הפקולטה למדעי החברה ע״ש גרשון גורדון

החוג לכלכלה

עבודת גמר

קורס : מבוא לאקונומטריקה

ד״ר אנליה שלוסר וגב׳ נועה דה לה וגה

עידו קנדיוטי 318814449 נוי סגל סויסה 318851821

החוג לכלכלה

תואר ראשון

חלק א'

המאמר של קרוגר (1999) הוא מאמר אשר עוסק בשאלה כיצד גודל הכיתה משפיע על ההישגים הלימודיים של התלמידים בה. בנוסף המאמר בוחן כיצד הקצאת משאבים באופנים שונים על ידי בתי ספר משפיעה על הישגי התלמידים. המשתנה המסביר במאמר הוא משתנה קטגוריאלי אשר מגדיר האם מדובר על כיתה קטנה, כיתה גדולה או כיתה גדולה עם סייעות. המשתנה המוסבר במאמר הוא ציון מנורמל המשקלל בתוכו כמה מבחנים שונים הבוחנים את הישגי התלמידים. המשתתפים בניסוי היו תלמידים מכיתה א' עד כיתה ג', כאשר בכיתה א' התלמידים הוקצו לכיתה באופן אקראי והישגיהם נבחנו עד אשר הגיעו לכיתה ג'. בנוסף גם המורים שהשתתפו בניסוי הוקצו לכיתות באופן אקראי כך שמאפייני המורים לא ישפיעו על תוצאות הניסוי. הניסוי נערך על 11,600 תלמידים מ80 בתי ספר שונים אשר הוקצו לכיתות בצורה אקראית.

לפי המאמר, מחקרים דומים שנערכו בעבר הגיעו למסקנות חלשות מבחינה סטטיסטית או למסקנות כלליות. המאמר נותן כמה סיבות לקשיים הללו. ראשית, קשה לנסח פונקציית מדידה טובה לתופעה כיוון שגם בחירת המשתנה המוסבר היא בעייתית כיוון שיש גישות שונות של מדידת הצלחה לימודית וכל אחת מהן מסתמכת על מדידה של פרמטרים שונים וגם בחירת המשתנים המסבירים מורכבת כיוון שלא ברור אילו משתנים מסבירים את המשתנה המוסבר ואילו רק מתואמים איתו, כלומר צריך להחליט גם אילו משתנים ניתן להשמיט וגם לחשוב האם יש משתנים מסבירים חשובים אשר הושמטו מהמודל. אם המשתנה שהושמט משפיע על המשתנה המוסבר ומתואם עם משתנה שאינו הושמט יכולה להיווצר הטיה. הטיה אפשרית היא כאשר משתמשים במודל בו המשתנה המסביר הוא גודל הכיתה והמשתנה המוסבר הוא הציון הממוצע של התלמידים. כאשר בוחנים את המסבירים לציוני התלמידים ניתן לקחת בחשבון לדוגמה גם את מספר הסייעות בכיתה, מספר הסייעות בכיתה משפיע לטובה על ציוני התלמידים ככל הנראה כיוון שהמשמעות של יותר סייעות היא יותר יחס אישי לכל תלמיד בכיתה, אך המשתנה של מספר הסייעות גם מקיים קורלציה חיובית עם מספר התלמידים בכיתה, כלומר ניתן לשער שבכיתה בה יש יותר תלמידים גם יהיו יותר סייעות, לכן אם המודל ישמיט את המשתנה של מספר הסייעות ייגרם מצב בו האומד למקדם של מספר התלמידים בכיתה יהיה מוטה כלפי מעלה. על מנת להתגבר על הקשיים בהם נתקלו חוקרים אשר חקרו את הנושא הזה בעבר, המחקר ניתח ניסוי בגודל חסר תקדים בין 80 בתי ספר שונים ועל 11,600 תלמידים, אשר במהלכו הוא הקצה את התלמידים לכיתות מסוגים שונים בבתי ספר שונים בצורה אקראית, כך שמשתנים רבים אשר הפריעו לקבלת מסקנות חד משמעיות בעבר נוטרלו. קושי נוסף אשר עלה במהלך הניסוי הוא תנועה של תלמידים בין הכיתות לאחר תחילת הניסוי, אך במאמר נטען כי תנועה זו לא השפיעה באופן מובהק על התוצאות ובנוסף, המודלים במאמר הסתמכו על כמה משתנים הקשורים לגודל הכיתה כך שהתנועות לאחר ההקצאה הראשונית נלקחו בחשבון ולא הטו את תוצאות המודל.

לסיכום, המסקנה העיקרית של המאמר היא שלמידה בכיתה קטנה משפרת את הישגי התלמידים באופן מובהק והשיפור הוא לא רק נקודתי אלא מתמשך, כלומר השיפור בהישגים לא יהיה חד פעמי רק לאחר מעבר לכיתה קטנה אלא גם בשנים הבאות. בנוסף המאמר מכריע כי גודל הכיתה משמעותי יותר לשיפור הישגי התלמידים מאשר איכויות המורה וממספר הסייעות בכיתה. בעוד שתלמידים בכיתה קטנה שיפרו את הישגיהם בבין 5 ל7 נקודות האחוז, תלמידים שלמדו בכיתות עם מורים בעלי ניסיון רב הראו שיפור קל בהישגיהם אך תלמידים אשר למדו אצל מורים בעלי תואר שני ומעלה לא הראו איזשהו שיפור משמעותי בהישגים ותלמידים שלמדו בכיתה עם מספר סייעות שיפרו את הישגיהם רק בבין 1 ל2 נקודות האחוז.

חלק ב'

.1

	mean_col	sd_col	median_col	min_col	max_col
sbirthq	2.45352464129757	1.07909	2	1	4
sbirthy	1979.61410076431	0.56321	1980	1977	1981
cltype1	2.03570314936077	0.78688	2	1	3
schtype1	2.43607733083879	0.91382	3	1	4
trace1	1.16643115088711	0.37250	1	1	2
hdeg1	1.3611371446423	0.51277	1	1	4
totexp1	11.7605435801312	8.99298	11	0	42
treadss1	521.435934258816	55.16230	514	404	651
tmathss1	530.842269813975	43.12952	529	404	676
ses1	1.49664429530201	0.50003	1	1	2
schid1n	41.1239476145931	22.47330	41	1	80
score	525.926333021515	45.74180	522.5	413.5	663.5
cs	20.6774243841597	3.73902	22	12	27
sc1	0.292329279700655	0.45487	0	0	1
white	0.666718652526513	0.47142	1	0	1
black	0.326575171553337	0.46900	0	0	1
boy	0.5177736202058	0.49972	1	0	1

לפי ממצאי הטבלה ניתן לראות שגודל כיתה ממוצעת הוא 20.67 תלמידים. בכל כיתה כמות הבנים והבנות היא כמעט שווה. הניקוד הממוצע במבחן שנערך היה 525.92 והניקוד המקסימלי היה 663.5. בממוצע למורות היו כמעט 12 שנות ניסיון ולרובן היה תואר ראשון. יש פחות או יותר אותה כמות של תלמידים ממעמד סוציו-אקונומי גבוה ונמוך.

.2

```
lm(formula = score ~ boy + cs + cs_sqr + hdeg1 + totexp1 + ses1,
    data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-105.87  -29.90   -3.33   25.79  131.49

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 498.1078    70.2298    7.09      0.0000000000017 ***
boy          -5.7288     1.6601   -3.45              0.00057 ***
cs           -2.3024     6.2613   -0.37              0.71311
cs_sqr        0.0363     0.1396    0.26              0.79465
hdeg1         4.1636     1.8776    2.22              0.02668 *
totexp1       0.1699     0.0986    1.72              0.08492 .
ses1         34.3074     1.6644   20.61 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 40.5 on 2380 degrees of freedom
  (48 observations deleted due to missingness)
Multiple R-squared:  0.161,     Adjusted R-squared:  0.159
F-statistic: 76.1 on 6 and 2380 DF,  p-value: <0.0000000000000002
```

במודל אמדנו את גודל הכיתה כמשפיע על הציון של התלמידים בה. השתמשנו במספר משתנים מפקחים בכדי לנטרל את השפעתם מהמשתנה "גודל כיתה", כאשר מדובר על כיתה גדולה (cltype1=2).

מקדמים	גודל המקדם	הסבר	מובהקות המקדם
בטא 0	498.1078	ציון ממוצע התחלתי בהינתן כל המשתנים המפקחים 0	מובהק.
בטא 1	-5.7288	ההבדל בממוצע בציון בין תלמידה לתלמיד בהינתן כל שאר המשתנים המפקחים 0	מובהק
בטא 2	-2.3024	על כל תלמיד שמתווסף לכיתה הציון הממוצע יורד ב2.3024 נקודות בהינתן כל שאר המשתנים המפקחים 0	לא מובהק, כלומר גודל כיתה לא משנה בצורה מובהקת
בטא 3	0.0363	ההשפעה השולית של עליית גודל הכיתה בהינתן כל שאר המשתנים המפקחים 0	לא מובהק וגם שואף ל0, כלומר ההשפעה השולית אפסית ולא משנה את הציון באופן מובהק
בטא 4	4.1636	על כל השגת תעודה אקדמאית נוספת על ידי המורה, הציון הממוצע יעלה ב4.1636 נקודות בהינתן כל שאר המשתנים המפקחים 0 , אנו מניחים כי לכל תעודה יש השפעה קבועה.	מובהק
בטא 5	0.1699	על כל עלייה בשנה בניסיון של המורה, הציון הממוצע של התלמידים יעלה ב0.1699 בהינתן כל שאר המשתנים המפקחים 0	לא מובהק וגם שואף ל0, כלומר ההשפעה אפסית ולא משנה את הציון באופן מובהק
בטא 6	34.3074	ההפרש הממוצע בציוני המבחן בין מעמד סוציו-אקונומי נמוך לגבוה	מובהק

המשתנים שבחרנו מתארים בעינינו את המאפיינים שמשפיעים בצורה הכי משמעותית על גובה הציון. בנוסף, הכנו משתנה נוסף שהוא גודל הכיתה בריבוע, כדי לבחון האם הוספת תלמיד אחד לגדלי כיתות שונות משפיע באותה מידה.

3. המודל שתיארנו לא מבטא השפעה של גודל הכיתה על ציוני התלמידים, כיוון שקיבלנו שהמקדמים של המשתנים המקושרים לגודל הכיתה אינם מובהקים. הסיבה לכך לדעתנו היא שכאשר בוחנים כיתה גדולה, ההשפעה של הוספת תלמיד נוסף לכיתה אינה משמעותית לציון כמו המשתנים האחרים בהם השתמשנו.

4. הקבוצה המטופלת היא הילדים אשר הוקצו באופן אקראי לכיתות קטנות וקבוצת הביקורת תהיה מורכבת מהילדים אשר הוקצו באופן אקראי לכיתה גדולה.

5. השתמשנו במספר רגרסיות בהן אנו רוצים לבדוק הבדלים עיקריים בין קבוצת הטיפול לקבוצת הביקורת. בחרנו שתי תכונות שמתארות את המורים ושתי תכונות שמתארות את התלמידים.

רגרסיה	**הפרש בין קבוצת הביקורת לקבוצת הטיפול**	**סטיית תקן של ההפרש**	**הפרש לאחר שימוש באפקט קבוע (FE)**	**סטיית תקן לאחר FE**	**ממוצע**
Boy ~ sc1	-0.005 לא מובהק	0.013	-0.0033 לא מובהק	0.01408	0.51777
Hdeg1 ~ sc1	0.0115 לא מובהק	0.0141	0.009 לא מובהק	0.01217	1.36114
Totexp1 ~ sc1 *	0.672 מובהק	0.2455	0.905 מובהק	0.221	11.76054
Ses1 ~ sc1 *	0.029 מובהק	0.013	0.0308 מובהק	0.0308	1.49664

ממצאים מעניינים מהטבלה:

- בכיתות קטנות בממוצע למורות יש ניסיון גדול יותר מאשר בכיתות קטנות באופן מובהק.
- בכיתות קטנות בממוצע יש יותר תלמידים ממעמד סוציו-אקונומי גבוה מאשר בכיתות גדולות באופן מובהק.
- מתוך תוצאות המודל עם האפקט הקבוע שמנו לב שגם במדגם רוחב ללא פיקוח fe התוצאות יהיו זהות מבחינת מובהקות.
- מתוך עמודת הממוצעים ניתן לראות כי אין הבדלים משמעותיים בכמות הבנים והבנות ובין כמות התלמידים ממעמד סוציו-אקונומי גבוה לנמוך. דבר התורם לנו במהימנות התוצאות האחרות.

6. העובדה שההקצאה לכיתות בניסוי נעשתה בצורה אקראית תרם רבות לתקפות הניסוי. התוצאה שקיבלנו הינה צפויה מכיוון שאיננו יכולים לצפות מהם הגורמים העיקריים שישפיעו בין קבוצות

הטיפול על סמך האקראיות שבמדגם הרוחב. בנוסף, ניתן לראות כי למרות ההקצאה האקראית יש שוויון בין התכונות הנבדקות של האנשים (עמודת הממוצעים).

.7

```
Call:
   felm(formula = sc1 ~ black + white + boy + sbirthy + ses1 | schid1n,      data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-0.572 -0.304 -0.226  0.584  0.876 

Coefficients:
        Estimate Std. Error t value  Pr(>|t|)    
black   -0.03081    0.07173   -0.43     0.668    
white   -0.01903    0.06996   -0.27     0.786    
boy      0.00062    0.01139    0.05     0.957    
sbirthy  0.04892    0.01033    4.74 0.0000022 ***
ses1     0.02359    0.01373    1.72     0.086 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.446 on 6173 degrees of freedom
  (160 observations deleted due to missingness)
Multiple R-squared(full model): 0.056   Adjusted R-squared: 0.0438 
Multiple R-squared(proj model): 0.00473   Adjusted R-squared: -0.00817 
F-statistic(full model):4.58 on 80 and 6173 DF, p-value: <0.0000000000000002 
F-statistic(proj model): 5.87 on 5 and 6173 DF, p-value: 0.0000203
```

מקדמים	גודל המקדם	מובהקות המקדם
בטא 1	-0.03081	לא מובהק
בטא 2	-0.01903	לא מובהק
בטא 3	0.00062	לא מובהק
בטא 4 ***	0.04892	מובהק
בטא 5	0.02359	לא מובהק

חסרונות המודל:

מכיוון שמדובר במודל LPM ישנם מספר חסרונות:

- ההסתברות החזויה אינה חסומה בין 0 ל1 וזה לא אפשרי מכיוון שהסתברות חייבת להיות בין 0 ל1
- בגלל שלמשתנה המוסבר יש שני ערכים אז הטעויות יכולות לגדול או לקטון בהינתן X ספציפי, לכן תהיה לנו בעיה של הטרוסקדסטיות. כתוצאה מכך האומדים לא יעילים, השונות לא נכונה ולא נוכל להשתמש בבדיקות ההשערה, לכן נרצה לעשות תיקון WHITE, פתרון שטוב לנו כי אנו משתמשים במדגם גדול.

- ההפרעה המקרית אינה מתפלגת נורמלית מכיוון שלטעויות יהיו שני ערכים בהינתן ערך X ספציפי וכתוצאה מכך, תהיה לנו בעיה במבחני ההשערות. למזלנו, המדגם מספיק גדול כך שעל פי משפט הגבול המרכזי נוכל לבצע את מבחני ההשערה
- כדי לוודא את השערתנו הרצנו מבחן BP למודל וקיבלנו PV בגודל 0.000000023 כלומר ניתן לדחות את השערת ה0 ברמת מובהקות של 95%. כלומר עדות חזקה להטרוסקדסטיות.

8. $Yi = \beta0 + \beta1 * black + \beta2 * white + \beta3 * boy + \beta4 * sbirthy + \beta5 * scs1 + Ui$

$H0: \beta1 = 0, \beta2 = 0, \beta3 = 0, \beta4 = 0, \beta5 = 0 \rightarrow\rightarrow\rightarrow Yi = \beta0 + Ui$

$H1: else$

$$F_S = \frac{R_U^2 \,/\, q}{(1 - R_U^2) \,/\, (N - K)}$$

$R^2u = 0.00473$

$q = 5$

$(N-K) = 6173$

$Fs = 5.87$

$F_C \sim (q\,,\ N - K\,,\ 1 - \alpha)$

$Fc = 2.21$

כלל החלטה ברמת מובהקות של 95% : Fs>Fc לכן נדחה את H0 כלומר המודל מובהק.

אנו לומדים כי לרגרסיה שאמדנו יש כוח הסבר על ההסתברות להגיע לכיתה הקטנה. התוצאה אינה צפויה כיוון שמטרת ההקצאה האקראית לכיתות היא לנטרל את ההשפעות האלו על ההסתברות ללמוד בכיתה קטנה.

9. Score ~ Sc1 | schid1n

משתנה	גודל	משמעות	מובהקות
SC1	12.8	עבור תלמיד בכיתה קטנה הציון הממוצע גבוה ב12.8 נקודות לעומת תלמיד בכיתה גדולה.	מובהק

```
Call:
   felm(formula = score ~ sc1 + ses1 + ses1_sc1 + totexp1 | schid1n,      data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-157.29  -26.33   -1.97   24.72  126.28 

Coefficients:
         Estimate Std. Error t value            Pr(>|t|)    
sc1         8.158      3.463    2.36               0.019 *  
ses1       24.256      1.316   18.43 <0.0000000000000002 ***
ses1_sc1    2.434      2.172    1.12               0.262    
totexp1     0.120      0.063    1.90               0.057 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.6 on 6166 degrees of freedom
  (168 observations deleted due to missingness)
Multiple R-squared(full model): 0.297   Adjusted R-squared: 0.288 
Multiple R-squared(proj model): 0.0904   Adjusted R-squared: 0.0788 
F-statistic(full model):32.9 on 79 and 6166 DF, p-value: <0.0000000000000002
F-statistic(proj model):  153 on 4 and 6166 DF, p-value: <0.0000000000000002
```

משתנה	הסבר לבחירת המשתנה
Ses1	אנו מאמינים כי המצב הסוציו-אקונומי משפיע על הישגי התלמידים בגלל שמבחינת משאבים הם מוגבלים ביכולת שלהם לקבל עזרה חיצונית בלימודים. לדוגמה היכולת להיעזר במורה פרטי.
Ses1_sc1	הפרש ההפרשים. רצינו לבחון האם ההשפעה של המצאות בכיתה קטנה מצמצמת את הפערים בין תלמידים ממצב סוציו-אקונומי גבוה לנמוך.
Totexp1	רצינו לבחון מה גובה ההשפעה של שנות הניסיון של המורה על ההישגים של התלמידים בכיתה שלה.

$$Yi = \beta0 + \beta1 * sc1 + \beta2 * ses1 + \beta3 * ses1\_sc1 + \beta4 * totexp1 + Ui$$

$$H0: \beta1 = 0, \beta2 = 0, \beta3 = 0, \beta4 = 0 \;\rightarrow\rightarrow\rightarrow Yi = \beta0 + Ui$$

$$H1: else$$

$$F_S = \frac{R_U^2 \,/\, q}{(1 - R_U^2) \,/\, (N - K)}$$

Fs = 153

$$F_C \sim (q\ ,\ N-K\ ,\ 1-\alpha)$$

Fc = 2.21

החלטה: נדחה את H0 ברמת מובהקות של 95%. כלומר המודל מובהק.

```
> #hypothesis test for Statistical significance
> linearHypothesis(model12,c('sc1 = 0','ses1=0','ses1_sc1=0','totexp1=0'))
Linear hypothesis test

Hypothesis:
sc1 = 0
ses1 = 0
ses1_sc1 = 0
totexp1 = 0

Model 1: restricted model
Model 2: score ~ sc1 + ses1 + ses1_sc1 + totexp1 | schid1n

  Res.Df Df Chisq          Pr(>Chisq)
1   6170
2   6166  4   613 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

11. השוואה בין מודל 11 (סעיף 9) למודל 12 (סעיף 10)

	Model 1	Model 2
sc1	12.804	8.157
	(1.128)	(3.463)
ses1		24.256
		(1.316)
ses1_sc1		2.434
		(2.172)
totexp1		0.120
		(0.063)
Num.Obs.	6414	6246
R2	0.241	0.297
R2 Adj.	0.232	0.288

המודל הראשון הוא המודל המצומצם משאלה 9 והמודל השני הוא המודל המורחב משאלה 10.
בשני המודלים המקדם של המשתנה הדמי שמתאר את גודל הכיתה מובהק. ההשפעה של האומד

במודל השני קטנה לעומת המודל הראשון כיוון שבמודל הראשון חלק מההשפעה של המשתנים האחרים במודל המורחב נכנסה למשתנה הדמי של גודל הכיתה.
בנוסף ניתן לראות כי מדד טיב ההתאמה המתוקנן ADJUSTED-R במודל המורחב גבוה יותר כלומר המודל המורחב מסביר בצורה טובה יותר את הציון הממוצע. סטיות התקן במודל הקצר קטנות יותר מכיוון שבמודל זה יש פחות דרגות חופש.

12. המודל המדובר הוא המודל אשר אמדנו בסעיף 10. כלומר, כבר בחנו את ההשפעה של גודל הכיתה על תלמידים מרמה סוציו-אקונומית שונה וקיבלנו שהפרש ההפרשים הוא 2.434, כלומר שעבור תלמידים מרמה סוציו-אקונומית גבוהה מעבר לכיתה קטנה יוביל לעלייה גדולה יותר בהישגים.

$H0: \beta 3 = 0$

$H1: else$

```
> linearHypothesis(model12, c('ses1_sc1=0'))
Linear hypothesis test

Hypothesis:
ses1_sc1 = 0

Model 1: restricted model
Model 2: score ~ sc1 + ses1 + ses1_sc1 + totexp1 | schid1n

  Res.Df Df Chisq Pr(>Chisq)
1   6167
2   6166  1  1.26       0.26
```

ברמת מובהקות של 95% נקבל את H0 כלומר האומד אינו מובהק. כלומר, אין הבדל בהישגים של תלמידים ממעמד סוציו-אקונומי נמוך לגבוה כאשר הם מוקצים לכיתה קטנה.

<u>חלק ג'</u>

13. התנאים לתקפות משתנה עזר הם:

- משתנה עזר מתואם עם המשתנה האנדוגני. COV(cs,sc1)≠0

בדקנו תנאי זה בעזרת חישוב השונות המשותפת בין שני המשתנים וקיבלנו -1.462, בנוסף ביצענו רגרסיה כאשר cs הוא המוסבר ו sc1 הוא המסביר, השיפוע במודל מובהק. חייב להיות

מתאם בין ההקצאה לכיתה קטנה או גדולה לבין מספר הסטודנטים בכיתה, גם אם בפועל היו העברות בין הכיתות לאחר ההקצאות, סביר שהמתאם ישאר שונה מ0.

- משתנה עזר אינו מתואם עם ההפרעה האקראית COV(sc1,Ui)=0

כלומר sc1 הוא משתנה אקסוגני. אנו לא יכולים לבדוק את התנאי מכיוון שאין לנו נתונים על ההפרעות האקראיות. אין קשר בין גודל הכיתה לדברים אקראיים הקשורים לסטודנטים כגון מין, עדה וכו'. המדגם במאמר נערך כך שהתלמידים הוקצו בצורה אקראית לכיתות בגודל שונה ולכן אין תיאום בין גודל הכיתה להפרעות האקראיות.

- אין קשר ישיר בין משתנה העזר למשתנה המוסבר.

אנו מניחים כי הקצאה לכיתה קטנה או גדולה לא משפיעה באופן ישיר על ציון התלמידים. כמות התלמידים בכיתה היא שמשפיעה על הציון של התלמידים והיא מתואמת בצורה חלקית עם ההקצאה לכיתות.

.14

משתנה	גודל מקדם	סטיית תקן	הסבר
מודל score ~ sc1 בפיקוח אפקט קבוע (FE)	12.8	1.13	כאשר תלמיד מוקצה לכיתה קטנה ציונו יהיה גדול בממוצע מתלמיד שהוקצה לכיתה גדולה ב12.8 נקודות
מודל cs ~ sc1 בפיקוח אפקט קבוע (FE)	-7.1140	0.0338	תלמיד המוקצה לכיתה קטנה בממוצע כיתתו תהיה קטנה ב7.114 מתלמיד המוקצה לכיתה גדולה

```
> b1_yz = as.numeric(model13$coefficients[1])
> b1_xz = as.numeric(model14$coefficients[1])
> b1IV = b1_yz/b1_xz
> paste("beta IV = ",b1IV)
[1] "beta IV =  -1.79980513924389"
```

.15

משתנה	גודל מקדם	סטיית תקן
Sc1 ***מובהק	11.871	1.091
Black	-21.346	6.162
white	-6.213	6.01
Ses1	22.736	1.17
boy	-5.862	0.974

המקדם של ההקצאה ירד במעט ככל הנראה כיוון שההקצאה היא אקראית ואין קשר חזק בין ההקצאה לכיתה בגודל שונה לתכונות המפקחות.

.16 שלבי אמידת TSLS:

- משוואת הבסיס הינה: score ~ β1cs + β2black + β3white + β4ses1 + β5boy | schid1n
- משוואת השלב הראשון: אנו מריצים את המשתנה האנדוגני כפונקציה של המשתנים האקסוגניים הקיימים במודל ובנוסף את משתנה העזר שלנו.

Cs ~ α1sc1 + α2black + α3white + α4ses1 + α5boy | schid1n

כעת קיבלנו את $\widehat{cs}$, נציבו במשוואת השלב השני ונקבל:

score ~ β1$\widehat{cs}$ + β2black + β3white + β4ses1 + β5boy | schid1n

הסבר התהליך: השתמשנו בשיטת *TSLS*. בשלב הראשון אמדנו את המשתנה המסביר האנדוגני *CS* כפונקציה של משתנה העזר *SC*1 ושל יתר המשתנים המפקחים האקסוגניים ואת האפקט הקבוע של כל בית ספר (*FE*).

בשלב השני אמדנו את המשתנה המוסבר *SCORE* כפונקציה של האמידה משלב 1 ($\widehat{cs}$) ויתר המשתנים המפקחים האקסוגניים.

שיטה זו הניבה לנו אומד מוטה אך עקיב.

משתנה	גודל מקדם	סטיית תקן	הסבר
$\widehat{cs}$***	-1.674	0.154	כשגודל הכיתה עולה ב1 הציון הממוצע של תלמיד בכיתה יורד ב1.674. האומד מובהק סטטיסטית.
*Black***	-21.072	6.162	כשהתלמיד הוא שחור, הציון הממוצע שלו קטן ב21.072 בממוצע מתלמיד

			שאינו שחור. האומד מובהק סטטיסטית.
White	-5.851	6.0161	כשהתלמיד הוא לבן, הציון הממוצע שלו קטן ב5.851 בממוצע מתלמיד שאינו לבן. כיוון שהאומד אינו מובהק ברמת מובהקות של 95% אנו יכולים לראות בהפרש בין תלמיד לבן לתלמיד שאינו לבן כדבר לא מוחלט
Ses1	22.696	1.17	הציון הממוצע עבור תלמיד מרמה סוציו-אקונומית גבוהה גדול מהציון הממוצע עבור תלמיד מרמה סוציו-אקונומית נמוכה ב22.696 נקודות בממוצע. האומד מובהק סטטיסטית.
Boy	-5.9	0.974	הציון הממוצע עבור תלמידים נמוך ב5.9 נקודות בממוצע מהציון הממוצע עבור תלמידות . האומד מובהק סטטיסטית.

- ניתן היה להשתמש בפקודת *CLUSTER* במידה והיינו מאמינים שיש גורם נוסף שמתואם ההפרעות של קוד בית ספר או מתואם עם ההפרעות הרגילות, במודלים שאמדנו לא השתמשנו, אך במידה והיינו משתמשים היינו מקבלים את שונויות האומדים בצורה נכונה ומותאמת להפרת ההנחה שההפרעות האקראיות אינן מתואמות בין התצפיות השונות.
לאחר תיקון זה, האומדים לא ישתנו אך השונויות כן.

final_project_R_appendix

Noy & Ido

2023-02-18

```
knitr::opts_chunk$set(echo = TRUE)
#Authors - Ido Candiotti &  Noy Segal Swisa

# Settings -------------------------------------------------------
-------

rm(list=ls()) # del all objects and functions
gc() #cleans memory

##          used (Mb) gc trigger (Mb) limit (Mb) max used (Mb)
## Ncells 533002 28.5    1188224 63.5         NA   669277 35.8
## Vcells 958606  7.4    8388608 64.0      16384  1839756 14.1

options(scipen=999) # tell R not to use Scientific notation
options(digits = 5) # controls how many digits are printed by defaul
t
options(na.action=na.exclude)


# Import ---------------------------------------------------------
-------
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

library(ggplot2)
library(car)

## Loading required package: carData

##
## Attaching package: 'car'

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

```
library(sandwich)
library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(lfe)

## Loading required package: Matrix

##
## Attaching package: 'lfe'

## The following object is masked from 'package:lmtest':
##
##     waldtest

library(stargazer)

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and S
ummary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=star
gazer

library(AER)

## Loading required package: survival

library(rstatix)

##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##     filter

library(plotrix)
library(lmtest)
library(AER)
library(modelsummary)


#Data-------------------------------------------------------------
```

```
-------
data <- read.csv("~/Desktop/TAU/Economics/Econometrics/data/term_pap
er.csv")
#summary(data)

#Part B

#Summary and standard deviation ------------------------------------
-------------------------------------

summary_df = sapply(data, summary)
summary_df = sapply(summary_df,head)

min_col <- summary_df[1, ]
median_col <- summary_df[3, ]
mean_col <- summary_df[4, ]
max_col <- summary_df[6, ]
sd_col <- apply(data, 2, sd , na.rm = T)

#q1 <- summary_stats[2, ] #25%
#q3 <- summary_stats[5, ] #75%

summary_df = data.frame(mean_col,sd_col,median_col,min_col,max_col,
row.names = names(data))
summary_df <- slice(summary_df, -1)
summary_df

##            mean_col   sd_col median_col min_col max_col
## sbirthq     2.45352  1.07909        2.0     1.0     4.0
## sbirthy  1979.61410  0.56321     1980.0  1977.0  1981.0
## cltype1     2.03570  0.78688        2.0     1.0     3.0
## schtype1    2.43608  0.91382        3.0     1.0     4.0
## trace1      1.16643  0.37250        1.0     1.0     2.0
## hdeg1       1.36114  0.51277        1.0     1.0     4.0
## totexp1    11.76054  8.99298       11.0     0.0    42.0
## treadss1  521.43593 55.16230      514.0   404.0   651.0
## tmathss1  530.84227 43.12952      529.0   404.0   676.0
## ses1        1.49664  0.50003        1.0     1.0     2.0
## schid1n    41.12395 22.47330       41.0     1.0    80.0
## score     525.92633 45.74180      522.5   413.5   663.5
## cs         20.67742  3.73902       22.0    12.0    27.0
## sc1         0.29233  0.45487        0.0     0.0     1.0
## white       0.66672  0.47142        1.0     0.0     1.0
## black       0.32658  0.46900        0.0     0.0     1.0
## boy         0.51777  0.49972        1.0     0.0     1.0

#main_varibles = c('score','sc1','black','white','boy','cltype1',
                  #'schtype1','hdeg1','totexp1','ses1','cs')
#df_subset <- summary_df[row.names(summary_df) %in% main_varibles,]
#df_subset
```

```
#Question 2-------------------------------------------------------
----------------

data$cs_sqr = data$cs * data$cs
data1=subset(data, data$cltype1=="2")

model1 = lm(score ~ boy + cs + cs_sqr + hdeg1 + totexp1 + ses1 , dat
a1)
summary(model1)

##
## Call:
## lm(formula = score ~ boy + cs + cs_sqr + hdeg1 + totexp1 + ses1,
##     data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -105.87  -29.90   -3.33   25.79  131.49
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 498.1078    70.2298    7.09      0.0000000000017 ***
## boy          -5.7288     1.6601   -3.45              0.00057 ***
## cs           -2.3024     6.2613   -0.37              0.71311
## cs_sqr        0.0363     0.1396    0.26              0.79465
## hdeg1         4.1636     1.8776    2.22              0.02668 *
## totexp1       0.1699     0.0986    1.72              0.08492 .
## ses1         34.3074     1.6644   20.61 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.5 on 2380 degrees of freedom
##   (48 observations deleted due to missingness)
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.159
## F-statistic: 76.1 on 6 and 2380 DF,  p-value: <0.0000000000000002

#Question 5 -------------------------------------------------------
-----------------

model2 = lm(boy ~ sc1  , data)
model3 = felm(boy ~ sc1 | schid1n , data)
fixed_effects = getfe(model3)
summary(model2)

##
## Call:
## lm(formula = boy ~ sc1, data = data)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.519 -0.519  0.481  0.481  0.486
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.51928    0.00742   70.00 <0.0000000000000002 ***
## sc1         -0.00514    0.01372   -0.37                0.71
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5 on 6412 degrees of freedom
## Multiple R-squared:  2.19e-05,   Adjusted R-squared:  -0.000134
## F-statistic: 0.141 on 1 and 6412 DF,  p-value: 0.708
```

```
summary(model3)
```

```
##
## Call:
##    felm(formula = boy ~ sc1 | schid1n, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.709 -0.516  0.397  0.481  0.611
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## sc1 -0.00338    0.01408   -0.24     0.81
##
## Residual standard error: 0.5 on 6337 degrees of freedom
## Multiple R-squared(full model): 0.00947   Adjusted R-squared: -0.
00241
## Multiple R-squared(proj model): 9.1e-06   Adjusted R-squared: -0.
012
## F-statistic(full model):0.797 on 76 and 6337 DF, p-value: 0.901
## F-statistic(proj model): 0.0577 on 1 and 6337 DF, p-value: 0.81
```

```
model4 = lm(hdeg1 ~ sc1  , data)
model5 = felm(hdeg1 ~ sc1 | schid1n , data)
fixed_effects = getfe(model5)
summary(model4)
```

```
##
## Call:
## lm(formula = hdeg1 ~ sc1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.369 -0.358 -0.358  0.631  2.642
##
```

```
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  1.35779    0.00761  178.39 <0.0000000000000002 ***
## sc1          0.01151    0.01411    0.82                0.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.513 on 6400 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.000104,   Adjusted R-squared:  -5.23e-05
## F-statistic: 0.665 on 1 and 6400 DF,  p-value: 0.415
```

```
summary(model5)
```

```
##
## Call:
##    felm(formula = hdeg1 ~ sc1 | schid1n, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.333 -0.292 -0.152  0.324  1.677
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## sc1  0.00957    0.01217    0.79     0.43
##
## Residual standard error: 0.432 on 6325 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared(full model): 0.298   Adjusted R-squared: 0.289
## Multiple R-squared(proj model): 9.78e-05   Adjusted R-squared: -0
.0119
## F-statistic(full model):35.3 on 76 and 6325 DF, p-value: <0.00000
00000000002
## F-statistic(proj model): 0.618 on 1 and 6325 DF, p-value: 0.432
```

```
model6 = lm(totexp1 ~ sc1  , data)
model7 = felm(totexp1 ~ sc1 | schid1n , data)
fixed_effects = getfe(model7)
summary(model6)
```

```
##
## Call:
## lm(formula = totexp1 ~ sc1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.24  -7.57  -1.24   5.43  30.43
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
```

```
## (Intercept)   11.565      0.133   86.68 <0.0000000000000002 ***
## sc1            0.672      0.247    2.72              0.0066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.99 on 6400 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.00115,    Adjusted R-squared:  0.000995
## F-statistic: 7.37 on 1 and 6400 DF,  p-value: 0.00663
```

```
summary(model7)
```

```
##
## Call:
##    felm(formula = totexp1 ~ sc1 | schid1n, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -23.84  -5.32  -1.08   4.06  27.33
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## sc1    0.905      0.221     4.1 0.000042 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.84 on 6325 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared(full model): 0.249   Adjusted R-squared: 0.24
## Multiple R-squared(proj model): 0.00265   Adjusted R-squared: -0.
00933
## F-statistic(full model):27.5 on 76 and 6325 DF, p-value: <0.00000
00000000002
## F-statistic(proj model): 16.8 on 1 and 6325 DF, p-value: 0.000042
```

```
model8 = lm(ses1 ~ sc1  , data)
model9 = felm(ses1 ~ sc1 | schid1n , data)
fixed_effects = getfe(model9)
summary(model8)
```

```
##
## Call:
## lm(formula = ses1 ~ sc1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.517 -0.488 -0.488  0.512  0.512
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
```

```
## (Intercept)  1.48800    0.00752  197.81 <0.0000000000000002 ***
## sc1          0.02937    0.01387    2.12               0.034 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5 on 6256 degrees of freedom
##   (156 observations deleted due to missingness)
## Multiple R-squared:  0.000717,   Adjusted R-squared:  0.000557
## F-statistic: 4.49 on 1 and 6256 DF,  p-value: 0.0342
```

```
summary(model9)
```

```
##
## Call:
##    felm(formula = ses1 ~ sc1 | schid1n, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9924 -0.3988  0.0078  0.3731  0.9951
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## sc1   0.0308     0.0122    2.52    0.012 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.428 on 6181 degrees of freedom
##   (156 observations deleted due to missingness)
## Multiple R-squared(full model): 0.275   Adjusted R-squared: 0.266
## Multiple R-squared(proj model): 0.00103   Adjusted R-squared: -0.
0113
## F-statistic(full model):30.8 on 76 and 6181 DF, p-value: <0.00000
00000000002
## F-statistic(proj model): 6.37 on 1 and 6181 DF, p-value: 0.0116
```

```
modelsummary(models = list(model2, model4, model6, model8))
```

```
## Warning in w * res^2: longer object length is not a multiple of s
horter object
## length
```

```
## Warning in log(w) - (log(2 * pi) + log(s2) + (w * res^2)/s2): lon
ger object
## length is not a multiple of shorter object length
```

```
## Warning in w * res^2: longer object length is not a multiple of s
horter object
## length
```

```
## Warning in log(w) - (log(2 * pi) + log(s2) + (w * res^2)/s2): lon
ger object
## length is not a multiple of shorter object length

## Warning in w * res^2: longer object length is not a multiple of s
horter object
## length

## Warning in log(w) - (log(2 * pi) + log(s2) + (w * res^2)/s2): lon
ger object
## length is not a multiple of shorter object length
```

	Model 1	Model 2	Model 3	Model 4
(Intercept)	0.519	1.358	11.565	1.488
	(0.007)	(0.008)	(0.133)	(0.008)
sc1	-0.005	0.012	0.672	0.029
	(0.014)	(0.014)	(0.247)	(0.014)
Num.Obs.	6414	6402	6402	6258
R2	0.00002	0.0001	0.001	0.0007
R2 Adj.	-0.0001	-0.00005	0.001	0.0006
AIC	9308.2			
BIC	9328.5			
Log.Lik.	-4651.100			
F	0.141	0.665	7.374	4.488
RMSE	0.50	0.51	8.99	0.50

```r
modelsummary(models = list(model3, model5, model7, model9))

 #fe models
```

	Model 1	Model 2	Model 3	Model 4
sc1	-0.003	0.010	0.905	0.031
	(0.014)	(0.012)	(0.221)	(0.012)
Num.Obs.	6414	6402	6402	6258
R2	0.009	0.298	0.249	0.275
R2 Adj.	-0.002	0.289	0.240	0.266
AIC	9397.3	7508.0	44616.1	7228.8
BIC	9925.1	8035.7	45143.7	7754.7
RMSE	0.50	0.43	7.79	0.43

```r
# Question 7 -----------------------------------------------------------
-------

model10 = felm(sc1 ~ black + white + boy + sbirthy + ses1 | schid1n
, data )
summary(model10)

##
## Call:
##    felm(formula = sc1 ~ black + white + boy + sbirthy + ses1 | sc
hid1n,      data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.572 -0.304 -0.226  0.584  0.876
##
## Coefficients:
##         Estimate Std. Error t value  Pr(>|t|)
## black   -0.03081    0.07173   -0.43     0.668
## white   -0.01903    0.06996   -0.27     0.786
## boy      0.00062    0.01139    0.05     0.957
## sbirthy  0.04892    0.01033    4.74 0.0000022 ***
## ses1     0.02359    0.01373    1.72     0.086 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.446 on 6173 degrees of freedom
##   (160 observations deleted due to missingness)
## Multiple R-squared(full model): 0.056   Adjusted R-squared: 0.043
8
```

```
## Multiple R-squared(proj model): 0.00473   Adjusted R-squared: -0.
00817
## F-statistic(full model):4.58 on 80 and 6173 DF, p-value: <0.00000
00000000002
## F-statistic(proj model): 5.87 on 5 and 6173 DF, p-value: 0.000020
3
```

```
#fixed_effects = getfe(model10)

#hypothesis test for Statistical significance
linearHypothesis(model10,c('black = 0','white=0','boy=0','sbirthy=0'
,'ses1=0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## black = 0
## white = 0
## boy = 0
## sbirthy = 0
## ses1 = 0
##
## Model 1: restricted model
## Model 2: sc1 ~ black + white + boy + sbirthy + ses1 | schid1n
##
##   Res.Df Df Chisq Pr(>Chisq)
## 1   6178
## 2   6173  5  29.4    0.00002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Breusch-Pagan test
bptest(model10)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model10
## BP = 44, df = 5, p-value = 0.000000023
```

```
#White correction for heteroscedasticity
coeftest(model10,vcov = vcovHC(model10,type = "HC"))
```

```
##
## t test of coefficients:
##
##         Estimate Std. Error t value   Pr(>|t|)
## black   -0.03081    0.07226   -0.43       0.67
## white   -0.01903    0.07072   -0.27       0.79
## boy      0.00062    0.01136    0.05       0.96
## sbirthy  0.04892    0.00978    5.00 0.00000058 ***
```

```
## ses1     0.02359    0.01389    1.70       0.09 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#modelsummary(models = list(model10, coeftest(model10,vcov = vcovHC(
model10,type = "HC"))))

# Question 9 ----------------------------------------------------
-------

model11 = felm(score ~ sc1 | schid1n, data)
summary(model11)
```

```
##
## Call:
##    felm(formula = score ~ sc1 | schid1n, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -143.95  -27.15   -2.06   25.34  137.06
##
## Coefficients:
##     Estimate Std. Error t value            Pr(>|t|)
## sc1    12.80       1.13    11.3 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.1 on 6337 degrees of freedom
## Multiple R-squared(full model): 0.241   Adjusted R-squared: 0.232
## Multiple R-squared(proj model): 0.0199   Adjusted R-squared: 0.00
816
## F-statistic(full model):26.5 on 76 and 6337 DF, p-value: <0.00000
00000000002
## F-statistic(proj model):  129 on 1 and 6337 DF, p-value: <0.00000
00000000002
```

```
fixed_effects = getfe(model11)

# Questions 10 + 11 + 12 ----------------------------------------
-------------------

data$ses1_sc1 = data$ses1 * data$sc1
model12 = felm(score ~ sc1 + ses1 + ses1_sc1 + totexp1  |schid1n , d
ata)
summary(model12)
```

```
##
## Call:
##    felm(formula = score ~ sc1 + ses1 + ses1_sc1 + totexp1 | schid
1n,      data = data)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -157.29  -26.33   -1.97   24.72  126.28
##
## Coefficients:
##          Estimate Std. Error t value            Pr(>|t|)
## sc1         8.158      3.463    2.36               0.019 *
## ses1       24.256      1.316   18.43 <0.0000000000000002 ***
## ses1_sc1    2.434      2.172    1.12               0.262
## totexp1     0.120      0.063    1.90               0.057 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.6 on 6166 degrees of freedom
##   (168 observations deleted due to missingness)
## Multiple R-squared(full model): 0.297   Adjusted R-squared: 0.288
## Multiple R-squared(proj model): 0.0904   Adjusted R-squared: 0.07
88
## F-statistic(full model):32.9 on 79 and 6166 DF, p-value: <0.00000
00000000002
## F-statistic(proj model):  153 on 4 and 6166 DF, p-value: <0.00000
00000000002
```

```
#fixed_effects = getfe(model12)

#Hypothesis test for statistical significance
linearHypothesis(model12,c('sc1 = 0','ses1=0','ses1_sc1=0','totexp1=
0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## sc1 = 0
## ses1 = 0
## ses1_sc1 = 0
## totexp1 = 0
##
## Model 1: restricted model
## Model 2: score ~ sc1 + ses1 + ses1_sc1 + totexp1 | schid1n
##
##   Res.Df Df Chisq          Pr(>Chisq)
## 1   6170
## 2   6166  4   613 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Comparison between models
modelsummary(models = list(model11, model12))
```

	Model 1	Model 2
sc1	12.804	8.157
	(1.128)	(3.463)
ses1		24.256
		(1.316)
ses1_sc1		2.434
		(2.172)
totexp1		0.120
		(0.063)
Num.Obs.	6414	6246
R2	0.241	0.297
R2 Adj.	0.232	0.288
AIC	65628.4	63440.4
BIC	66156.1	63986.3
RMSE	39.84	38.34

```r
linearHypothesis(model12, c('ses1_sc1=0'))
```

```
## Linear hypothesis test
##
## Hypothesis:
## ses1_sc1 = 0
##
## Model 1: restricted model
## Model 2: score ~ sc1 + ses1 + ses1_sc1 + totexp1 | schid1n
##
##   Res.Df Df Chisq Pr(>Chisq)
```

```
## 1   6167
## 2   6166  1  1.26       0.26
```

```
#Part C


# Question 13 -----------------------------------------------------
-------
model13 = lm (cs ~ sc1, data )
summary(model13)
```

```
##
## Call:
## lm(formula = cs ~ sc1, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.743 -1.676  0.257  1.257  4.324
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  22.7433     0.0283     802 <0.0000000000000002 ***
## sc1          -7.0671     0.0524    -135 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.91 on 6412 degrees of freedom
## Multiple R-squared:  0.739,  Adjusted R-squared:  0.739
## F-statistic: 1.82e+04 on 1 and 6412 DF,  p-value: <0.000000000000
0002
```

```
model14 = felm (cs ~ sc1 | schid1n, data )
summary(model14)
```

```
##
## Call:
##    felm(formula = cs ~ sc1 | schid1n, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.997 -0.703  0.000  0.810  3.153
##
## Coefficients:
##     Estimate Std. Error t value            Pr(>|t|)
## sc1  -7.1140     0.0338    -211 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 6337 degrees of freedom
## Multiple R-squared(full model): 0.898   Adjusted R-squared: 0.897
```

```
## Multiple R-squared(proj model): 0.875   Adjusted R-squared: 0.874
## F-statistic(full model): 737 on 76 and 6337 DF, p-value: <0.00000
00000000002
## F-statistic(proj model): 4.44e+04 on 1 and 6337 DF, p-value: <0.0
000000000000002
```

```
cov(data$cs,data$sc1)
```

```
## [1] -1.4622
```

```
# Question 14 ----------------------------------------------------
-------

#IV
model15 = felm (score ~ sc1 | schid1n, data )  #y~z
summary(model15)
```

```
##
## Call:
##    felm(formula = score ~ sc1 | schid1n, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -143.95  -27.15   -2.06   25.34  137.06
##
## Coefficients:
##     Estimate Std. Error t value            Pr(>|t|)
## sc1    12.80       1.13    11.3 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.1 on 6337 degrees of freedom
## Multiple R-squared(full model): 0.241   Adjusted R-squared: 0.232
## Multiple R-squared(proj model): 0.0199   Adjusted R-squared: 0.00
816
## F-statistic(full model):26.5 on 76 and 6337 DF, p-value: <0.00000
00000000002
## F-statistic(proj model):  129 on 1 and 6337 DF, p-value: <0.00000
00000000002
```

```
model16 = felm (cs ~ sc1 | schid1n, data ) #x~z
summary(model16)
```

```
##
## Call:
##    felm(formula = cs ~ sc1 | schid1n, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.997 -0.703  0.000  0.810  3.153
##
```

```
## Coefficients:
##     Estimate Std. Error t value            Pr(>|t|)
## sc1  -7.1140     0.0338    -211 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.2 on 6337 degrees of freedom
## Multiple R-squared(full model): 0.898   Adjusted R-squared: 0.897
## Multiple R-squared(proj model): 0.875   Adjusted R-squared: 0.874
## F-statistic(full model): 737 on 76 and 6337 DF, p-value: <0.00000
00000000002
## F-statistic(proj model): 4.44e+04 on 1 and 6337 DF, p-value: <0.0
000000000000002
```

```
b1_yz = as.numeric(model15$coefficients[1])
b1_xz = as.numeric(model16$coefficients[1])

b1IV = b1_yz/b1_xz
paste("beta IV = ",b1IV)
```

```
## [1] "beta IV =  -1.79980513924389"
```

```
# Question 15 ----------------------------------------------------
-------

model17 = felm(score ~ sc1 + black + white + ses1 + boy  | schid1n,
data)
summary(model17)
```

```
##
## Call:
##    felm(formula = score ~ sc1 + black + white + ses1 + boy | schi
d1n,      data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -159.37  -25.95   -1.67   24.71  124.59
##
## Coefficients:
##       Estimate Std. Error t value             Pr(>|t|)
## sc1     11.871      1.091   10.88 < 0.0000000000000002 ***
## black  -21.346      6.162   -3.46              0.00054 ***
## white   -6.213      6.010   -1.03              0.30125
## ses1    22.736      1.170   19.43 < 0.0000000000000002 ***
## boy     -5.862      0.974   -6.02         0.0000000018 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.3 on 6175 degrees of freedom
##   (158 observations deleted due to missingness)
```

```
## Multiple R-squared(full model): 0.307   Adjusted R-squared: 0.298
## Multiple R-squared(proj model): 0.104   Adjusted R-squared: 0.092
1
## F-statistic(full model):34.2 on 80 and 6175 DF, p-value: <0.00000
00000000002
## F-statistic(proj model):  143 on 5 and 6175 DF, p-value: <0.00000
00000000002
```

```
# Question 16 ----------------------------------------------------
-------

#tsls

model18 = felm(score ~ cs + black + white+ ses1 + boy | schid1n, dat
a) #base
summary(model18)
```

```
##
## Call:
##    felm(formula = score ~ cs + black + white + ses1 + boy | schid
1n,      data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -155.44  -25.96   -1.92   24.69  126.83
##
## Coefficients:
##       Estimate Std. Error t value             Pr(>|t|)
## cs      -1.494      0.144  -10.37 < 0.0000000000000002 ***
## black  -21.140      6.167   -3.43              0.00061 ***
## white   -5.914      6.015   -0.98              0.32551
## ses1    22.741      1.171   19.42 < 0.0000000000000002 ***
## boy     -5.902      0.974   -6.06         0.0000000015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.3 on 6175 degrees of freedom
##   (158 observations deleted due to missingness)
## Multiple R-squared(full model): 0.306   Adjusted R-squared: 0.297
## Multiple R-squared(proj model): 0.102   Adjusted R-squared: 0.090
6
## F-statistic(full model):  34 on 80 and 6175 DF, p-value: <0.00000
00000000002
## F-statistic(proj model):  141 on 5 and 6175 DF, p-value: <0.00000
00000000002
```

```
model19 = felm(cs ~ sc1 + black + white + ses1 + boy |schid1n, data)
#first stage
summary(model19)
```

```
##
## Call:
##    felm(formula = cs ~ sc1 + black + white + ses1 + boy | schid1n
,      data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.001 -0.694  0.005  0.807  3.170
##
## Coefficients:
##       Estimate Std. Error t value            Pr(>|t|)
## sc1    -7.0927     0.0341 -208.29 <0.0000000000000002 ***
## black   0.1634     0.1923    0.85                0.40
## white   0.2162     0.1875    1.15                0.25
## ses1   -0.0238     0.0365   -0.65                0.51
## boy    -0.0226     0.0304   -0.74                0.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.19 on 6175 degrees of freedom
##   (158 observations deleted due to missingness)
## Multiple R-squared(full model): 0.899   Adjusted R-squared: 0.898
## Multiple R-squared(proj model): 0.876   Adjusted R-squared: 0.874
## F-statistic(full model): 689 on 80 and 6175 DF, p-value: <0.00000
00000000002
## F-statistic(proj model): 8.69e+03 on 5 and 6175 DF, p-value: <0.0
000000000000002
```

```
data$cs_hat = fitted(model19)#get the cs hat values

model20 = felm(score ~ cs_hat + black + white + ses1 + boy |schid1n
, data) #second stage
summary(model20)
```

```
##
## Call:
##    felm(formula = score ~ cs_hat + black + white + ses1 + boy |
schid1n, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -159.37  -25.95   -1.67   24.71  124.59
##
## Coefficients:
##        Estimate Std. Error t value             Pr(>|t|)
## cs_hat   -1.674      0.154  -10.88 < 0.0000000000000002 ***
## black   -21.072      6.162   -3.42              0.00063 ***
## white    -5.851      6.010   -0.97              0.33029
## ses1     22.696      1.170   19.40 < 0.0000000000000002 ***
## boy      -5.900      0.974   -6.06         0.0000000014 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.3 on 6175 degrees of freedom
##   (158 observations deleted due to missingness)
## Multiple R-squared(full model): 0.307   Adjusted R-squared: 0.298
## Multiple R-squared(proj model): 0.104   Adjusted R-squared: 0.092
1
## F-statistic(full model):34.2 on 80 and 6175 DF, p-value: <0.00000
00000000002
## F-statistic(proj model):  143 on 5 and 6175 DF, p-value: <0.00000
00000000002
```