# Project 2 - Randomized Trials in Economics


## Alexander Jeremijenko - Econ 123


In [465]:

```
import pandas as pd
import matplotlib as plt
import statsmodels.api as sm
import numpy as np
import scipy.stats as st
from scipy.stats import ttest_ind as ttest
import warnings
from statsmodels.tools.sm_exceptions import ValueWarning
from scipy.stats import f as f_test
from scipy.stats import tstd
warnings.simplefilter('ignore', ValueWarning)
```

In [466]:

```
df = pd.read_stata('AEJApp-20090168_data.dta')
print('Field names are:')
display(pd.DataFrame(df.columns).rename(columns = {0: 'column'}))
display(df)
```

```
Field names are:
```

| | column |
|---|---|
| **0** | age_s |
| **1** | dmarried_s |
| **2** | empl_06 |
| **3** | salary_06 |
| **4** | profit_06 |
| **5** | tenure_06 |
| **6** | days_06 |
| **7** | hours_06 |
| **8** | contract_06 |
| **9** | dformal_06 |
| **10** | educ_s |
| **11** | lsalary_06 |
| **12** | lprofit_06 |
| **13** | lhours_06 |
| **14** | ldays_06 |
| **15** | city |
| **16** | age_lb |
| **17** | dmarried_lb |
| **18** | empl_04 |
| **19** | salary_04 |
| **20** | profit_04 |
| **21** | tenure_04 |
| **22** | days_04 |
| **23** | hours_04 |
| **24** | contract_04 |
| **25** | dformal_04 |
| **26** | educ_lb |
| **27** | ldays_04 |
| **28** | lhours_04 |
| **29** | select |
| **30** | pempl_06 |
| **31** | pempl_04 |
| **32** | dcontinue |
| **33** | codigo_ecap |

| | column |
|---|---|
| 34 | codigo_curs |
| 35 | dwomen |
| 36 | p_selecap |
| 37 | formalsal_06 |
| 38 | informalsal_06 |
| 39 | coursefixe |

| | age_s | dmarried_s | empl_06 | salary_06 | profit_06 | tenure_06 | days_06 | hours_06 | contract_( |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 22.0 | 0.0 | 1.0 | 0.0 | 240000.0 | 15.233334 | 22.0 | 84.0 | 0 |
| 1 | 22.0 | 0.0 | 1.0 | 116000.0 | 0.0 | 1.866667 | 10.0 | 14.0 | 0 |
| 2 | 24.0 | 0.0 | 1.0 | 650000.0 | 0.0 | 1.866667 | 28.0 | 91.0 | 0 |
| 3 | 24.0 | 0.0 | 1.0 | 408000.0 | 0.0 | 0.100000 | 28.0 | 48.0 | 0 |
| 4 | 22.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 3951 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 3952 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 3953 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 3954 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |
| 3955 | NaN | NaN | NaN | NaN | NaN | NaN | NaN | NaN | Na |

3956 rows × 40 columns

# Question 1

The power of the experiment is defined by the proportions sampled (1/2, 1/2) because it is the sample size which is a determinant of standard error of our estimates.

# Question 2

In [467]:

```
df = pd.read_stata('AEJApp-20090168_data.dta')
# First remove people who did not continue in the sample
df = df[df['dcontinue'] == 1]
df = df.assign(select = lambda x: pd.to_numeric(x.select.map({'selected': 1, 'contr
ol': 0})))

dummies = pd.get_dummies(df.coursefixe, prefix = 'fe')
dummy_list = list(dummies.columns)
# Remove the last item from the dummies
dummy_list.pop()
df = pd.concat([df, dummies], axis = 1).drop('coursefixe', axis = 1)

# Next split into men and women
df_w = df[df['dwomen'] == 1]
df_m = df[df['dwomen'] == 0]

# Now split into control (c) and treatment (t), also grouped by male and female
df_w_t = df_w[df_w.select == 1]
df_w_c = df_w[df_w.select == 0]
df_m_t = df_m[df_m.select == 1]
df_m_c = df_m[df_m.select == 0]


# Now, conducting t-tests at 5%, with the Null that the difference in means is 0, a
nd the alternative being a difference greater than 0. The variables below were chos
en because
# they are those which were chosen in the paper that we are replicating. Further, t
hey all plausibly will effect future labor outcomes, since they are related to curr
ent labor
# conditions
variables = ['empl_04', 'contract_04', 'dformal_04', 'pempl_04', 'salary_04', 'prof
it_04', 'days_04', 'hours_04', 'age_lb', 'dmarried_lb']
genders = ['MALE', 'FEMALE']
ttest_df = pd.DataFrame(columns = ['gender', 'field_name', 't_val', 'percentage_dif
f', 'significance', 'indicator'])
for gender in genders:
    for variable in variables:
        if gender == 'MALE':
            t_val, sig_level = ttest(df_m_t[variable], df_m_c[variable])
            pct_diff = (df_m_t[variable].mean() - df_m_c[variable].mean())/df_m_t[v
ariable].mean()*100
        else:
            t_val, sig_level = ttest(df_w_t[variable], df_w_c[variable])
            pct_diff = (df_w_t[variable].mean() - df_w_c[variable].mean())/df_w_t[v
ariable].mean()*100
        significance = 'significant' if sig_level < 0.05 else 'unsignificant'
        new_data = {'gender': gender, 'field_name': variable, 't_val': t_val, 'perc
entage_diff': pct_diff, 'significance': sig_level, 'indicator': significance}
        ttest_df = ttest_df.append(new_data, ignore_index='True')
ttest_df
```

Out[467]:

| | gender | field_name | t_val | percentage_diff | significance | indicator |
|---|---|---|---|---|---|---|
| **0** | MALE | empl_04 | 2.515859 | 10.732105 | 0.011980 | significant |
| **1** | MALE | contract_04 | 0.081946 | 1.256891 | 0.934701 | unsignificant |
| **2** | MALE | dformal_04 | -0.075694 | -1.087616 | 0.939673 | unsignificant |
| **3** | MALE | pempl_04 | 3.501277 | 20.473390 | 0.000477 | significant |
| **4** | MALE | salary_04 | 1.377236 | 9.650209 | 0.168649 | unsignificant |
| **5** | MALE | profit_04 | -1.830769 | -34.155462 | 0.067338 | unsignificant |
| **6** | MALE | days_04 | 1.984302 | 9.087635 | 0.047408 | significant |
| **7** | MALE | hours_04 | 2.080990 | 10.277980 | 0.037608 | significant |
| **8** | MALE | age_lb | -1.435012 | -0.731849 | 0.151497 | unsignificant |
| **9** | MALE | dmarried_lb | -2.242761 | -40.836707 | 0.025061 | significant |
| **10** | FEMALE | empl_04 | 0.292971 | 1.480857 | 0.769579 | unsignificant |
| **11** | FEMALE | contract_04 | -0.133585 | -2.342194 | 0.893746 | unsignificant |
| **12** | FEMALE | dformal_04 | 0.720357 | 11.918608 | 0.471401 | unsignificant |
| **13** | FEMALE | pempl_04 | 1.178168 | 7.555955 | 0.238888 | unsignificant |
| **14** | FEMALE | salary_04 | 0.346718 | 2.678929 | 0.728844 | unsignificant |
| **15** | FEMALE | profit_04 | 0.433581 | 7.668272 | 0.664646 | unsignificant |
| **16** | FEMALE | days_04 | 0.327419 | 1.758574 | 0.743389 | unsignificant |
| **17** | FEMALE | hours_04 | 0.265031 | 1.538361 | 0.791016 | unsignificant |
| **18** | FEMALE | age_lb | -1.690681 | -0.772594 | 0.091074 | unsignificant |
| **19** | FEMALE | dmarried_lb | -0.162461 | -1.293607 | 0.870961 | unsignificant |

We have no statistically significant fields for females, but several for males. Here, conducting a further joint F-test to see if these difference are jointly significant.

In [468]:

```
y = df_w.select
X = sm.add_constant(df_w[variables])
f_value = sm.OLS(y, X.astype(float)).fit().fvalue
print('The f_test for women is:', f_value)

y = df_m.select
X = sm.add_constant(df_m[variables])
f_value = sm.OLS(y, X.astype(float)).fit().fvalue
print('The f_test for men is:', f_value)
```

```
The f_test for women is: 1.117445685509792
The f_test for men is: 3.480850005495692
```

We have a significant f statistic for men and an unsignificant one for women. This is in-line with the results of the paper and with the table above.

Before we begin the regression analysis, a check for heteroskedasticity. The check will be made by seeing if there is statistically significant covariance between the squared residuals of the OLS regression and the X independents.

In [469]:

```
# First make predictions for y (we use employment as y here, and the characteristic
s above as X)
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'empl_06'

data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model_1 = sm.OLS(y, X).fit()
predict = model_1.predict(X)
# Generate the residuals
residuals_sq = np.square((predict - y).values)
# Regress X and e
model_summary = sm.OLS(residuals_sq, X).fit()
model_summary = model.summary().as_text().split('\n')
# removing dummies
for li in model_summary:
    if li.startswith('Notes'):
        break
    if not li.startswith('fe'):
        print(li)
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:             dformal_06   R-squared:
0.322
Model:                            OLS   Adj. R-squared:
0.093
Method:                 Least Squares   F-statistic:
37.42
Date:                Fri, 04 Dec 2020   Prob (F-statistic):
0.00
Time:                        22:15:11   Log-Likelihood:
-616.50
No. Observations:                1669   AIC:
2075.
Df Residuals:                    1248   BIC:
4357.
Df Model:                         420
Covariance Type:                  HC3
==============================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const          -0.1181      0.161     -0.734      0.463      -0.434
0.197
select          0.0538      0.026      2.076      0.038       0.003
0.105
empl_04        -0.1873      0.094     -1.987      0.047      -0.372
-0.003
pempl_04        0.0641      0.085      0.753      0.451      -0.103
0.231
contract_04     0.0863      0.086      1.005      0.315      -0.082
0.254
dformal_04      0.2431      0.089      2.716      0.007       0.068
0.418
salary_04    4.994e-08   2.02e-07      0.247      0.805   -3.46e-07
4.46e-07
profit_04    6.498e-09   3.98e-07      0.016      0.987   -7.73e-07
7.86e-07
tenure_04       0.0015      0.001      1.361      0.174      -0.001
0.004
days_04         0.0067      0.004      1.852      0.064      -0.000
0.014
hours_04       -0.0013      0.001     -1.084      0.278      -0.004
0.001
educ_lb         0.0340      0.009      3.854      0.000       0.017
0.051
age_lb         -0.0030      0.007     -0.452      0.651      -0.016
0.010
dmarried_lb    -0.0224      0.030     -0.749      0.454      -0.081
0.036
==============================================================================
=======
Omnibus:                      132.732   Durbin-Watson:
2.108
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
165.190
Skew:                           0.771   Prob(JB):
1.35e-36
Kurtosis:                       3.022   Cond. No.
1.44e+23
=======================================================================
=======
```

As we are seeing statistically significant correlations between the residuals and the X values (for example on select, and salary_04), we will be using heteroskedastic robust standard errors.

## Question 3

Here we estimate the treatment effect on employment and earnings for men and women. The exogenous variables are our treatment dummy and the baseline characteristics that we included above.

In [470]:

```
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'empl_06'
for gender in genders:
    if gender == "MALE":
        data = df_m[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    else:
        data = df_w[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    print(f'***{gender}***')
    model = sm.OLS(y, X).fit(cov_type = 'HC3')
    model_summary = model.summary().as_text().split('\n')
    # removing dummies
    for li in model_summary:
        if li.startswith('Notes'):
            break
        if not li.startswith('fe'):
            print(li)
```

```
***MALE***
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:                empl_06   R-squared:
0.353
Model:                            OLS   Adj. R-squared:
0.079
Method:                 Least Squares   F-statistic:
16.46
Date:                Fri, 04 Dec 2020   Prob (F-statistic):           1.
65e-258
Time:                        22:15:11   Log-Likelihood:
-322.31
No. Observations:                1329   AIC:
1437.
Df Residuals:                     933   BIC:
3493.
Df Model:                         395
Covariance Type:                  HC3
==============================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const           0.2419      0.346      0.698      0.485      -0.437
0.921
select         -0.0223      0.030     -0.746      0.456      -0.081
0.036
empl_04         0.0741      0.112      0.661      0.508      -0.146
0.294
pempl_04       -0.0094      0.100     -0.094      0.926      -0.206
0.187
contract_04     0.0673      0.059      1.138      0.255      -0.049
0.183
dformal_04     -0.0134      0.057     -0.236      0.813      -0.125
0.098
salary_04   -4.454e-08   1.86e-07     -0.239      0.811    -4.1e-07
3.21e-07
profit_04    1.572e-07   2.95e-07      0.533      0.594   -4.21e-07
7.35e-07
tenure_04       0.0013      0.001      0.926      0.354      -0.001
0.004
days_04         0.0041      0.004      0.995      0.320      -0.004
0.012
hours_04       -0.0009      0.001     -0.737      0.461      -0.003
0.002
educ_lb         0.0020      0.010      0.213      0.832      -0.017
0.021
age_lb          0.0183      0.007      2.569      0.010       0.004
0.032
dmarried_lb     0.0322      0.047      0.691      0.490      -0.059
0.123
==============================================================================
=======
Omnibus:                      159.925   Durbin-Watson:
```

```
2.179
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
221.693
Skew:                          -0.919   Prob(JB):
7.24e-49
Kurtosis:                       3.791   Cond. No.
5.84e+22
========================================================================
=======

***FEMALE***
                            OLS Regression Results
========================================================================
=======
Dep. Variable:                empl_06   R-squared:
0.329
Model:                            OLS   Adj. R-squared:
0.103
Method:                 Least Squares   F-statistic:
2236.
Date:                Fri, 04 Dec 2020   Prob (F-statistic):
0.00
Time:                        22:15:12   Log-Likelihood:
-788.39
No. Observations:                1669   AIC:
2419.
Df Residuals:                    1248   BIC:
4701.
Df Model:                         420
Covariance Type:                  HC3
========================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------
--------
const          -0.0845      0.189     -0.447      0.655      -0.455
0.286
select          0.0537      0.028      1.887      0.059      -0.002
0.109
empl_04        -0.0410      0.111     -0.370      0.711      -0.258
0.176
pempl_04        0.0673      0.098      0.690      0.490      -0.124
0.259
contract_04     0.0294      0.090      0.329      0.742      -0.146
0.205
dformal_04      0.0677      0.087      0.778      0.437      -0.103
0.238
salary_04    3.178e-08   1.99e-07      0.159      0.873   -3.59e-07
4.23e-07
profit_04    4.758e-07   4.33e-07      1.098      0.272   -3.73e-07
1.32e-06
tenure_04       0.0004      0.002      0.200      0.841      -0.003
0.004
days_04         0.0042      0.004      1.124      0.261      -0.003
0.012
hours_04       -0.0008      0.001     -0.675      0.500      -0.003
```

```
0.002
educ_lb          0.0229      0.011       2.136      0.033       0.002
0.044
age_lb          -0.0018      0.008      -0.237      0.812      -0.017
0.013
dmarried_lb     -0.0581      0.034      -1.692      0.091      -0.125
0.009
==============================================================================
=======
Omnibus:                       97.420   Durbin-Watson:
2.111
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
70.866
Skew:                          -0.402   Prob(JB):
4.09e-16
Kurtosis:                       2.391   Cond. No.
1.44e+23
==============================================================================
=======
```

The treatment effect for women is 0.0629 (~6.3%), and is statistically significant (t-value = 2.737). For men we do not get a statistically significant effect, and the coefficient obtained is slightly negative (but near 0 (-0.015)).

## For Salary

In [471]:

```
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'salary_06'
for gender in genders:
    if gender == "MALE":
        data = df_m[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    else:
        data = df_w[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    print(f'***{gender}***')
    model = sm.OLS(y, X).fit(cov_type = 'HC3')
    model_summary = model.summary().as_text().split('\n')
    # removing dummies
    for li in model_summary:
        if li.startswith('Notes'):
            break
        if not li.startswith('fe'):
            print(li)
```

```
***MALE***
                            OLS Regression Results                            
==============================================================================
=======
Dep. Variable:              salary_06   R-squared:                       
0.357
Model:                            OLS   Adj. R-squared:                  
0.085
Method:                 Least Squares   F-statistic:                     
24.99
Date:                Fri, 04 Dec 2020   Prob (F-statistic):              
0.00
Time:                        22:15:13   Log-Likelihood:                  
-17938.
No. Observations:                1329   AIC:                           3.
667e+04
Df Residuals:                     933   BIC:                           3.
872e+04
Df Model:                         395                                         
Covariance Type:                  HC3                                         
==============================================================================
========
                 coef    std err          z      P>|z|      [0.025      
0.975]
------------------------------------------------------------------------------
--------
const        8.501e+04   1.74e+05      0.489      0.625   -2.56e+05    
4.26e+05
select       1.814e+04   1.61e+04      1.129      0.259   -1.33e+04    
4.96e+04
empl_04     -1.808e+04   6.98e+04     -0.259      0.796   -1.55e+05    
1.19e+05
pempl_04     3.684e+04   6.38e+04      0.577      0.564   -8.83e+04    
1.62e+05
contract_04  1.297e+04   3.64e+04      0.356      0.722   -5.84e+04    
8.44e+04
dformal_04   1.618e+04   3.62e+04      0.447      0.655   -5.48e+04    
8.71e+04
salary_04       0.0791      0.105      0.753      0.451      -0.127      
0.285
profit_04       0.0978      0.237      0.412      0.680      -0.368      
0.563
tenure_04     481.9628    841.988      0.572      0.567   -1168.303    
2132.229
days_04      2032.9981   2430.587      0.836      0.403   -2730.865    
6796.861
hours_04     -647.7235    783.728     -0.826      0.409   -2183.802     
888.355
educ_lb      5025.1093   5116.372      0.982      0.326   -5002.796    
1.51e+04
age_lb       5548.2635   4544.316      1.221      0.222   -3358.432    
1.45e+04
dmarried_lb   2.94e+04    2.8e+04      1.050      0.294   -2.55e+04    
8.43e+04
==============================================================================
=======
Omnibus:                       38.231   Durbin-Watson:                   
```

```
2.254
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
91.520
Skew:                           0.008   Prob(JB):
1.34e-20
Kurtosis:                       4.285   Cond. No.
5.84e+22
========================================================================
=======

***FEMALE***
                            OLS Regression Results
========================================================================
=======
Dep. Variable:              salary_06   R-squared:
0.370
Model:                            OLS   Adj. R-squared:
0.158
Method:                 Least Squares   F-statistic:
19.82
Date:                Fri, 04 Dec 2020   Prob (F-statistic):
0.00
Time:                        22:15:13   Log-Likelihood:
-22370.
No. Observations:                1669   AIC:                           4.
558e+04
Df Residuals:                    1248   BIC:                           4.
786e+04
Df Model:                         420
Covariance Type:                  HC3
========================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------
--------
const       -5.678e+04   7.87e+04     -0.722      0.470   -2.11e+05
9.74e+04
select       3.175e+04   1.17e+04      2.725      0.006    8911.781
5.46e+04
empl_04     -4.126e+04   4.31e+04     -0.957      0.339   -1.26e+05
4.32e+04
pempl_04    -1.054e+04   3.92e+04     -0.269      0.788   -8.73e+04
6.62e+04
contract_04  2.786e+04   3.67e+04      0.759      0.448   -4.41e+04
9.98e+04
dformal_04   2.406e+04   3.68e+04      0.654      0.513    -4.8e+04
9.62e+04
salary_04       0.1582      0.099      1.605      0.108      -0.035
0.351
profit_04      -0.0977      0.184     -0.532      0.595      -0.458
0.262
tenure_04     897.5497    643.719      1.394      0.163    -364.117
2159.216
days_04      2048.3531   1708.892      1.199      0.231   -1301.014
5397.720
hours_04      213.3943    538.169      0.397      0.692    -841.399
```

```
1268.187
educ_lb        1.81e+04   4286.089      4.224      0.000    9702.983
2.65e+04
age_lb      -1860.2450   3146.305     -0.591      0.554   -8026.889
4306.399
dmarried_lb -2.649e+04   1.43e+04     -1.847      0.065   -5.46e+04
1617.051
======================================================================
=======
Omnibus:                       15.203   Durbin-Watson:
2.052
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
15.384
Skew:                           0.224   Prob(JB):
0.000456
Kurtosis:                       2.856   Cond. No.
1.44e+23
======================================================================
=======
```

Here we have statistically significant and positive effects of treatment women (31750 +04, t = 2.725), and an unsignificant, but positive coefficient for men (18140, t = 1.129)

## Overall

In [472]:

```
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'salary_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
model_summary = model.summary().as_text().split('\n')
# removing dummies
for li in model_summary:
    if li.startswith('Notes'):
        break
    if not li.startswith('fe'):
        print(li)
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:              salary_06   R-squared:
0.257
Model:                            OLS   Adj. R-squared:
0.126
Method:                 Least Squares   F-statistic:
8.525
Date:                Fri, 04 Dec 2020   Prob (F-statistic):           4.
70e-285
Time:                        22:15:14   Log-Likelihood:
-40617.
No. Observations:                2998   AIC:                          8.
214e+04
Df Residuals:                    2546   BIC:                          8.
485e+04
Df Model:                         451
Covariance Type:                  HC3
==============================================================================
========
                 coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const       2.315e+04   1.18e+05      0.196      0.845   -2.08e+05
2.54e+05
select      3.268e+04   8646.138      3.780      0.000    1.57e+04
4.96e+04
empl_04    -4.393e+04   3.45e+04     -1.272      0.204   -1.12e+05
2.38e+04
pempl_04    6718.0509   3.14e+04      0.214      0.831   -5.48e+04
6.83e+04
contract_04 1.129e+04   2.34e+04      0.482      0.630   -3.46e+04
5.72e+04
dformal_04   2.95e+04   2.24e+04      1.316      0.188   -1.44e+04
7.34e+04
salary_04      0.1769      0.062      2.857      0.004       0.056
0.298
profit_04      0.1410      0.133      1.059      0.289      -0.120
0.402
tenure_04    543.7351    564.369      0.963      0.335    -562.408
1649.878
days_04     1398.0363   1270.690      1.100      0.271   -1092.470
3888.542
hours_04     164.4700    415.883      0.395      0.692    -650.646
979.586
educ_lb     1.132e+04   2937.716      3.855      0.000    5567.152
1.71e+04
age_lb      2109.9358   2356.672      0.895      0.371   -2509.057
6728.929
dmarried_lb -3.13e+04   1.15e+04     -2.718      0.007   -5.39e+04   -
8732.698
==============================================================================
=======
Omnibus:                      165.998   Durbin-Watson:
2.095
```

```
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
347.889
Skew:                           0.372   Prob(JB):
2.86e-76
Kurtosis:                       4.494   Cond. No.
9.05e+07
======================================================================
=======
```

Overall we have a positive and significant effect of treatment (32680 at t = 3.780).

To test whether the impact is the same for men and women we will use the original dataframe, incorporating the female dummy, and interacting the select and female dummy variables. Then we perform a t-test on the coefficient dwomen*select.

In [473]:

```
# Create the ineraction
df = df.assign(select_w = lambda x: x.select * x.dwomen)

# For employment
variables = dummy_list + ['select_w', 'select', 'dwomen', 'empl_04', 'pempl_04', 'c
ontract_04', 'dformal_04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours
_04', 'educ_lb', 'age_lb', 'dmarried_lb']
target = 'empl_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
model_summary = model.summary().as_text().split('\n')
# removing dummies
for li in model_summary:
    if li.startswith('Notes'):
        break
    if not li.startswith('fe'):
        print(li)
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:                empl_06   R-squared:
0.257
Model:                            OLS   Adj. R-squared:
0.125
Method:                 Least Squares   F-statistic:
24.46
Date:                Fri, 04 Dec 2020   Prob (F-statistic):
0.00
Time:                        22:15:15   Log-Likelihood:
-1369.2
No. Observations:                2998   AIC:
3646.
Df Residuals:                    2544   BIC:
6373.
Df Model:                         453
Covariance Type:                  HC3
==============================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const           0.1673      0.246      0.679      0.497      -0.316
0.650
select_w        0.0771      0.035      2.177      0.029       0.008
0.146
select         -0.0198      0.025     -0.788      0.431      -0.069
0.029
dwomen         -0.1790      0.029     -6.245      0.000      -0.235
-0.123
empl_04         0.0121      0.069      0.177      0.860      -0.122
0.147
pempl_04        0.0330      0.059      0.559      0.576      -0.083
0.149
contract_04     0.0260      0.044      0.589      0.556      -0.061
0.113
dformal_04      0.0387      0.041      0.934      0.350      -0.043
0.120
salary_04    4.163e-08   1.15e-07      0.362      0.717   -1.84e-07
2.67e-07
profit_04    3.767e-07   1.94e-07      1.940      0.052   -3.89e-09
7.57e-07
tenure_04       0.0003      0.001      0.210      0.833      -0.002
0.003
days_04         0.0025      0.002      1.016      0.310      -0.002
0.007
hours_04       -0.0003      0.001     -0.368      0.713      -0.002
0.001
educ_lb         0.0136      0.006      2.111      0.035       0.001
0.026
age_lb          0.0105      0.005      2.264      0.024       0.001
0.020
dmarried_lb    -0.0376      0.025     -1.517      0.129      -0.086
0.011
```

```
==============================================================================
=======
Omnibus:                      219.610   Durbin-Watson:
2.114
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
241.081
Skew:                          -0.662   Prob(JB):
4.47e-53
Kurtosis:                       2.579   Cond. No.
9.05e+07
==============================================================================
=======
```

We have a positive, significant value on select_w, our interacted variable (0.0771, 2.177). This is consistent with the result that we have uncovered above, namely, that women have larger employment benefits from treatment then men do. Further, there is similarly no strong evidence here to say that men have employment benefits from participating in the program at all.

In [474]:

```
# For wages
variables = dummy_list + ['select_w', 'select', 'dwomen', 'empl_04', 'pempl_04', 'c
ontract_04', 'dformal_04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours
_04', 'educ_lb', 'age_lb', 'dmarried_lb']
target = 'salary_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
model_summary = model.summary().as_text().split('\n')
# removing dummies
for li in model_summary:
    if li.startswith('Notes'):
        break
    if not li.startswith('fe'):
        print(li)
```

```
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:              salary_06   R-squared:
0.275
Model:                            OLS   Adj. R-squared:
0.146
Method:                 Least Squares   F-statistic:
6.909
Date:                Fri, 04 Dec 2020   Prob (F-statistic):           2.
13e-227
Time:                        22:15:16   Log-Likelihood:
-40581.
No. Observations:                2998   AIC:                          8.
207e+04
Df Residuals:                    2544   BIC:                          8.
480e+04
Df Model:                         453
Covariance Type:                  HC3
==============================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const        3.484e+04   1.12e+05      0.311      0.756   -1.85e+05
2.55e+05
select_w     9277.3097   1.71e+04      0.544      0.587   -2.42e+04
4.27e+04
select       2.432e+04   1.35e+04      1.798      0.072   -2187.634
5.08e+04
dwomen      -7.653e+04   1.36e+04     -5.637      0.000   -1.03e+05    -
4.99e+04
empl_04      -3.97e+04   3.43e+04     -1.158      0.247   -1.07e+05
2.75e+04
pempl_04     1.185e+04   3.11e+04      0.381      0.704   -4.92e+04
7.29e+04
contract_04  1.936e+04    2.3e+04      0.840      0.401   -2.58e+04
6.45e+04
dformal_04   2.124e+04   2.21e+04      0.959      0.337   -2.22e+04
6.46e+04
salary_04       0.1157      0.064      1.822      0.068      -0.009
0.240
profit_04       0.0717      0.132      0.544      0.586      -0.186
0.330
tenure_04     648.4693    545.366      1.189      0.234    -420.428
1717.366
days_04      1513.8711   1263.113      1.199      0.231    -961.784
3989.527
hours_04      123.2553    411.749      0.299      0.765    -683.757
930.268
educ_lb      1.087e+04   2881.908      3.774      0.000    5226.554
1.65e+04
age_lb       3406.5783   2336.077      1.458      0.145   -1172.049
7985.206
dmarried_lb -1.568e+04   1.14e+04     -1.370      0.171   -3.81e+04
6751.988
```

```
==============================================================================
=======
Omnibus:                      147.783   Durbin-Watson:
2.091
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
308.526
Skew:                           0.333   Prob(JB):
1.01e-67
Kurtosis:                       4.424   Cond. No.
9.05e+07
==============================================================================
=======
```

Here we uncover a positive but insignificant relationship between the dummy select_w and salary outcomes. This implies that we cannot say with a high degree of certainty that the female treated cohort differs in a material way from the male cohort in terms of salary benefits. What this regression is showing is that both cohorts have a positive salary benefit from participating in the program.

## Part 5. Testing whether treatment effects are the same with and without accounting for fixed effects.

**using the following z-test:**

$$Z = \frac{\beta_1 - \beta_2}{\sqrt{(SE\beta_1)^2 + (SE\beta_2)^2}}$$

First testing with salary as our target

In [475]:

```
# Using salary as our target, first with fixed effects (dummy_list)
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'salary_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
unrestricted_coef = model.params['select']
unrestricted_stderr = model.bse['select']

# Now removing the fixed effect
variables = ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_04', 'salary_
04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_lb', 'dmarrie
d_lb']
target = 'salary_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
restricted_coef = model.params['select']
restricted_stderr = model.bse['select']

print(f"unrestricted coef and stderr: \ncoef: {unrestricted_coef}, stderr: {unrestr
icted_stderr}" )
print(f"restricted coef and stderr: \ncoef: {restricted_coef}, stderr: {restricted_
stderr}" )

Z = (restricted_coef - unrestricted_coef)/(unrestricted_stderr**2 + unrestricted_st
derr**2)**1/2
print("Z-Score =", Z)
```

```
unrestricted coef and stderr:
coef: 32682.6076388949, stderr: 8646.137896047987
restricted coef and stderr:
coef: 35797.529520567245, stderr: 7592.122734255852
Z-Score = 1.0417004523100241e-05
```

Using salary as our target, we find no evidence at all that including or discluding fixed effects impacts the results of the treatment effect. This implies that the dummies are likely uncorrelated with the effect of treatment on salary.
Testing again with employment as our target.

In [476]:

```
# Using employment as our target, first with fixed effects (dummy_list)
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'empl_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
unrestricted_coef = model.params['select']
unrestricted_stderr = model.bse['select']

# Now removing the fixed effect
variables = ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_04', 'salary_
04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_lb', 'dmarrie
d_lb']
target = 'empl_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
restricted_coef = model.params['select']
restricted_stderr = model.bse['select']

print(f"unrestricted coef and stderr: \ncoef: {unrestricted_coef}, stderr: {unrestr
icted_stderr}" )
print(f"restricted coef and stderr: \ncoef: {restricted_coef}, stderr: {restricted_
stderr}" )

Z = (restricted_coef - unrestricted_coef)/(unrestricted_stderr**2 + unrestricted_st
derr**2)**1/2
print("Z-Score =", Z)
```

```
unrestricted coef and stderr:
coef: 0.02987637883618876, stderr: 0.018316547201872577
restricted coef and stderr:
coef: 0.032936753972035465, stderr: 0.0160475480423568
Z-Score = 2.28048623179176
```

Here we find a statistically significant score (Z = 2.28). This implies that the dummies are quite probably correlated with the treatment effect in effecting employment outcomes. There are two potential reasons behind this.
First, to enter the pool of people who would be randomly chosen from, each course had a different acceptance rate. Thus, it is possible that variance between the classes where introduced at this pre-randomization faze.
Second, it is very plausible that there would be difference in outcomes based on the classes that each individual has taken. Perhaps, for example, taking a marketing course resulted in a higher improvement in employability then a hairdressing course.
Overall, it seems like a rational idea to keep the fixed effects incorporated in the regression.

# Question 6

As stated above, the fixed effects are designed to capture both the impact of variable selectivity between courses, as well as the different outcomes which might be associated with taking different courses.

In [477]:

```
df.shape[0]
```

Out[477]:

```
3237
```

In [478]:

```
# using a F-Test to see if the fixed effects have an impact, using employment as ou
r target.
# Resricted model
# Using employment as our target, first with fixed effects (dummy_list)
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'empl_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
URSS = model.ssr
URsquared = model.rsquared

# Now removing the fixed effect (Unrestricted)
variables = ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_04', 'salary_
04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_lb', 'dmarrie
d_lb']
target = 'empl_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
RRSS = model.ssr
RRsquared = model.rsquared
# Applying the F-stat formula
q = len(dummy_list)
N = df.shape[0]
k = q + len(variables)
Fstat = ((RRSS-URSS)/q)/(URSS/(N-k))
# Surival function
p_val = {f_test.sf(Fstat, q, N-k)}

print('***Employment***')
print(f'Unrestricted R-squared: {URsquared}')
print(f'Restricted R-squared: {RRsquared}')
print(f'This yields an Fstat of {Fstat}, with dof {q}, {N-k} for the numerator and
 denominator respectively')
print(f'The survival function for this Fstat (pvalue) = {p_val}')
```

```
***Employment***
Unrestricted R-squared: 0.24023153489691318
Restricted R-squared: 0.045415710955158284
This yields an Fstat of 1.630985001854482, with dof 438, 2786 for the n
umerator and denominator respectively
The survival function for this Fstat (pvalue) = {4.328148813842117e-13}
```

Doing the same for Salary

In [479]:

```
# using a F-Test to see if the fixed effects have an impact, using employment as ou
r target.
# Resricted model
# Using employment as our target, first with fixed effects (dummy_list)
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'salary_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
URSS = model.ssr
URsquared = model.rsquared

# Now removing the fixed effect (Unrestricted)
variables = ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_04', 'salary_
04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_lb', 'dmarrie
d_lb']
target = 'salary_06'
data = df[[target]+variables].dropna()
y = data[target]
X = sm.add_constant(data[variables])
model = sm.OLS(y, X).fit(cov_type = 'HC3')
RRSS = model.ssr
RRsquared = model.rsquared

# Applying the F-stat formula
q = len(dummy_list)
N = df.shape[0]
k = q + len(variables)
Fstat = ((RRSS-URSS)/q)/(URSS/(N-k))
# Surival function
p_val = {f_test.sf(Fstat, q, N-k)}

print(f'Unrestricted R-squared: {URsquared}')
print(f'Restricted R-squared: {RRsquared}')
print(f'This yields an Fstat of {Fstat}, with dof {q}, {N-k} for the numerator and
 denominator respectively')
print(f'The survival function for this Fstat (pvalue) = {p_val}')
```

```
Unrestricted R-squared: 0.2572686352951493
Restricted R-squared: 0.07626586953601466
This yields an Fstat of 1.5501026136807574, with dof 438, 2786 for the
numerator and denominator respectively
The survival function for this Fstat (pvalue) = {9.401756053713795e-11}
```

In both regressions, the pvalue on the fixed effects being insignificant is very small, and they also significantly reduce the R2 in both regressions, so we reject the null that the fixed effects are insignificant.

This is interesting given that the fixed effects were found to be uncorrelated with the effect of the treatment on salary. It is thus probable that the fixed effects in the salary regression are uncorrelated with the treatment variable, and so do not effect its outcome, but are still highly correlated with the outcome variable.

In the general case, this is how fixed effects can influence can be significant themselves but not matter for estimating the treatment effect - by being correlated with the dependent variable, but not the treatment variable.

## Problem 8

Here, the treatment unit is the individual. Further, in this case, we do not need to account for clustering when computing standard errors. The reason for this is because (under a few assumptions) we are selecting a random sample from our population. The assumption that is being made is that the population who have been accepted to be a part of this experiment is representative of the population that would be the target of future initiatives of this sort - i.e. they are of a high enough education/character standard to be accepted in to a training program and are willing to participate. If this assumption was not made, then the individuals in this report would indeed by highly clustered. However, as this is an initiative targeted at a specific cohort which very plausibly is fully represented by our sample, there is no need here.

## Problem 9

In [480]:

```
df
```

Out[480]:

| | age_s | dmarried_s | empl_06 | salary_06 | profit_06 | tenure_06 | days_06 | hours_06 | contract_( |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 22.0 | 0.0 | 1.0 | 0.0 | 240000.0 | 15.233334 | 22.0 | 84.0 | 0 |
| **1** | 22.0 | 0.0 | 1.0 | 116000.0 | 0.0 | 1.866667 | 10.0 | 14.0 | 0 |
| **2** | 24.0 | 0.0 | 1.0 | 650000.0 | 0.0 | 1.866667 | 28.0 | 91.0 | 0 |
| **3** | 24.0 | 0.0 | 1.0 | 408000.0 | 0.0 | 0.100000 | 28.0 | 48.0 | 0 |
| **4** | 22.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | |
| **3232** | 20.0 | 0.0 | 1.0 | 408000.0 | 0.0 | 3.866667 | 28.0 | 84.0 | 1 |
| **3233** | 20.0 | 0.0 | 1.0 | 320000.0 | 0.0 | 10.733334 | 30.0 | 46.0 | 0 |
| **3234** | 27.0 | 1.0 | 1.0 | 408000.0 | 0.0 | 12.566667 | 26.0 | 48.0 | 1 |
| **3235** | 26.0 | 1.0 | 1.0 | 408000.0 | 0.0 | 2.866667 | 20.0 | 40.0 | 1 |
| **3236** | 21.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0 |

3237 rows × 479 columns

In [481]:

```
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'dmarried_s'
for gender in genders:
    if gender == "MALE":
        data = df_m[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    else:
        data = df_w[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    print(f'***{gender}***')
    model = sm.OLS(y, X).fit(cov_type = 'HC3')
    model_summary = model.summary().as_text().split('\n')
    # removing dummies
    for li in model_summary:
        if li.startswith('Notes'):
            break
        if not li.startswith('fe'):
            print(li)
```

```
***MALE***
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:             dmarried_s   R-squared:
0.531
Model:                            OLS   Adj. R-squared:
0.332
Method:                 Least Squares   F-statistic:
5.454
Date:                Fri, 04 Dec 2020   Prob (F-statistic):          6.
08e-100
Time:                        22:15:20   Log-Likelihood:
-86.777
No. Observations:                1329   AIC:
965.6
Df Residuals:                     933   BIC:
3022.
Df Model:                         395
Covariance Type:                  HC3
==============================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const          -0.2674      0.146     -1.827      0.068      -0.554
0.019
select          0.0017      0.023      0.074      0.941      -0.044
0.047
empl_04        -0.0685      0.085     -0.808      0.419      -0.235
0.098
pempl_04       -0.0499      0.082     -0.605      0.545      -0.211
0.112
contract_04     0.1089      0.064      1.711      0.087      -0.016
0.234
dformal_04     -0.0717      0.059     -1.207      0.227      -0.188
0.045
salary_04   -1.275e-08   1.65e-07     -0.077      0.938   -3.36e-07
3.11e-07
profit_04   -8.293e-08   2.11e-07     -0.393      0.694   -4.97e-07
3.31e-07
tenure_04      -0.0007      0.002     -0.447      0.655      -0.004
0.002
days_04         0.0032      0.004      0.812      0.417      -0.005
0.011
hours_04        0.0015      0.001      1.202      0.230      -0.001
0.004
educ_lb        -0.0139      0.007     -1.859      0.063      -0.029
0.001
age_lb          0.0154      0.006      2.440      0.015       0.003
0.028
dmarried_lb     0.6764      0.049     13.818      0.000       0.580
0.772
==============================================================================
=======
Omnibus:                      305.153   Durbin-Watson:
```

```
2.054
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            
791.138
Skew:                           1.204   Prob(JB):                    1.
61e-172
Kurtosis:                       5.914   Cond. No.                    
5.84e+22
==============================================================================
=======

***FEMALE***
                            OLS Regression Results                            
==============================================================================
=======
Dep. Variable:             dmarried_s   R-squared:                   
0.570
Model:                            OLS   Adj. R-squared:              
0.425
Method:                 Least Squares   F-statistic:                 
2318.
Date:                Fri, 04 Dec 2020   Prob (F-statistic):          
0.00
Time:                        22:15:21   Log-Likelihood:              
-403.93
No. Observations:                1669   AIC:                         
1650.
Df Residuals:                    1248   BIC:                         
3932.
Df Model:                         420                                         
Covariance Type:                  HC3                                         
==============================================================================
========
                  coef    std err          z      P>|z|      [0.025      
0.975]
------------------------------------------------------------------------------
--------
const          -0.1379      0.149     -0.923      0.356      -0.431      
0.155
select          0.0047      0.023      0.209      0.835      -0.040      
0.049
empl_04         0.1016      0.086      1.186      0.236      -0.066      
0.270
pempl_04       -0.0754      0.079     -0.957      0.339      -0.230      
0.079
contract_04    -0.0647      0.062     -1.042      0.298      -0.186      
0.057
dformal_04      0.0846      0.067      1.266      0.206      -0.046      
0.216
salary_04   -2.472e-08   1.63e-07     -0.152      0.879   -3.44e-07      
2.95e-07
profit_04   -5.861e-07   3.81e-07     -1.537      0.124   -1.33e-06      
1.61e-07
tenure_04      -0.0001      0.001     -0.089      0.929      -0.003      
0.002
days_04         0.0013      0.003      0.414      0.679      -0.005      
0.008
hours_04       -0.0004      0.001     -0.408      0.683      -0.003      
```

```
0.002
educ_lb        -0.0011      0.009     -0.122      0.903      -0.019
0.017
age_lb          0.0063      0.006      1.034      0.301      -0.006
0.018
dmarried_lb     0.6777      0.028     24.318      0.000       0.623
0.732
========================================================================
=======
Omnibus:                      137.823   Durbin-Watson:
2.016
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
233.509
Skew:                           0.593   Prob(JB):
1.97e-51
Kurtosis:                       4.397   Cond. No.
1.44e+23
========================================================================
=======
```

The effect does not appear to be significant in either regressions t = (0.074, 0.202). For both genders, although we don't have significance, we do have positive coefficients in each case. It is entirely plausible that participation in this sort of a program would increase chance of marriage. The reason would be high wages, and job security, leading to potentially both more financial comfortability with starting a family and also high attractiveness in the dating market.

## Problem 10

Here we will analyze the effect of the program on formal employment.

In [482]:

```
variables = dummy_list + ['select', 'empl_04', 'pempl_04', 'contract_04', 'dformal_
04', 'salary_04', 'profit_04', 'tenure_04', 'days_04', 'hours_04', 'educ_lb', 'age_
lb', 'dmarried_lb']
target = 'dformal_06'
for gender in genders:
    if gender == "MALE":
        data = df_m[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    else:
        data = df_w[[target]+variables].dropna()
        y = data[target]
        X = sm.add_constant(data[variables])
    print(f'***{gender}***')
    model = sm.OLS(y, X).fit(cov_type = 'HC3')
    model_summary = model.summary().as_text().split('\n')
    # removing dummies
    for li in model_summary:
        if li.startswith('Notes'):
            break
        if not li.startswith('fe'):
            print(li)
```

```
***MALE***
                            OLS Regression Results
==============================================================================
=======
Dep. Variable:             dformal_06   R-squared:
0.318
Model:                            OLS   Adj. R-squared:
0.030
Method:                 Least Squares   F-statistic:
457.1
Date:                Fri, 04 Dec 2020   Prob (F-statistic):
0.00
Time:                        22:15:21   Log-Likelihood:
-672.93
No. Observations:                1329   AIC:
2138.
Df Residuals:                     933   BIC:
4194.
Df Model:                         395
Covariance Type:                  HC3
==============================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------------
--------
const           0.1458      0.404      0.361      0.718      -0.647
0.938
select          0.0541      0.036      1.483      0.138      -0.017
0.126
empl_04        -0.2027      0.141     -1.433      0.152      -0.480
0.075
pempl_04        0.0109      0.133      0.082      0.934      -0.249
0.271
contract_04     0.0919      0.093      0.992      0.321      -0.090
0.273
dformal_04      0.0666      0.089      0.749      0.454      -0.108
0.241
salary_04    3.038e-07    2.6e-07      1.170      0.242   -2.05e-07
8.13e-07
profit_04     3.37e-07   4.49e-07      0.750      0.453   -5.44e-07
1.22e-06
tenure_04       0.0020      0.002      0.872      0.383      -0.002
0.006
days_04         0.0031      0.005      0.593      0.553      -0.007
0.013
hours_04        0.0007      0.002      0.451      0.652      -0.002
0.004
educ_lb         0.0207      0.013      1.646      0.100      -0.004
0.045
age_lb          0.0037      0.009      0.405      0.686      -0.014
0.022
dmarried_lb     0.0569      0.070      0.816      0.415      -0.080
0.194
==============================================================================
=======
Omnibus:                      187.730   Durbin-Watson:
```

```
2.149
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
64.476
Skew:                           0.304   Prob(JB):
9.98e-15
Kurtosis:                       2.109   Cond. No.
5.84e+22
========================================================================
=======

***FEMALE***
                            OLS Regression Results
========================================================================
=======
Dep. Variable:             dformal_06   R-squared:
0.322
Model:                            OLS   Adj. R-squared:
0.093
Method:                 Least Squares   F-statistic:
37.42
Date:                Fri, 04 Dec 2020   Prob (F-statistic):
0.00
Time:                        22:15:22   Log-Likelihood:
-616.50
No. Observations:                1669   AIC:
2075.
Df Residuals:                    1248   BIC:
4357.
Df Model:                         420
Covariance Type:                  HC3
========================================================================
========
                  coef    std err          z      P>|z|      [0.025
0.975]
------------------------------------------------------------------------
--------
const          -0.1181      0.161     -0.734      0.463      -0.434
0.197
select          0.0538      0.026      2.076      0.038       0.003
0.105
empl_04        -0.1873      0.094     -1.987      0.047      -0.372
-0.003
pempl_04        0.0641      0.085      0.753      0.451      -0.103
0.231
contract_04     0.0863      0.086      1.005      0.315      -0.082
0.254
dformal_04      0.2431      0.089      2.716      0.007       0.068
0.418
salary_04    4.994e-08   2.02e-07      0.247      0.805   -3.46e-07
4.46e-07
profit_04    6.498e-09   3.98e-07      0.016      0.987   -7.73e-07
7.86e-07
tenure_04       0.0015      0.001      1.361      0.174      -0.001
0.004
days_04         0.0067      0.004      1.852      0.064      -0.000
0.014
hours_04       -0.0013      0.001     -1.084      0.278      -0.004
```

```
0.001
educ_lb          0.0340       0.009       3.854       0.000        0.017
0.051
age_lb          -0.0030       0.007      -0.452       0.651       -0.016
0.010
dmarried_lb     -0.0224       0.030      -0.749       0.454       -0.081
0.036
========================================================================
=======
Omnibus:                     132.732   Durbin-Watson:
2.108
Prob(Omnibus):                 0.000   Jarque-Bera (JB):
165.190
Skew:                          0.771   Prob(JB):
1.35e-36
Kurtosis:                      3.022   Cond. No.
1.44e+23
========================================================================
=======
```

For males, we have a positive but not statistically significant effect (0.0541, 1.4873). For women, we have a positive and statistically significant effect (.0538, 2.0768). This is consistent with the results of the paper.

## Part 10

In [483]:

```
# First calc standard deviation for outcome variable employment, since we are presu
pposing we don't know the outcomes of the
# experiment, using employment in period 04:
def getTestSize(yStd, c, power, a):
# Using the formula: N = ((2*sd(y)(ppf(1 - a)-ppf(1 - power)))/c)**2
    N = ((2*yStd*(st.norm.ppf(1-a)-st.norm.ppf(1-power)))/c)**2
    return N

N1 = getTestSize(tstd(df.empl_04), 0.03, 0.8 , 0.025)
N2 = getTestSize(tstd(df.empl_04), 0.03, 0.9 , 0.025)
N3 = getTestSize(tstd(df.empl_04), 0.03, 0.9 , 0.005)

print("N at power = 80%, sig level = 5%:", N1)
print("N at power = 90%, sig level = 5%:", N2)
print("N at power = 90%, sig level = 1%:", N3)

# For power 90%, sig level 5%
```

```
N at power = 80%, sig level = 5%: 8714.493529049863
N at power = 90%, sig level = 5%: 11666.234338537397
N at power = 90%, sig level = 1%: 16520.36055988788
```

### Part 11

In the first paper, the key finding was that the intervention had a significant positive effect on both employment and wages for women, with salaries increasing by close to 20%. For men, there was no significant effect found in employment and wages, with the only discernible change found being a shift from the informal to formal work sector. The second paper on the other hand examined the full data set (approx 10x bigger) and looking at a significantly larger time span found that men as well as women received lasting positive employment and wages effects from participating in the training program.

Looking at the cost-benefit analysis of the first paper, a 20% internal rate of return for women was found (with indeterminate effects for men), and in the second paper, a lower bound on internal rate of return of 10% was found. The main discrepancies between the two analysis were first of course different data, second a different cash discount rate (5% in the first paper versus 6% in the second), and lastly the second paper did not have access to the data of those working in the informal sector (which most likely, but not for certain, would have pushed up the IRR further). Other assumptions about costs are all similar. Importantly, the result were based on an assumption of benefits deprecating over time.

Another key finding in the second paper is that there is no evidence that the program resulted in displacement of other workers, which was not explicitly incorporated in the first paper but was implicitly included. Thus the assumptions of the first paper seem strong, and consistent with the second. Discrepancy is most likely due to natural variation in the data, a different discount rate, and conservative assumptions in the second paper.

**Part 12**

The first paper finds that apart from a shift from informal to formal employment, there are no significant observed treatment effects for men, which stands in stark contrast to the significant effects found for women. The follow up paper, using a significantly larger data set which allowed for much more accurate estimates over a longer period of time, found that the effects were indeed consistent across men and women. In the first paper, as can be seen in Figure 1 in paper 2, the smaller data set was resulting in much more variable estimates because of the smaller sample size. As opposed to the larger data set, which saw low variability in estimates due to the high sample size.

## Part 13 - Conclusion

Overall, this experiment provided a unique insight into the economic dynamics of training and education within developing countries, and a strong framework from which to build future educational interventions, particularly for those who belong to the lower economic deciles of society. The experiment showed the positive and long-term effects that training intervention can have in a developing society, particularly in boosting formal employment. High rates of formal employment are extremely critical for the healthy and fast growth of an economy. Only when an economy has a strong legal framework that is able to track and ensure business activity within an economy can an economy develop with the massive gains that come from fair, safe and open internal trade. Further, high rates of formal employment also increase government revenue and the ability to carry out interventions such as the one analyzed in this report.

There are a number of follow up experiments that would be useful to carry out in the light of the information provided by this experiment. Firstly, it would be very useful to have more granular information on the benefits brought by different types of training programs, as opposed to simply some training program (i.e. training in which technical fields results in the greatest benefit at the lowest cost - and without causing displacement). This would allow policy maker to make more informed choices and extract greater benefit at lower cost from running this type of intervention. Secondly, it would be an interesting experiment to see if the additional tax earnings created by introducing this program resulted in the costs of the program being offset, and in what sort of a time period that occurred. This would give a strong incentive to implement this type of training program.

Scaling this program up to the whole economy will of course have scaling issues. Namely, there would quickly be an over-qualification problem, with too many people earning training certificates in industries with limited demand. This would simply result in the displacement, without adding any real jobs to the economy. Second, this experiment was carried out on the assumption that only a certain subset of people will be selected to participate in training programs (the selectivity that occurred before randomization). Therefore these results to not generalize to the average person, so it's hard to say whether it would be as economical of a program when offered ubiquitously. Last, an over supply of trained workers would end up pushing wages down in many industries, without necessarily raising the wages of the informal jobs that these newly trained workers are attempting to exit out of.

In [ ]:

```

```