Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 9 Exercises

Alan S. Gerber and Donald P. Green*

Question 1

Important concepts:

a) Define CATE. Is a Complier average causal effect (CACE) an example of a CATE?
Answer:
CATE stands for conditional average treatment effect, or the ATE among a subgroup. Typically, the subgroup in question is defined by some observable covariate(s), such as the CATE for women over 40 years of age. One could, however, define a CATE for a latent group such as Compliers (those who take the treatment if and only if assigned to the treatment group). Therefore, a CACE is a CATE.

b) What is an interaction effect?
Answer:
An interaction refers to systematic variation in treatment effects. A treatment-by-covariate interaction refers to variation in ATEs that is a function of covariates. A treatment-by-treatment interaction refers to variation in the average effect of one randomized intervention that occurs as a function of other assigned treatments.

c) Describe the multiple comparisons problem and the Bonferroni correction.
Answer:
The multiple comparisons problem refers to the disortion in p-values that occurs when researchers conduct a series of hypothesis tests. When several hypothesis tests are conducted, the chances that at least one of them appears significant may be substantially greater than 0.05, the nominal size of each test. The Bonferroni correction reestablishes the proper size of each test when several hypothesis tests are conducted. If k tests are conducted at the 0.05 level, the Bonferroni-corrected target significance level is $0.05/k$.

Question 2

The standard error formula given in equation (3.4) suggests that, all else being equal, reducing variance in $Y_i(0)$ helps reduce sampling uncertainty. Referring to the procedure outlined in section 9.2, explain why the same principle applies to estimating bounds on treatment effect heterogeneity.
Answer:
Nonparametric tests of heterogeneity put an estimated lower bound on the variance of the subject-level treatment effect by sorting the observed $Y_i(0)$ and $Y_i(1)$ in ascending order and calculating the

*Solutions prepared by Peter M. Aronow and revised by Alexander Coppock

difference between them. The variance of this difference is the estimated lower bound. When the variance of $Y_i(0)$ is small, the variance of the differences between $Y_i(0)$ and $Y_i(1)$ is scarcely affected by whether $Y_i(0)$ is sorted in ascending or descending order. In the limiting case where the variance of $Y_i(0)$ is zero, sorting makes no difference at all; if one were sure that $Y_i(0)$ were constant, one could estimate unit-level treatment effects by subtracting the mean of $Y_i(0)$ from $Y_i(1)$.

Question 3

One way to reduce variance in $Y_i(0)$ is to block on a prognostic covariate. When blocking is used, the joint distribution of $Y_i(0)$ and $Y_i(1)$ is simulated within blocks using the bounding procedure described in section 9.2. Using the schedule of potential outcomes below, show how the maximum and minimum values of the covariance of $Y_i(0)$ and $Y_i(1)$ compare to the maximum and minimum values of the covariance of $Y_i(0)$ and $Y_i(1)$ for the dataset as a whole (i.e., had blocking not been used).

Table 1: Question 3 Table

Block	Subject	Yi(0)	Yi(1)
A	A-1	0	2
A	A-2	1	5
A	A-3	1	3
A	A-4	2	1
B	B-1	2	3
B	B-2	3	3
B	B-3	4	9
B	B-4	4	7


```
In [1]: clear
        set obs 8
        egen block = repeat(), values("A")
        replace block ="B" in 5/8

In [3]: input int y0 int y1
                0 2
                1 5
                1 3
                2 1
                2 3
                3 3
                4 9
                4 7 end

In [6]: // function to calculate population covariance
        cap program drop cov_pop
        program define cov_pop, rclass
        args x y
```

```
        tempvar xy_dev
        qui sum `x'
        local avg_x = r(mean)
        local length = r(N)

        qui sum `y'
        local avg_y = r(mean)

        gen `xy_dev' = (`x'-`avg_x')*(`y'-`avg_y')
        qui tabstat `xy_dev', stat(sum) save
        return scalar cor_pop = el(r(StatTotal),1,1)/`length'
        end

        qui egen rank_y1=rank(y1), unique
        qui gen id=_n

In [7]: vlookup id, generate(y1_lowtohigh) key(rank_y1) value(y1)
        replace id = 9-id
        vlookup id, generate(y1_hightolow) key(rank_y1) value(y1)

In [8]: cov_pop y0 y1_hightolow
        di "cov.min ="%8.3f r(cor_pop)



cov.min =  -3.141


In [9]: cov_pop y0 y1_lowtohigh
        di "cov.min ="%8.3f r(cor_pop)



cov.min =   3.234


In [10]: qui replace id=_n
         qui replace id=. if block=="B"
         qui egen rank_y1_A = rank(y1), unique by(block)
         qui replace rank_y1_A =. if block=="B"
         qui vlookup id, generate(y1_hightolow_block_A) key(rank_y1_A) value(y1)
         qui replace id = _n-4
         qui replace id =. if block=="A"
         qui egen rank_y1_B = rank(y1), unique by(block)
         qui replace rank_y1_B =. if block=="A"
         qui vlookup id, generate(y1_hightolow_block_B) key(rank_y1_B) value(y1)
         qui gen y1_hightolow_block = y1_hightolow_block_A
         qui replace y1_hightolow_block = y1_hightolow_block_B if block=="B"
```

```
        qui replace id=5-_n
        qui replace id=. if block=="B"
        qui vlookup id, generate(y1_lowtohigh_block_A) key(rank_y1_A) value(y1)
        qui replace id = 9-_n
        qui replace id =. if block=="A"
        qui vlookup id, generate(y1_lowtohigh_block_B) key(rank_y1_B) value(y1)
        qui gen y1_lowtohigh_block = y1_lowtohigh_block_A
        qui replace y1_lowtohigh_block = y1_lowtohigh_block_B if block=="B"

In [11]: cov_pop y0 y1_lowtohigh_block
         di "cov.min ="%8.4f r(cor_pop)



cov.min = -0.0156


In [12]: cov_pop y0 y1_hightolow_block
         di "cov.min ="%8.3f r(cor_pop)



cov.min =   2.984
```

The lowest and highest covariances under simple random assignment are -3.14 and 3.23. In order to find the lowest and highest covariances under blocked assignment, sort the potential outcomes within blocks before calculating the covariances for all observations. Under blocked random assignment, the lowest covariance is -0.02, and the highest covariance is 2.98. Taking advantange of the blocks reduces the range of possible covariances.

Question 4

Suppose that a researcher compares the CATE among two subgroups, men and women. Among men (N = 100), the ATE is estimated to be 8.0 with a standard error of 3.0, which is significant at $p < 0.05$. Among women (N = 25), the CATE is estimated to be 7.0 with an estimated standard error of 6.0, which is not significant, even at the 10% significance level. Critically evaluate the researcher's claim that "the treatment only works for men; for women, the effect is statistically indistinguishable from zero." In formulating your answer, address the distinction between testing whether a single CATE is different from zero and testing whether two CATEs are different from each other.

Answer:

The researcher's interpretation of the results ignores the fact that the estimated CATE for women is almost as large (7 versus 8) as the estimated CATE among men. The difference between the estimated CATEs (8-7=1) is much smaller than the apparent standard error of the difference (which is the square root of the sum of the estimated standard errors, or 6.7). An alternative interpretation is that both of the CATEs are the same, but the CATE among men is estimated with greater precision because the male sample is much larger than the female sample.

Question 5

The table below shows hypothetical potential outcomes for an experiment in which low-income subjects in a developing country are randomly assigned to receive (i) loans to aid their small businesses; (ii) business training to improve their accounting, hiring, and inventory-management skills; (iii) both; or (iv) neither. The outcome measure in business income during the subsequent year. The table also includes a pre-treatment covariate, an indicator scored 1 if the subject was judged to be proficient in these basic business skills.

Table 2: Question 5 Table

Subject	$Y_i(loan)$	$Y_i(training)$	$Y_i(both)$	$Y_i(Neither)$	Prior business skills
1	2	2	3	2	0
2	2	3	2	1	0
3	5	6	6	4	1
4	3	1	5	1	1
5	4	4	5	0	0
6	10	8	11	10	1
7	1	3	3	1	0
8	5	5	5	5	1
Average	4	4	5	3	0.5

a) What is the ATE of the loan if all subjects were also to receive training?
Answer:
The relevant comparison is the average potential outcomes under "both" to the average potential outcome under only "training." The ATE is 5-4=1.

b) What is the ATE of the loan if no subjects receive training?
Answer:
The relevant comparison is the average potential outcomes under "loan" to the average potential outcome under only "neither." The ATE is 4-3=1.

c) What is the ATE of the training if all subjects also receive a loan?
Answer:
The relevant comparison is the average potential outcomes under "both" to the average potential outcome under only "loan." The ATE is 5-4=1.

d) What is the ATE of the training if no subjects receive a loan?
Answer:
The relevant comparison is the average potential outcomes under "training" to the average potential outcome under only "neither." The ATE is 4-3=1.

e) Suppose subjects were randomly assigned to one of the four experimental treatments in equal proportions. Use the table above to fill in the expected values of the four regression coefficients for the model and interpret the results:

$$Y_i = \alpha_0 + \alpha_1 Loan_i + \alpha_2 Training_i + \alpha_3 (Loan_i * Training_i) + e_i \quad (1)$$

The four coefficients are $\alpha_0 = 3$, the average outcome under "neither"; $\alpha_1 = 1$, the ATE of loan when there is no training; $\alpha_2 = 1$, the ATE of training when there is no loan; and $\alpha_3 = 0$ the change in the effect of training that occurs when our focus switches from those who receive no loan to those who receive a loan. Note that this interaction term can also be interpreted as the change in the ATE of loans that we observe when we move from the untrained subgroup to the trained subgroup.

$$Y_i = 3 + 1 * Loan_i + 1 * Training_i + 0 * (Loan_i * Training_i) + e_i \tag{2}$$

f) Suppose a researcher were to implement a block randomized experiment, such that two subjects with business skills are assigned to receive loans, and two subjects without business skills are assigned to receive loans, and the rest are assigned to control. No subjects are assigned to receive training. The researcher estimates the model

$$Y_i = \gamma_0 + \gamma_1 Loan_i + \gamma_2 Skills_i + \gamma_3 (Loan_i * Skills_i) + e_i \tag{3}$$

Over all 36 possible random assignments, the average estimated regression is as follows:

$$Y_i = 1.00 + 1.25 Loan_i + 4.00 Skills_i - 0.50 (Loan_i * Skills_i) \tag{4}$$

Interpret the results and contrast them with the results from part (e). (Hint: the block randomized design does not affect the interpretation. Focus on the distinction between treatment-by-treatment and treatment-by-covariate interactions.)
Answer:
The key thing to bear in mind when interpreting these results is that the interaction between loans and skills is a treatment-by-covariate interaction because skills are not randomly assigned. The results seem to suggest that loans are more effective amongst those without skills (CATE = 1.25) than among those with skills (CATE = 1.25 - 0.5 = 0.75). These CATEs may describe the ATEs in these two skill groups, but the change in CATEs does not necessarily imply that a random increase in skill would diminish the effects of loans.

Question 6

Rind and Bordia studied the tipping behavior of lunchtime patrons of an "upscale Philadelphia restaurant" who were randomly assigned to four experimental groups.[1] One factor was server sex (male or female), and a second factor was whether the server draws a "happy face" on the back of the bill presented to customers.[2] Download the data located at http://isps.research.yale.edu/FEDAI.

a) Suppose you ignored the sex of the server and simply analyzed whether the happy face treatment has heterogeneous effects. Use randomization inference to test whether $Var(\tau_i) = 0$ by testing whether $Var(Y_i(1)) = Var(Y_i(0))$. Construct the full schedule of potential outcomes by assuming that the treatment effect is equal to the observed difference-in-means between $Y_i(1)$

[1] Rind and Bordia 1996.

[2] The authors took steps to ensure the blindness of the servers to the happy face condition, which was determined only moments before the bill was delivered. The authors also instructed waitstaff to deliver bills and walk away, so that there would be no additional interaction with customers. It is not clear whether the sex of the server was randomly assigned.

and $Y_i(0)$. Interpret your results.

```
In [1]: qui import delim ./data/chapter09/Rind_Bordia_JASP_1996, clear

In [2]: gen Z =.
        qui replace Z = 1 if happyface==1
        qui replace Z = 0 if happyface==0

        rename tip Y

        capture program drop var_difference
        program define var_difference, rclass
                sum Y if Z==1, detail
                local var_treat = r(Var)
                sum Y if Z==0, detail
                local var_control = r(Var)
                return scalar vardiff= `var_treat'-`var_control'
        end

        tsrtest Z r(vardiff): var_difference

Two-sample randomization test for theta=r(vardiff) of var_difference by Z

Combinations:   5.19137106438e+25 = (89 choose 44)
Assuming null=0
Observed theta: 53.31

Minimum time needed for exact test (h:m:s):  1.66e+18:00:00
Reverting to Monte Carlo simulation.
Mode: simulation (10000 repetitions)

progress: |...|

 p=0.23448 [one-tailed test of Ho:  theta(Z==0)<=theta(Z==1)]
 p=0.76542 [one-tailed test of Ho:  theta(Z==0)>=theta(Z==1)]
 p=0.47205 [two-tailed test of Ho:  theta(Z==0)==theta(Z==1)]


In [3]: //  p-value for var(Y1)>Var(Y0)
        di %8.4f r(uppertail)

  0.2345


In [4]: //  p-value for var(Y1)<>Var(Y0)
        di %8.3f r(twotail)

   0.472
```

We constructed a simulation of 10,000 random assignments and for each assessed the difference in variances between treatment and control group. The observed difference is 53.31. However, this absolute difference has a p-value of 0.472. We cannot reject the null hypothesis that the observed difference in variances is the produce of random samping variability. The failure to reject the null is not surprising given the low power of this test, which does not focus on any specific model of heterogeneous treatment effects.

b) Write down a regression model that depicts the effect of the sex of the waitstaff, whether they write a happy face on the bill, and the interaction of these factors.
Answer:
Using tip percentage as the outcome and a binary variable for sex (female=1) and for the use of a happy face (face=1), a regression model is as follows:

$$Y_i = \gamma_0 + \gamma_1 Sex_i + \gamma_2 Face_i + \gamma_3 (Sex_i * Face_i) + e_i \tag{5}$$

c) Estimate the regression model in (b) and test the interaction between waitstaff sex and the happy face treatment. Is the interaction significant?
Answer:


```
In [5]: rename female female_factor
        gen female = .
        replace female = 1 if female_factor==1
        replace female = 0 if female_factor==0

        gen zfemale = Z*female

In [6]: //lmmodelint: regression with interaction between
        //happyface and waitstaff sex
        regress Y Z female zfemale


      Source |       SS           df       MS      Number of obs   =        89
-------------+----------------------------------   F(3, 85)        =      9.32
       Model |  3072.39611         3  1024.13204   Prob > F        =    0.0000
    Residual |  9335.52582        85  109.829716   R-squared       =    0.2476
-------------+----------------------------------   Adj R-squared   =    0.2211
       Total |  12407.9219        88  140.999113   Root MSE        =     10.48

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Z |  -3.629627   3.163098    -1.15   0.254    -9.918714     2.65946
      female |   6.378199   3.163098     2.02   0.047      .089112    12.66729
     zfemale |   8.887078   4.446646     2.00   0.049     .0459551     17.7282
       _cons |   21.40571   2.286916     9.36   0.000     16.85871    25.95272
------------------------------------------------------------------------------
```

```
In [7]: //lmmodel: regression model without interaction
        regress Y Z female


      Source |       SS           df       MS      Number of obs   =        89
-------------+----------------------------------   F(2, 86)        =     11.59
       Model |  2633.69091         2  1316.84545   Prob > F        =    0.0000
    Residual |  9774.23103        86  113.653849   R-squared       =    0.2123
-------------+----------------------------------   Adj R-squared   =    0.1939
       Total |  12407.9219        88  140.999113   Root MSE        =    10.661

------------------------------------------------------------------------------
           Y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           Z |   .8673363   2.261535     0.38   0.702    -3.628446    5.363119
      female |   10.87516   2.261535     4.81   0.000      6.37938    15.37094
       _cons |   19.05503   1.995134     9.55   0.000     15.08883    23.02122
------------------------------------------------------------------------------


In [8]: scalar coeff_z = _b[Z]
        cap drop Y0 Y1
        gen Y0 = Y - coeff_z * Z
        gen Y1 = Y + coeff_z*(1- Z)


        qui regress Y Z female zfemale
        qui test zfemale
        global f_obs = r(F)

In [9]: capture program drop wald_f
        program define wald_f, rclass
                tempvar Y_sim zsimfemale
                gen `Y_sim' = Y1 * Z + Y0 * (1 - Z)
                gen  `zsimfemale' = female*Z
                qui reg `Y_sim' Z female `zsimfemale'
                test `zsimfemale'
                return scalar f_sims = r(F)
        end

In [10]: tsrtest Z r(f_sims) using 9_6_fsims.dta, overwrite: wald_f

Two-sample randomization test for theta=r(f_sims) of wald_f by Z

Combinations:   5.19137106438e+25 = (89 choose 44)
Assuming null=0
Observed theta: 3.994
```

```
Minimum time needed for exact test (h:m:s):  6.84e+19:00:00
Reverting to Monte Carlo simulation.
Mode: simulation (10000 repetitions)

progress: |...|

 p=0.04740 [one-tailed test of Ho:  theta(Z==0)<=theta(Z==1)]
 p=0.95250 [one-tailed test of Ho:  theta(Z==0)>=theta(Z==1)]
 p=0.04740 [two-tailed test of Ho:  theta(Z==0)==theta(Z==1)]

Saving log file to 9_6_fsims.dta...done.


In [11]: di %8.4f r(uppertail)

  0.0474
```

The regression reported above suggests a positive interaction between the happyface treatment and female, implying that female waitstaff receive much more return from happyfaces than male waitstaff. The two-sided p-value from the regression is 0.049, which is similar to the result from randomization inference ($p = 0.0474$). A two-sided test is appropriate here because the direction of the effect was not predicted ex ante. Thinking back to section (a), the specific interaction posited by this regression sets the stage for a more powerful test of treatment effect heterogeneity.

Question 7

In their 2004 study of racial discrimination in employment markets, Bertrand and Mullainathan sent resumes with varying characteristics to firms advertising job openings. Some firms were sent resumes with putative African American names, while other firms received resumes with putatively Caucasian names. The researchers also varied other attributes of the resume, such as whether the resume was judged to be of high or low quality (based on labor market experience, career profile, gaps in employment, and skills listed).[3] The table below shows the rate at which applicants were called back by employers, by the city in which the experiment took place and by the randomly assigned attributes of their applications.

Table 3: Question 7 Table

	Boston				Chicago			
	Low-quality resume		High-quality resume		Low-quality resume		High-quality resume	
	Black	White	Black	White	Black	White	Black	White
% Received Call	7.01	10.15	8.5	13.12	5.52	7.16	5.28	8.94
(N)	(542)	(542)	(541)	(541)	(670)	(670)	(682)	(682)

[3] Bertrand and Mullainathan 2004, p. 994.

a) For each city, interpret the apparent treatment effects of race and resume quality on the probability of receiving a follow-up call.
Answer:
For Boston, the effect of (white) race is 10.15 - 7.01 = 3.14 when resume quality is low and 13.12 - 8.50 = 4.62 when resume quality is high. For Chicago, the effect of (white) race is 7.16 - 5.52 = 1.64 when resume quality is low and 8.94 - 5.28 = 3.66 when resume quality is high. Note that another, equally valid way to interpret the table is to assess the effect of resume quality for each race, but the substantive focus of this study is on race effects.

b) Propose a regression model that assesses the effects of the treatments, interaction between them, and interactions between the treatments and the covariate, city.
Answer:
This model is similar to the interactive regression specifications described above, but it contains treatment-by-treatment interactions (race x resume) and treatment-by-covariate interactions (race x city, resume x city) and a higher order interaction (race x resume x city) that allows for the possibility that the race x resume interaction differs by city. Here, City is scored 1 if Chicago. Race = 1 if white. Resume =1 if high quality. Notice that the "saturated" regression model contains eight parameters, one for each cell of the table.

$$Y_i = \gamma_0 + \gamma_1 Race_i + \gamma_2 Resume_i + \gamma_3 City_i +$$
$$\gamma_4(Race_i * Resume_i) + \gamma_5(Race_i * City_i) + \gamma_6(Resume_i * City_i) + \gamma_7(Race_i * Resume_i * City_i) + e_i$$

c) Estimate the parameters in your regression model. Interpret the results (This can be done by hand based on the percentages given in the table.)
Answer:
Because there as many parameters as experimental groups, the estimated coefficients reproduce the percentages given in the table:

$$Y_i = 7.01 + 3.14 Race_i + 1.49 Resume_i - 1.49 City_i +$$
$$1.48(Race_i * Resume_i) - 1.50(Race_i * City_i) - 1.73(Resume_i * City_i) + 0.54(Race_i * Resume_i * City_i) + e_i$$

Additional response (Boston = 1; Black = 1; Low quality = 1)

$$Y_i = 8.94 - 3.66 Race_i - 1.78 Resume_i + 4.18 City_i +$$
$$2.02(Race_i * Resume_i) - 0.96(Race_i * City_i) - 1.19(Resume_i * City_i) - 0.54(Race_i * Resume_i * City_i) + e_i$$

Additional response (Boston = 1; Black = 1; High quality = 1)

$$Y_i = 7.16 - 1.64 Race_i + 1.78 Resume_i + 2.99 City_i -$$
$$2.02(Race_i * Resume_i) - 1.50(Race_i * City_i) + 1.19(Resume_i * City_i) + 0.54(Race_i * Resume_i * City_i) + e_i$$

```
In [1]: clear
        qui set obs 4870
        qui egen y = fill(1,1)
```

```
        qui replace y = 0 in 39/542
        qui replace y = 0 in 598/1084
        qui replace y = 0 in 1131/1625
        qui replace y = 0 in 1697/2166
        qui replace y=0 in 2204/2836
        qui replace y=0 in 2885/3506
        qui replace y=0 in 3543/4188
        qui replace y=0 in 4250/4870

        qui egen boston = fill(1,1)
        qui replace boston = 0 in 2167/4870
        qui gen chicago = 1-boston
        qui egen lowquality = fill(1,1)
        qui replace lowquality = 0 in 1085/2166
        qui replace lowquality = 0 in 3507/4870
        qui gen highquality = 1-lowquality
        qui egen black = fill(1,1)
        qui replace black = 0 in 543/1084
        qui replace black = 0 in 1626/2166
        qui replace black = 0 in 2837/3506
        qui replace black = 0 in 4188/4870
        qui replace black = 0 in 4189/4870
        qui gen white = 1-black

        qui gen whitehighquality = white*highquality
        qui gen whitechicago = white*chicago
        qui gen highqualitychicago = highquality*chicago
        qui gen whitehighqualitychicago = white * highquality * chicago
```

In [2]:
```
        // fit_1
        qui regress y white highquality chicago whitehighquality ///
        whitechicago highqualitychicago whitehighqualitychicago
        estimates store m1, title(Model 1)

        gen blackhighquality = black*highquality
        gen blackchicago = black*chicago
        gen blackhighqualitychicago = black * highquality * chicago
```

In [3]:
```
        // fit_2
        qui regress y black highquality chicago blackhighquality ///
        blackchicago highqualitychicago blackhighqualitychicago
        estimates store m2, title(Model 2)
```

In [4]:
```
        // fit_3
        gen whiteboston = white*boston
        gen highqualityboston = highquality*boston
        gen whitehighqualityboston = white * highquality * boston
        qui regress y white highquality boston whitehighquality ///
```

```
        whiteboston highqualityboston whitehighqualityboston

        estimates store m3, title(Model 3)
```

In [5]:
```
// fit_4
gen blacklowquality = black*lowquality
gen lowqualitychicago = lowquality*chicago
gen blacklowqualitychicago = black * lowquality * chicago
qui regress y black lowquality chicago blacklowquality ///
blackchicago lowqualitychicago blacklowqualitychicago

estimates store m4, title(Model 4)
```

In [6]:
```
estout m1 m2 m3 m4, cells(b(star fmt(3)) se(par fmt(3))) ///
starlevels( * 0.10 ** 0.05 *** 0.010) ///
legend label varlabels(_cons constant) ///
stats(N r2 r2_a rmse F, fmt(0 3 3 3 3) ///
label(N R-squared Ajusted-R2 Residual_Std_Error F-Statistic))
```

```
------------------------------------------------------------------------------------
                      Model 1         Model 2         Model 3         Model 4
                         b/se            b/se            b/se            b/se
------------------------------------------------------------------------------------
white                   0.031*                          0.016
                       (0.016)                         (0.015)
highquality              0.015          0.030*          -0.002
                       (0.016)         (0.016)         (0.015)
chicago                 -0.015          -0.030*                         -0.042***
                       (0.016)         (0.016)                         (0.016)
whitehighquality         0.015                          0.020
                       (0.023)                         (0.021)
whitechicago            -0.015
                       (0.022)
highqualitychicago      -0.017          -0.012
                       (0.022)         (0.022)
whitehighqualitych~o     0.005
                       (0.031)
black                                   -0.031*                         -0.046***
                                       (0.016)                         (0.016)
blackhighquality                        -0.015
                                       (0.023)
blackchicago                             0.015                           0.010
                                       (0.022)                         (0.022)
blackhighqualitych~o                    -0.005
                                       (0.031)
boston                                                  0.015
                                                       (0.016)
whiteboston                                             0.015
                                                       (0.022)
highqualityboston                                       0.017
                                                       (0.022)
whitehighqualitybo~n                                   -0.005
                                                       (0.031)
lowquality                                                              -0.030*
```

```
                                                                          (0.016)
blacklowquality                                                            0.015
                                                                          (0.023)
lowqualitychicago                                                          0.012
                                                                          (0.022)
blacklowqualitychi~o                                                       0.005
                                                                          (0.031)
constant                    0.070***       0.101***       0.055***       0.131***
                           (0.012)        (0.012)        (0.010)        (0.012)
-------------------------------------------------------------------------------------
N                           4870           4870           4870           4870
R-squared                   0.008          0.008          0.008          0.008
Ajusted-R2                  0.006          0.006          0.006          0.006
Residual_Std_Error          0.271          0.271          0.271          0.271
F-Statistic                 5.349          5.349          5.349          5.349
-------------------------------------------------------------------------------------
* p<0.10, ** p<0.05, *** p<0.010
```

Question 8

In Chapter 3, we analyzed data from Clingingsmith, Khwaja, and Kremer's study of Pakistani Muslims who participated in a lottery to obtain a visa for the pilgrimage to Mecca.[4] By comparing lottery winners to lottery losers, the authors are able to estimate the effects of the pilgrimage on various attitudes, including views about people from other countries. Winners and losers were asked to rate the Saudi, Indonesian, Turkish, African, European, and Chinese people on a five-point scale ranging from very negative (-2) to very positive ($+2$). Adding the responses to all six items creates an index ranging from -12 to $+12$. The key results are presented in the table below.

Table 4: Question 8 table

	Control group	Treatment group
N	448	510
Mean	1.868	2.343
Variance	5.793	6.902
Absolute difference in variances	1.109	

a) Explain the meaning of "absolute difference in variances."
Answer:
The term "difference in variances" is the observed difference between the variance of outcomes in the treatment group and the variance of outcomes in the control group. The term "absolute" refers to the absolute value of this difference.

b) Describe how one could use randomization inference to test the null hypothesis of constant treatment effects.
Answer:
One method is to create a full schedule of potential outcomes under the null hypothesis of

[4] Clingingsmith, Khwaja, and Kremer 2009.

constant treatment effects. For example, we could assume that all subjects have a treatment effect equal to the observed ATE. In order to obtain untreated potential outcomes for the treatment group, we subtract off the ATE from the observed treated potential outcomes. In order to obtain the treated potential outcomes for the control group, we add the apparent ATE. We then simulate a large number of possible random assignments; for each random assignment, we calculate the absolute difference between the variance of the treatment group and the variance of the control group. We obtain p-values by determining where the observed absolute difference falls in the sampling distribution under the null hypothesis.

c) Assume that researchers applied the method you proposed in part (b) and simulated 100,000 random assignments, each time calculating the absolute difference in variances; they find that 25,220 of these differences are as large or larger than 1.109, the absolute difference in variances observed in the original sample. Calculate the p-value implied by these results. What do you conclude about treatment effect heterogeneity in this example?
Answer:
The p-value is $25220/100000 = 0.2522$. The difference in observed variances is consistent with (and thus we cannot reject) the null hypothesis of homogeneous effects.

d) Suppose that this experiment were partitioned into subgroups defined according to whether the subjects had travelled abroad in the past. Suppose that the CATE among those who had previously travelled abroad were 0 and that the CATE among those who had not travelled abroad were 1.0. Suppose this difference in CATEs were significant at $p < .05$. Does this result imply that randomly encouraging people to travel abroad eliminates the Hajj's effect?
Answer:
Not necessarily. This regression reports a treatment-by-covariate interaction, which describes the CATEs for two subgroups that may or may not have similar potential outcomes. Randomly encouraging travel is designed to create groups with the same expected potential outcomes; this design tests whether travel causes the effect of the Hajj to change.

Question 9

An example of a two-factor design that encounters one-sided noncompliance may be found in Fieldhouse et al.'s study of voter mobilization in the United Kingdom.[5] In this study, the first factor is whether each voter was mailed a letter encouraging him or her to vote in the upcoming election. The second factor is whether each voter was called with an encouragement to vote. Noncompliance occurs in the case of phone calls, as some targeted voters cannot be reached when called. The experimental design consists of four groups: a control group, a mail-only group, a phone-only group, and a group targeted for both mail and phone. The following table shows the results by assigned experimental group.

a) Show that, under certain assumptions, this experimental design allows one to identify the following parameters: (i) the ATE of mail, (ii) the Complier average causal effect (CACE) of phone calls, (iii) the CATE of mail among those who comply with the phone call treatment, (iv) the CATE of mail among those who do not comply with the phone call treatment, and (v) the CACE of phone calls among those who receive mail.
Answer:

[5] Fieldhouse et al. 2010.

Table 5: Question 9 Table

	Control	Mail Only	Phone Only	Mail and Phone
N	5179	4367	3466	2287
Number Contacted by Phone	0	0	2003	1363
Among those Assigned to this Experimental Group, Percent who Voted	0.397	0.403	0.397	0.418
Among those Contacted by Phone, Percent who Voted	NA	NA	0.465	0.468

(i) **ATE of mail.** The ATE of mail is identified using the core assumptions of chapter 2 (random assignment, non-interference, and excludability). Excludability in this holds that the only way that random assignment of mail affects outcomes is through the mail treatment itself.

(ii) **Complier average causal effect (CACE) of phone calls.** In order identify the CACE of phone calls, we must invoke the assumptions of Chapter 5, since this is a case of one-sided non-compliance. Again, the exclusion restriction holds that the only way that the assignment of phone calls affects outcomes is through actual phone contacts. The CACE here is ATE among those who receive phone calls if assigned to the treatment group.

(iii) **CATE of mail among those who comply with the phone call treatment.** The CATE of mail is identified in the same was as an ATE, except that it is restricted to those who actually receive phone calls

(iv) **CATE of mail among those who do not comply with the phone call treatment.** Same as above, but among those who are not treated when called.

(v) **CACE of phone calls among those who receive mail.** The CACE of phone calls among those who receive mail is identified among the same group of compliers as the CACE above, since mail is assigned and received randomly (because we assume full compliance with the mail treatment). However, the ATE of the calls among Compliers may differ from the ATE among Compliers who also receive mail, due to a treatment-by-treatment interaction.

b) Using the identification strategies you laid out in part (a), estimate each of the five parameters using the results in the table.

(i) **ATE of mail.** The estimated ATE of mail is 40.3 - 39.7 = 0.6 percentage points.

(ii) **Complier average causal effect (CACE) of phone calls.** The estimated CACE of phone calls is the ITT divided by the share of compliers: (39.7 - 39.7)/ (2003/3466) = 0.

(iii) **CATE of mail among those who comply with the phone call treatment.** The CATE of mail among those who receive a call is 46.8 - 46.5 = 0.3 percentage points.

(iv) **CATE of mail among those who do not comply with the phone call treatment.** In order to figure out the CATE of mail among those who did not comply when called, we must first back out the voting rates given the numbers presented above. For example, the overall voting rate in the treatment group of 41.8 is a weighted average of the voting rates among the contacted and uncontacted. Thus, 41.8 = 46.8(1363/2287) + X(923/2287).

Solving for X gives 34.4, and repeating the same calculation for the control group gives 30.4. Therefore, the estimated effect of mail for this subgroup is 4.0 percentage points.

(v) **CACE of phone calls among those who receive mail.** The CACE of phone calls among those who receive mail is the ITT divided by the contact rate: (41.8 - 40.3) / (1363/2287) = 2.5 percentage points

c) In Chapters 5 and 6, we discussed the use of instrumental variables regression to estimate CACEs when experiments involve noncompliance. Here, we can apply instrumental variables regression to a factorial experiment in which one factor encounters noncompliance. With the replication dataset at http://isps.research.yale.edu/FEDAI, use instrumental variables regression to estimate the parameters of the Vote equation in the following three-equation regression model:

$$PhoneContact_i = \alpha_0 + \alpha_1 Mail_i + \alpha_2 PhoneAssign_i + \alpha_3(PhoneAssign_i * Mail_i) + e_i$$
$$PhoneContact_i * Mail_i = \gamma_0 + \gamma_1 Mail_i + \gamma_2 PhoneAssign_i + \gamma_3(PhoneAssign_i * Mail_i) + \epsilon_i$$
$$Vote_i = \beta_0 + \beta_1 Mail_i + \beta_2 PhoneContact_i + \beta_3(PhoneContact_i * Mail_i) + u_i$$

Interpret the regression estimates in light of the five parameters you estimated in part (b). Which causal parameters does instrumental variables regression estimate or fail to estimate?


```
In [1]: import delim ./data/chapter09/Fieldhouse_et_al_unpublished_2010_expanded,clear

In [2]: rename m mail
        rename p phone_assign
        rename c phone_contact
        rename y vote
        rename c_m phone_contact_mail
        rename p_m phone_assign_mail


        gen mail_phone_contact = mail*phone_contact
        gen mail_phone_assign = mail*phone_assign

        ivregress 2sls vote mail ///
        (mail_phone_contact phone_contact = mail phone_assign mail_phone_assign)

Instrumental variables (2SLS) regression          Number of obs   =     15,300
                                                  Wald chi2(3)    =       3.38
                                                  Prob > chi2     =     0.3367
                                                  R-squared       =     0.0010
                                                  Root MSE        =     .49001

------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
mail_phone_~t |   .0253338   .0282274     0.90   0.369    -.0299908    .0806584
phone_contact |  -.0001781   .0186117    -0.01   0.992    -.0366565    .0363002
        mail |   .0060348   .0100671     0.60   0.549    -.0136962    .0257659
       _cons |   .3969878    .006809    58.30   0.000     .3836424    .4103332
```

```
------------------------------------------------------------------------------
Instrumented:  mail_phone_contact phone_contact
Instruments:   mail phone_assign mail_phone_assign
```

The intercept is the voting rate in the control group. The coefficient for "phone contact" is the estimated CACE for phones when no mail is assigned. The effect for "mail" is the ATE for mail when no phone calls are assigned. The coefficient for "mail:phone contact" is the extent to which the apparent CACE of phone calls increases when we move from the no-mail to the mail group. These estimates reproduce the estimates generated by hand above. Notice that IV regression does not report the effect of mail for non-compliers.