Econ203B HW1

Chris Ackerman*

January 19, 2021

Contents

*I worked on this problem set with Luna Shen.

1 Question 1

In our sample, we have the minimization problem

$$\begin{aligned}
&\beta_0 = \arg\min_{b\in\mathbb{R}^2} \mathbb{E}[(Y-(1,X)b)^2] \\
\implies &\beta_0 = \mathbb{E}[(1,X)'(1,X)]^{-1}\mathbb{E}[Y,(1,X)'].
\end{aligned}$$

Let's build the matrices we need to perform this calculation.

$$\begin{aligned}
\mathbb{E}\left[\begin{bmatrix}1\\X\end{bmatrix}[1,X]\right] &= \begin{bmatrix}1 & \mathbb{E}[X]\\ \mathbb{E}[X] & \mathbb{E}[X^2]\end{bmatrix} \\
&= \begin{bmatrix}1 & \frac{1}{2}\\ \frac{1}{2} & \frac{1}{3}\end{bmatrix}
\end{aligned}$$

The finite variance allows us to invert this matrix:

$$E[(1,X)'(1,X)]^{-1} = \begin{bmatrix}4 & -6\\ -6 & 12\end{bmatrix}.$$

Now, onto the next matrix. We're going to use the Law of Iterated Expectations for this one, since we know $\mathbb{E}[Y \mid X]$.

$$\begin{aligned}
\mathbb{E}[Y(1,X)'] &= \mathbb{E}\begin{bmatrix}\mathbb{E}[Y\mid X]\\ X\mathbb{E}[Y\mid X]\end{bmatrix} \\
&= \mathbb{E}\begin{bmatrix}X^2\\ X^3\end{bmatrix} \\
&= \begin{bmatrix}\frac{1}{3}\\ \frac{1}{4}\end{bmatrix}.
\end{aligned}$$

We can plug these matrices into our FOC formula:

$$\begin{aligned}
\beta_0 &= \mathbb{E}[(1,X)'(1,X)]^{-1}\mathbb{E}[Y,(1,X)'] && = \begin{bmatrix}4 & -6\\ -6 & 12\end{bmatrix}\begin{bmatrix}\frac{1}{3}\\ \frac{1}{4}\end{bmatrix} \\
&= \begin{bmatrix}-\frac{1}{6}\\ 1\end{bmatrix}
\end{aligned}$$

2 Question 2

$$
\begin{aligned}
\nabla\mathbb{E}[\mathbb{E}[Y \mid X] &= \left(\frac{\partial\mathbb{E}[Y \mid X]}{\partial X_1} \cdots \frac{\partial\mathbb{E}[Y \mid X]}{\partial X_d}\right)' \\
\mathbb{E}[\|\nabla\mathbb{E}[Y \mid X] - b\|^2] &= \mathbb{E}\left[\sum_{i=1}^{d}\left(\frac{\partial\mathbb{E}[Y \mid X]}{\partial X_i} - b_i\right)^2\right] \\
\frac{\partial}{\partial b_i}\mathbb{E}[\|\nabla\mathbb{E}[Y \mid X] - b\|^2] &= -2\left[\frac{\partial\mathbb{E}[Y \mid X]}{\partial x_i} - b_i\right] \\
&= 0 \\
\implies b^* &= \mathbb{E}\left[\frac{\partial\mathbb{E}[Y \mid X]}{\partial X_i}\right]
\end{aligned}
$$

This expression is the same as

$$b_0 = \nabla\mathbb{E}[\mathbb{E}[Y \mid X]].$$

TODO: Add counter example.

3 Question 3

a)

$$\begin{aligned}\mathbb{E}[Y_i \mid D_i] &= \mathbb{E}[D_i Y_i(1) + (1-D_i)Y_i(0) \mid D_i] \\ &= D_i \mathbb{E}[Y_i \mid D_i = 1] + (1-D_i)\mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= \mathbb{E}[Y_i \mid D_i = 0] + D_i(\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]) \\ &= \alpha_0 + D_i \beta_0\end{aligned}$$

Now define

$$\begin{aligned}\eta &= Y_i - \alpha_0 - D_i\beta_0 \\ &= Y_i - \mathbb{E}[Y_i \mid D_i] \\ \mathbb{E}[\eta \mid D_i] &= \mathbb{E}[Y_i - \alpha_0 - D_i\beta_0] \\ &= \mathbb{E}[Y_i \mid D_i] - \alpha_0 - D_i\beta_0 \\ &= \alpha_o + D_i\beta_0 - \alpha_0 - D_i\beta_0 \\ &= 0\end{aligned}$$

b)

$$\begin{aligned}\beta_0 &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] + \mathbb{E}[Y_i(0) \mid D_i = 0] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\ &= \mathbb{E}[Y_i - Y_i(0) \mid D_i = 1] + \mathbb{E}[Y_i(0) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]\end{aligned}$$

c) ATEU should be positive if college has a positive impact on earnings.

d) Selection bias should be positive. Regardless of whetehr they attended college, more talented individuals would have earned more, so we are conflating the effect of attending college with these individuals' innate abilities.

e) OLS is not consistent for ATE regardless of heterogeneity, because we will still have a bias term. Note that, even with heterogeneity, we are only trying to identify the *average* treatment effect.

4 Question 6

For this problem, it's easier to work with the demeaned data,

$$\tilde{\beta}_n = \arg\min_{b \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}((Y_i - \overline{Y}_n) - (X_i - \overline{X}_n)'b)^2.$$

Starting with the forward direction,

$$\begin{aligned}
R^2 = 1 \implies RSS &= 0 \\
\equiv 0 &= \sum_{i=1}^{n}((Y_i - \overline{Y}_n) - (X_i - \overline{X}_n)'\tilde{\beta}_n)^2 \\
0 &= Y_i - \overline{Y}_n - (X_i - \overline{X}_n)'\tilde{\beta}_n \ \forall i \\
Y_i &= \underbrace{\overline{Y}_n - \overline{X}'_n\tilde{\beta}_n}_{\alpha_0} + X'_i \underbrace{\tilde{\beta}_n}\, b_0 \forall i
\end{aligned}$$

To go the other way, suppose

$$Y_i = a_0 + X'_i b_0 \ \forall i$$

$$\begin{aligned}
\tilde{\beta}_n &= \arg\min_{b \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}((Y_i - \overline{Y}_n) - (X_i - \overline{X}_n)'b)^2 \\
&= \arg\min_{b \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}((X_i - \overline{X}_n)'b_0 - (X_i - \overline{X}_n)'b)^2.
\end{aligned}$$

The arg min for this expression is $b = b_0$.

$$\begin{aligned}
Y_i - \overline{Y}_n &= a_0 - a_0 + (X_i - \overline{X}_n)'b_0 \\
&= (X_i - \overline{X_n})'b_0 \\
&= (X_i - \overline{X_n})'\tilde{\beta}_n \\
\implies RSS &= \sum_{i=1}^{n}((Y_i - \overline{Y}_n) - (X_i - \overline{X}_n)'\tilde{\beta}_n)^2 = 0 \\
\implies R^2 &= 1
\end{aligned}$$

5 Question 8

a) See the `python` code below; the function that does this part of the problem is `drop_missing_observations`

b) The function that performs these calculations is `calculate_summary_statistics`. Our dataset contains 2620.0 boys. 2960 students were assigned to tracking schools. The average baseline original score was 0.028842416616841626, and our dataset contains 108 unique schools.

c) See the code below for the actual calculations; the code contains the outcome and covariates for each specification I report.

Table 1: Regression to estimate the treatment effect, run on the sample of only girls

Dep. Variable:	totalscore	**R-squared:**	0.005
Model:	OLS	**Adj. R-squared:**	0.004
Method:	Least Squares	**F-statistic:**	12.36
Date:	Mon, 18 Jan 2021	**Prob (F-statistic):**	0.000446
Time:	22:59:22	**Log-Likelihood:**	-3674.1
No. Observations:	2530	**AIC:**	7352.
Df Residuals:	2528	**BIC:**	7364.
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.0623	0.032	1.945	0.052	-0.001	0.125
tracking	0.1469	0.042	3.516	0.000	0.065	0.229

Omnibus:	185.750	**Durbin-Watson:**	1.427
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	172.304
Skew:	0.576	**Prob(JB):**	3.84e-38
Kurtosis:	2.447	**Cond. No.**	2.88

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 2: Regression to estimate the treatment effect, run on the sample of only boys

d)

Dep. Variable:	totalscore	**R-squared:**	0.003
Model:	OLS	**Adj. R-squared:**	0.002
Method:	Least Squares	**F-statistic:**	7.538
Date:	Mon, 18 Jan 2021	**Prob (F-statistic):**	0.00608
Time:	22:59:22	**Log-Likelihood:**	-3671.1
No. Observations:	2620	**AIC:**	7346.
Df Residuals:	2618	**BIC:**	7358.
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-0.0323	0.029	-1.115	0.265	-0.089	0.025
tracking	0.1063	0.039	2.746	0.006	0.030	0.182

Omnibus:	199.415	**Durbin-Watson:**	1.472
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	246.077
Skew:	0.748	**Prob(JB):**	3.67e-54
Kurtosis:	2.869	**Cond. No.**	2.79

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 3: Regression to estimate the treatment effect for both boys and girls, run on the whole sample

e)

Dep. Variable:	totalscore	**R-squared:**	0.007
Model:	OLS	**Adj. R-squared:**	0.007
Method:	Least Squares	**F-statistic:**	12.92
Date:	Mon, 18 Jan 2021	**Prob (F-statistic):**	2.09e-08
Time:	22:59:22	**Log-Likelihood:**	-7348.6
No. Observations:	5150	**AIC:**	1.471e+04
Df Residuals:	5146	**BIC:**	1.473e+04
Df Model:	3		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-0.0323	0.030	-1.086	0.277	-0.091	0.026
girl	0.0946	0.043	2.193	0.028	0.010	0.179
treated_boy	0.1063	0.040	2.676	0.007	0.028	0.184
treated_girl	0.1469	0.041	3.606	0.000	0.067	0.227

Omnibus:	351.267	**Durbin-Watson:**	1.399
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	397.937
Skew:	0.658	**Prob(JB):**	3.88e-87
Kurtosis:	2.647	**Cond. No.**	5.14

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

TODO: Add proof that this regression is giving us what we want.

$$
\begin{aligned}
Y_i &= \alpha_0 + \alpha_1 G_i + \beta_0 T_i \times (1 - G_i) + \beta_1 T_i \times G_i \\
\alpha_0 &= \text{ boy, untreated mean} \\
\alpha_1 &= \text{ girl, untreated mean} \\
\beta_0 &= \text{ boy, treatment effect} \\
\beta_1 &= \text{ girl, treatment effect}
\end{aligned}
$$

Let

$$
\begin{aligned}
Y_i(1) &= \text{ outcome with treatment} \\
\mathbb{E}[Y_i \mid T_i, G_i] &= (1 - T_i)(1 - G_i)\mathbb{E}[Y_i^{\text{boy}}(0) \mid T_i = 0, \text{ boy}] \\
&\quad + (1 - T_i)G_i\mathbb{E}[Y_i^{\text{girl}}(0) \mid T_i = 0, \text{ girl}] \\
&\quad + T_i(1 - G_i)\mathbb{E}[Y_i^{\text{boy}}(1) \mid T_i = 1, \text{ boy}] \\
&\quad + T_i G_i\mathbb{E}[Y_i^{\text{girl}}(1) \mid T_i = 1, \text{ girl}] \\
&= \underbrace{(1 - G_i)\mathbb{E}[Y_i^{\text{boy}(0)}]}_{\alpha_0} + \underbrace{G_i\mathbb{E}[Y_i^{\text{girl}(0)}]}_{\alpha_1} \\
&\quad + \underbrace{T_i(1 - G_i)\mathbb{E}[Y_i^{\text{boy}}(1) - Y_i^{\text{boy}}(0)]}_{\beta_0} + \underbrace{T_i G_i\mathbb{E}[Y_i^{\text{girl}}(1) - Y_i^{\text{girl}}(0)]}_{\beta_1}
\end{aligned}
$$

We can estimate these objects via OLS since conditional expectation is linear.

Table 4: Regression to estimate the treatment effect, run on the top half of the sample

f)

Dep. Variable:	totalscore	**R-squared:**	0.006
Model:	OLS	**Adj. R-squared:**	0.005
Method:	Least Squares	**F-statistic:**	14.60
Date:	Mon, 18 Jan 2021	**Prob (F-statistic):**	0.000136
Time:	22:59:22	**Log-Likelihood:**	-3748.5
No. Observations:	2642	**AIC:**	7501.
Df Residuals:	2640	**BIC:**	7513.
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	0.3882	0.030	13.133	0.000	0.330	0.446
tracking	0.1501	0.039	3.821	0.000	0.073	0.227

Omnibus:	200.720	**Durbin-Watson:**	1.468
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	116.240
Skew:	0.372	**Prob(JB):**	5.74e-26
Kurtosis:	2.291	**Cond. No.**	2.80

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Table 5: Regression to estimate the treatment effect, run on the bottom half of the sample

Dep. Variable:	totalscore	**R-squared:**	0.006
Model:	OLS	**Adj. R-squared:**	0.006
Method:	Least Squares	**F-statistic:**	15.49
Date:	Mon, 18 Jan 2021	**Prob (F-statistic):**	8.53e-05
Time:	22:59:22	**Log-Likelihood:**	-3139.7
No. Observations:	2508	**AIC:**	6283.
Df Residuals:	2506	**BIC:**	6295.
Df Model:	1		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
const	-0.3987	0.026	-15.225	0.000	-0.450	-0.347
tracking	0.1349	0.034	3.935	0.000	0.068	0.202

Omnibus:	387.199	**Durbin-Watson:**	1.504
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	588.956
Skew:	1.101	**Prob(JB):**	1.29e-128
Kurtosis:	3.885	**Cond. No.**	2.86

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```python
import numpy as np
import os
import pandas as pd
import scipy.io
import statsmodels.api as sm


def load_matlab_data(filename='DDKData.mat'):
    matlab_data = scipy.io.loadmat(filename)
    df = pd.DataFrame(
        columns=[data_field for data_field in matlab_data.keys() if data_field[0] != '_']
        )
    for column in df.columns:
        df[column] = matlab_data[column].flatten()
    return df


def drop_missing_observations(dataframe, obs_to_check=['girl', 'std_mark', 'totalscore', 'tra
    original_nObs = dataframe.shape[0]
    df = dataframe.dropna(subset=obs_to_check)
    print(f'We dropped {original_nObs - df.shape[0]} observations.')
    return df


def calculate_summary_statistics(dataframe, filename=False):
    num_boys = dataframe.shape[0] - dataframe.girl.sum()
    num_tracking = dataframe.tracking.sum()
    original_score = dataframe.std_mark.mean()
    unique_schools = len(dataframe.schoolid.unique())
    if filename:
        with open(filename, 'w') as text_file:
            print(f'Our dataset contains {num_boys} boys. \
                    {num_tracking} students were assigned to tracking schools. \
                    The average baseline original score was {original_score}, \
                    and our dataset contains {unique_schools} unique schools.',
                  file=text_file)


def prepare_datasets(dataframe):
    dataframe['const'] = 1
    girls = dataframe[dataframe.girl == 1]
    boys = dataframe[dataframe.girl == 0]
    dataframe['boy'] = pd.get_dummies(dataframe['girl'])[0.0]
    dataframe['treated_boy'] = dataframe['tracking'] * dataframe['boy']
    dataframe['treated_girl'] = dataframe['tracking'] * dataframe['girl']
```

```python
    top = dataframe[dataframe['tophalf'] == 1]
    bottom = dataframe[dataframe['bottomhalf'] == 1]
    return dataframe, girls, boys, top, bottom


def calculate_ATE(dataframe, outcome, exog, filename=False):
    reg = sm.OLS(endog=dataframe[outcome],
                 exog=dataframe[exog]
                 ).fit()
    if filename:
        with open(filename, 'w') as text_file:
            print(f'{reg.summary().as_latex()}', file=text_file)


if __name__ == '__main__':
    os.chdir('/home/chris/files/school/ucla/first_year/winter/203b/psets/pset1')
    df = load_matlab_data()
    df = drop_missing_observations(df)
    calculate_summary_statistics(df, filename='summary_statistics.tex')
    complete_dataset, girls, boys, top, bottom = prepare_datasets(df)
    girls_spec = [girls, 'totalscore', ['const', 'tracking'], 'girls.tex']
    boys_spec = [boys, 'totalscore', ['const', 'tracking'], 'boys.tex']
    all_spec = [complete_dataset, 'totalscore', ['const', 'girl', 'treated_boy', 'treated_gir
    top_spec = [top, 'totalscore', ['const', 'tracking'], 'top.tex']
    bottom_spec = [bottom, 'totalscore', ['const', 'tracking'], 'bottom.tex']
    specifications = [girls_spec, boys_spec, all_spec, top_spec, bottom_spec]
    for specification in specifications:
        calculate_ATE(*specification)
```