechanisms for the estimated ATE of being offered training. **Hint:** Can you think of the possible pathways through which on-the-job training may cause an increase in post-intervention earnings?

A potential mechanism for the estimated ATE of being offered training could be that if an individual who if offered training accepts and undergoes the training, they may received skill enhancement which will enable them to secure better-paying positions that have more skill requirements. Another possible mechanism for the estimated ATE of being offered training is suggested by the footnote, that an individual might become more optimistic about their prospects and therefore works harder to land a well-paying job, even without actually going through the training.

4. Starting with Hint 1, we express (y_{1i}^o, y_{0i}^o) in terms of (y_{1i}, y_{0i}). First, recognize that individuals that are not offered training, cannot take up training. This means that the earnings of individuals not offered training, are a subset of the group of individuals that do not take up training. We can assert that $y_{0i}^o = y_{0i}$ by recognizing this relationship.

   Then, recognize that individuals that are offered training, have a 0.5 probability of accepting or rejecting the offer of training. Individuals who accept the offer compose the group of individuals that receive training. Individuals who reject the offer are part of the a subset of the group of individuals that do not take up training. This can be expressed as

$$\begin{aligned} y_{1i}^o &= \mathbb{P}[\text{accept offer}]y_{1i} + \mathbb{P}[\text{reject offer}]y_{0i} \\ &= \frac{1}{2}y_{1i} + \frac{1}{2}y_{0i}. \end{aligned}$$

   We then find the following relationship between ATE and ATE^o:

$$\begin{aligned} ATE^o &= \mathbb{E}[y_{1i}^o - y_{0i}^o] \\ &= \mathbb{E}[(\frac{1}{2}y_{1i} + \frac{1}{2}y_{0i}) - y_{0i}], \text{ by definition,} \\ &= \mathbb{E}[\frac{1}{2}y_{1i} - \frac{1}{2}y_{0i}] \\ &= \frac{1}{2}\mathbb{E}[y_{1i} - y_{0i}], \text{ by linearity,} \\ &= \frac{1}{2}ATE. \end{aligned}$$

   Using this model we have shown that the two average treatment effects are different. Particularly that the Intent to Treat effect is of captured as a proportion of the Average Treatment Effect of receiving training, specifically $\frac{1}{2}$ in this case..