Field Experiments: Design, Analysis and Interpretation
Solutions for Chapter 4 Exercises

Alan S. Gerber and Donald P. Green*

April 8, 2019

Question 1

Important concepts: [10pts]

a) Define "covariate." Explain why covariates are (at least in principle) measured prior to the random allocation of subjects to treatment and control.
Answer:
A covariate is a variable that is (1) unaffected by the treatment and (2) used to predict outcomes. In order to increase the credibility of the claim that a given covariate is unaffected by the treatment, researchers typically restrict the set of covariates to those variables that are measured (or are measurable) prior to the random allocation of treatments.

b) Define "disturbance term."
Answer:
The disturbance term comprises all sources of variation in potential outcomes other than the average treatment effect. For example, in equation (4.7), the disturbance term is $u_i = Y_i(0) - \mu_{Y(0)} + [(Y_i(1) - \mu_{Y(1)}) - (Y_i(0) - \mu_{Y(0)})]D_i$. The disturbance term comprises the idiosyncratic variation in untreated responses $Y_i(0) - \mu_{Y(0)}$, plus the idiosyncratic variation in treatment effects $[(Y_i(1) - \mu_{Y(1)}) - (Y_i(0) - \mu_{Y(0)})]D_i$.

c) In equation (4.2), we demonstrated that rescaling the outcome by subtracting a pre-test leads to unbiased estimates of the ATE. Suppose that instead of subtracting the pre-test X_i, we subtracted a rescaled pretest cX_i, where c is some positive constant. Show that this procedure produces unbiased estimates of the ATE.
Answer:
The proof is similar to equation (4.2) and again makes use of the fact that the expected value of X_i is the same in the treatment and control groups when treatments are allocated randomly:

$$\begin{aligned}
E[\widehat{ATE}] &= E[Y_i - cX_i|D_i = 1] - E[Y_i - cX_i|D_i = 0)] \\
&= E[Y_i|D_i = 1] - E[cX_i|D_i = 1] - E[Y_i|D_i = 0] + E[cX_i|D_i = 0] \\
&= E[Y_i|D_i = 1] - cE[X_i|D_i = 1] - E[Y_i|D_i = 0] + cE[X_i|D_i = 0] \\
&= E[Y_i(1)] - E[Y_i(0)]
\end{aligned}$$

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

d) Show that the parameter b in equation (4.7) is identical to the ATE.
Answer:
Recall from Equation (4.7) that:

$$\begin{aligned} Y_i &= Y_i(0)(1 - D_i) + Y_i(1)D_i \\ &= Y_i(0) + (Y_i(1) - Y_i(0))D_i \\ &= \mu_{Y(0)} + [\mu_{Y(1)} - \mu_{Y(0)}]D_i + Y_i(0) - \mu_{Y(0)} + [(Y_i(1) - \mu_{Y(1)}) - (Y_i(0) - \mu_{Y(0)})]D_i \\ &= a + bD_i + u_i \end{aligned}$$

This equation implies that $b = \mu_{Y(1)} - \mu_{Y(0)}$, which is the ATE because the expected value of $Y_i(1)$ is $\mu_{Y(1)}$, and the expected value of $Y_i(0)$ is $\mu_{Y(0)}$.

Question 2

A researcher working with Israeli elementary school students sought to improve students' ability to solve logic puzzles.[1] Students in the treatment and control group initially took a computer-administered test, and the number of correctly solved puzzles was recorded. A few days later, students assigned to the control group were then given 30 minutes to improve their puzzle-solving skills by playing on a computer. During the same allotment of time, students in the treatment group listened to an instructor describe some rules of thumb to keep in mind when solving logic puzzles. All subjects then took a computer-administered post-test, and the number of correctly solved puzzles was recorded. The table below shows the results for each subject. [10pts]

Table 1: Question 2 Table

Subject	D	Pre-test	Post-test	Improvement
1	1	10	10	0
2	1	9	11	2
3	1	5	6	1
4	1	3	6	3
5	1	3	6	3
6	1	6	7	1
7	1	6	7	1
8	1	5	6	1
9	1	6	7	1
10	0	9	9	0
11	0	6	7	1
12	0	11	10	-1
13	0	4	5	1
14	0	3	3	0
15	0	10	10	0
16	0	7	8	1
17	0	7	7	0
18	0	8	10	2

[1] Dan Gendelman conducted this study in 2004 and shared it with us via personal communication.

a) As a randomization check, use randomization inference to test the null hypothesis that the pre-test scores are unaffected by treatment assignment.

```
        // download data from: http://hdl.handle.net/10079/xwdbs5h
        // copy and paste the url to your web browser

        clear
        use "Gendelman_2004.dta.dta"
        set seed 1234567
        rename treatment D
        rename posttest Y
        rename pretest X

        capture program drop Fstat
        program define Fstat, rclass
                regress D X
            return scalar Fs=e(F)
        end

        // calculate 48620 (18 choose 9) permutations
        tsrtest D r(Fs) using 4_2_Fstat.dta, overwrite: Fstat


        di "Fstat = "r(obsvStat)

        // p.value is different from R result due to
        // the rounding digits of F stats
        // the full permutation schedule is exactly the same
        di "p.value= "r(uppertail)

Two-sample randomization test for theta=r(Fs) of Fstat by D

Combinations:   48620 = (18 choose 9)
Assuming null=0
Observed theta: 1.274

Minimum time needed for exact test (h:m:s):  0:03:38
Mode: exact

progress: |...|

 p=0.31345 [one-tailed test of Ho:  theta(D==0)<=theta(D==1)]
 p=0.76851 [one-tailed test of Ho:  theta(D==0)>=theta(D==1)]
 p=0.31345 [two-tailed test of Ho:  theta(D==0)==theta(D==1)]

Saving log file to 4_2_Fstat.dta...done.

Fstat = 1.2743363
```

```
p.value= .31345125
```

We calculated the F-statistic of a regression of treatment assignment on the pretest score for all possible randomizations, and found that the observed F-statistic was larger than 31.35% of the simulated statistics, implying a p-value of 0.313. As expected,we fail to reject the null hypothesis that the treatment assignment is unrelated to the pretreatment covariate, pretest.

b) Use difference-in-means estimation to estimate the effect of the treatment on the post-test score. Form a 95% confidence interval.

```
// calculate ate
qui reg Y D
global tau = _b[D]
di "Estimated ate = " $tau

//
// RI under the null ate=ate

gen Y0_sim = Y
gen Y1_sim = Y
gen Y_sim = .
replace Y0_sim = Y - $tau if D==1
replace Y1_sim = Y + $tau if D==0

capture program drop ate_ci
program define ate_ci, rclass
        replace Y_sim = Y0_sim*(1-D) + Y1_sim*(D)
        regress Y_sim D
    return scalar Ys=_b[D]
end


tsrtest D r(Ys) using ate_ci.dta, overwrite: ate_ci

preserve
use "ate_ci.dta", clear
drop if _n==1

sort theta

// 95% confidence interval (CI)

// 95% CI is different from R result due to rounding
// the permutation test is exactly the same
```

```
        di "(" round(theta[floor(_N*0.025)], 0.001) ///
        ", "round(theta[floor(_N*0.975)], 0.001) ")"

        restore

Estimated ate = -.33333333



(18 missing values generated)

(9 real changes made)

(9 real changes made)



Two-sample randomization test for theta=r(Ys) of ate_ci by D

Combinations:   48620 = (18 choose 9)
Assuming null=0
Observed theta: -.3333

Minimum time needed for exact test (h:m:s):  0:03:30
Mode: exact

progress: |...|

 p=0.50428 [one-tailed test of Ho:  theta(D==0)<=theta(D==1)]
 p=0.50000 [one-tailed test of Ho:  theta(D==0)>=theta(D==1)]
 p=0.76450 [two-tailed test of Ho:  theta(D==0)==theta(D==1)]

Saving log file to ate_ci.dta...done.



(1 observation deleted)


(-2.259, 1.593)
```

We obtained a difference-in-means estimate of the ATE of -0.3333333 and a 95% confidence interval of $[-2.259, 1.593]$. This confidence interval is wide enough to include much larger and much smaller treatment effects – even crossing zero.

c) Use difference-in-differences estimation to estimate the effect of the treatment on the post-test

score. Form a 95% confidence interval, and compare it to the interval in part (b).

```
        rename improvement Y_improve

        // calculate ate.improve
        qui reg Y_improve D
        global tau_im = _b[D]
        di "ATE.improve = " $tau_im

        //
        // RI under the null ate=ate.improve

        replace Y0_sim = Y_improve
        replace Y1_sim = Y_improve
        replace Y_sim = .

        replace Y0_sim = Y_improve - $tau_im if D==1
        replace Y1_sim = Y_improve + $tau_im if D==0


        capture program drop ate_im_ci
        program define ate_im_ci, rclass
                replace Y_sim = Y0_sim*(1-D) + Y1_sim*(D)
                regress Y_sim D
            return scalar Ys_im=_b[D]
        end



        tsrtest D r(Ys_im) using ate_im_ci.dta, overwrite: ate_im_ci

        use "ate_im_ci.dta", clear
        drop if _n==1

        sort theta

        // 95% confidence interval (CI)

        // 95% CI is different from R result due to rounding
        // the permutation test is exactly the same

        di "95%CI = " "(" round(theta[floor(_N*0.025)], 0.001) " , " round(theta[floor(_N*0.
```

```
ATE.improve = 1

(18 real changes made)

(18 real changes made)
```

```
(0 real changes made)

(9 real changes made)

(9 real changes made)



Two-sample randomization test for theta=r(Ys_im) of ate_im_ci by D

Combinations:   48620 = (18 choose 9)
Assuming null=0
Observed theta: 1

Minimum time needed for exact test (h:m:s):  0:03:33
Mode: exact

progress: |...|

 p=0.58739 [one-tailed test of Ho:  theta(D==0)<=theta(D==1)]
 p=0.59726 [one-tailed test of Ho:  theta(D==0)>=theta(D==1)]
 p=0.58739 [two-tailed test of Ho:  theta(D==0)==theta(D==1)]

Saving log file to ate_im_ci.dta...done.


(1 observation deleted)


95%CI = (.111 , 1.889)
```

By subtracting a pre-test, we have sharpened our estimates. The difference-in-differences estimate of the ATE is 1 and the 95% confidence interval is [0.11, 1.89]. No longer does the 95% confidence interval cross zero, meaning we can be confidence at the 95% level that the estimated ATE is larger than zero. This contrasts with part b) where the background variability in test scores made the estimation of a small treatment effect more difficult.

Question 3

The table below illustrates the problems that may arise when researchers exercise discretion over what results to report to readers. Suppose the true ATE associated with a given treatment were 1.0. The table reports the estimated ATE from nine experiments, each of which involves approximately 200 subjects. Each study produces two estimates, one based on a difference-in-means and another using regression to control for covariates. In principle, both estimators generate unbiased estimates, and covariate adjustment has a slight edge in terms of precision. Suppose the researchers conducting each study use the following decision rule: "Estimate the ATE using both estimators and report

whichever estimate is larger." Under this reporting policy, are the reported estimates unbiased? Why or why not? [6 pts]
Answer:
This procedure leads to biased estimates. Although each estimator is unbiased, the greater of two unbiased estimates is not unbiased. One can think of this procedure as "Report the no-covariates estimate unless the with-covariates estimate is larger, in which case report the with-covariates estimate." On its own, the no-covariates estimate is unbiased, but it tends to be corrected when it generates a lower-than-average estimate. In this example, the average estimate generated by this reporting procedure is $12/9 = 1.33$, which is greater than the true ATE of 1.0.

Table 2: Question 3 table

Study	No covariates	With covariates	Greater of two estimates
1	5	4	5
2	3	3	3
3	2	2	2
4	6	5	6
5	1	1	1
6	0	0	0
7	-3	-1	-1
8	-5	-4	-4
9	0	-1	0
Average	1	1	1.33
Standard Deviation	3.54	2.83	3.08

Question 4

Table 4.1 contains a column of treatment assignments, D_i, that reflects a complete random assignment of 20 schools to treatment and 20 schools to control. [14pts]

a) Use equation (2.2) to generate observed outcomes based on these assigned treatments. Regress Y_i on D_i and interpret the slope and intercept. Is the estimated slope the same as the estimated ATE based on a difference-in-means?

```
D  <- teach$D
Y1 <- teach$y1
Y0 <- teach$y0
X <- teach$x

Y <- Y0*(1-D) + Y1*(D)     # Equation 2.2

fit <- lm(Y~D)
arm::display(fit)

## lm(formula = Y ~ D)
##             coef.est coef.se
```

```
## (Intercept) 26.85     3.33
## D           10.70     4.71
## ---
## n = 40, k = 2
## residual sd = 14.89, R-Squared = 0.12

diff_means <- mean(Y[D==1])-mean(Y[D==0])
diff_means

## [1] 10.7
```

The estimate obtained with OLS regression (10.7) is identical to the estimate obtained with difference-in-means (10.7).

b) Regress treated and untreated outcomes on X_i to see whether the condition in equation (4.6) appears to hold. What do you infer about the advisability of rescaling the dependent variable so that the outcome is a change (i.e., $Y_i - X_i$)?

```
fit.1 <- lm(Y~X,subset=D==1)
fit.0 <- lm(Y~X,subset=D==0)

sum_of_coefficients <- fit.1$coefficients[2] + fit.0$coefficients[2]
sum_of_coefficients

##     X
## 1.846

Ydiff <- Y-X
fit.diff <- lm(Ydiff~D)
arm::display(fit.diff)

## lm(formula = Ydiff ~ D)
##             coef.est coef.se
## (Intercept) -1.00     1.15
## D            4.85     1.62
## ---
## n = 40, k = 2
## residual sd = 5.13, R-Squared = 0.19
```

Substituting regression estimates for the true ratio of covariances to variances satisfies the inequality, suggesting that the use of this covariate will improve precision.

$$\frac{Cov(\widehat{Y_i(0)}, X_i)}{Var(X_i)} + \frac{Cov(\widehat{Y_i(1)}, X_i)}{Var(X_i)} = 0.8995221 + 0.9465386 = 1.8460607$$

We also see that the standard errors have shrunk substantially – the standard error for a regression of Y on D is 4.7093133, whereas the standard error for the regression of the change in Y on D is 1.6226603

c) Regress Y_i on D_i and X_i. Interpret the regression coefficients, contrasting these results with those obtained from a regression of Y_i on D_i alone.

```
fit.cov <- lm(Y~D+X)
arm::display(fit.cov)

## lm(formula = Y ~ D + X)
##             coef.est coef.se
## (Intercept) 1.22     1.88
## D           5.32     1.63
## X           0.92     0.05
## ---
## n = 40, k = 3
## residual sd = 5.05, R-Squared = 0.90
```

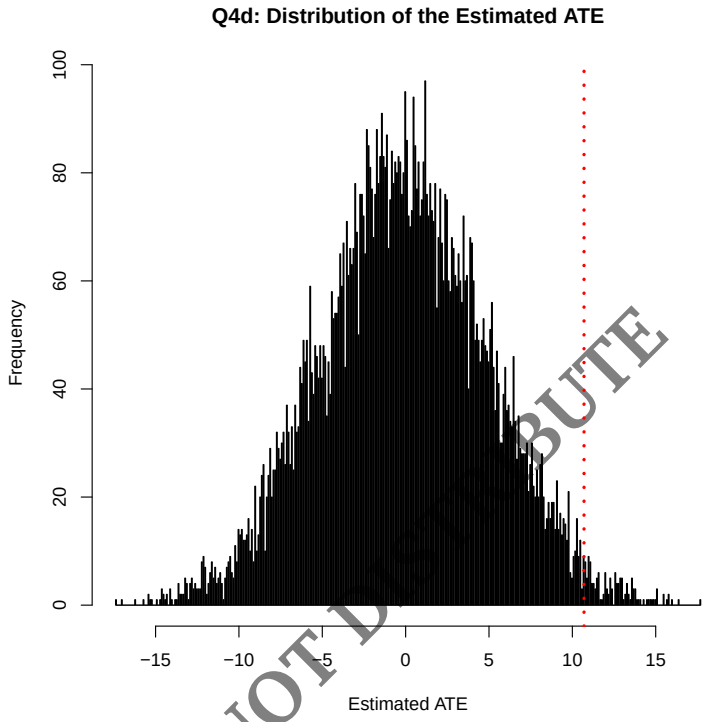
The estimated ATE (5.32) is now roughly half the size as the original difference-in-means. (This estimate also happens to be much closer to the true ATE of 4.0.) Comparing the estimated standard errors from both regressions suggests that the inclusion of a covariate has greatly improved precision.

d) With the estimates obtained in part (a), use randomization inference (as described in Chapter 3) to evaluate the sharp null hypothesis of no effect for any school. To obtain the sampling distribution under the sharp null hypothesis, simulate 100,000 random assignments, and for each simulated sample, estimate the ATE using a regression of Y_i on d_i. Interpret the results.


```
perms <- genperms(D,maxiter=10000)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 137846528820 to perform exact estimation.

probs <- genprobexact(D)
ate <- estate(Y,D,prob=probs)
Ys <- genouts(Y,D,ate=0)

distout <- gendist(Ys,perms,prob=probs)
p.value <- mean(abs(distout)>abs(ate))

ate

## [1] 10.7

p.value
```

```
## [1] 0.0291
```

```
hist(distout, breaks=1000,
     main="Q4d: Distribution of the Estimated ATE",
     xlab="Estimated ATE")
abline(v=ate, col="red", lty=3, lwd=3)
```



We use a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject. We find a two-tailed p-value of 0.029, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some positive effect.

e) Using the estimator in part (c), use randomization inference to evaluate the sharp null hypothesis of no effect for any school. To obtain the sampling distribution under the sharp null hypothesis, simulate 100,000 random assignments, and for each simulated sample, estimate the ATE using a regression of Y_i on D_i and X_i. Interpret the results.

```
ate_cov <- estate(Y,D,X,prob=probs)
distout_cov <- gendist(Ys,perms,X,prob=probs)
p.value_cov <- mean(abs(distout_cov)>abs(ate_cov))
ate_cov
```

```
##     Z
## 5.316
```

```
p.value_cov
```

```
## [1] 0.0026
```

We again use a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject. We find a two-tailed p-value of 0.003, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some effect.

f) Use the estimated ATE in part (a) to construct a full schedule of potential outcomes for all schools, assuming that every school has the same treatment effect. Using this simulated schedule of potential outcomes, construct a 95% confidence interval for the sample average treatment effect in the following way. First, assign each subject to treatment or control, and estimate the ATE by a regression of Y_i on D_i. Repeat this procedure until you have 100,000 estimates of the ATE. Order the estimates from smallest to largest. The 2,501st estimate marks the 2.5th percentile, and the 97,500th estimate marks the 97.5th percentile. Interpret the results.

```
Ys <- genouts(Y,D,ate=ate)
distout <- gendist(Ys,perms,prob=probs)
ci.95 <- quantile(distout, probs=c(0.025, .975))
ci.95

##  2.5% 97.5%
##  1.53 19.84
```

The confidence interval stretches from [1.53, 19.84] implying that the ATE is positive but its location is subject to a great deal of statistical uncertainty. Our best guess is 10.7, but the interval ranges from a small positive value to a truly massive effect.

g) Use the estimated ATE in part (c) to construct a full schedule of potential outcomes for all schools, assuming that every school has the same treatment effect. Using this simulated schedule of potential outcomes, simulate the 95% confidence interval for the sample average treatment effect estimated by a regression of Y_i on D_i and X_i. Interpret the results. Is this confidence interval narrower than one you generated in question (f)?

```
Ys_cov <- genouts(Y,D,ate=ate_cov)
distout_cov <- gendist(Ys_cov,perms,X,prob=probs)
ci.95_cov <- quantile(distout_cov, probs=c(0.025, .975))
ci.95_cov

##  2.5% 97.5%
## 2.225 8.442
```

The confidence interval now stretches from [2.23, 8.44]. Interestingly, this interval no longer contains the estimate obtained without controls for covariates. Our best guess is now 5.32, and our 95% interval is now roughly one-third as wide as before.

Question 5

Randomizations are said to be "restricted" when the set of all possible random allocations is narrowed to exclude allocations that have inadequate covariate balance. Suppose, for example, that the assignment of treatments (D_i) in Table 4.1 was conducted subject to the restriction that a regression of D_i on X_i (the pretest) does not allow the researcher to reject the sharp null hypothesis

of no effect of X_i on D_i at the 0.05 significance level) produces a p-value on that is greater than 0.05. In other words, had the researcher found that the assigned D_i were significantly predicted by X_i, the random allocation would have been conducted again, until the D_i met this criterion. [10pts]

a) Conduct a series of random assignments in order to calculate the weighting variable w_i; for units in the treatment group, this weight is defined as the inverse of the probability of being assigned to treatment, and for units in the control group, this weight is defined as the inverse of the probability of being assigned to control. See Table 4.2 for an example. Does w_i appear to vary within the treatment group or within the control group?

```
D  <- teach$D
Y1 <- teach$y1
Y0 <- teach$y0
X <- teach$x

Y <- Y0*(1-D) + Y1*(D)
N <- length(D)


randfun <- function() {
  teststat <- -1
  while (teststat < 0.05) {
                Zri <- sample(D)
                teststat <- summary(lm(Zri~X))$coefficients[2,4]
        }
        return(Zri)
}

# notice the use of the restricted randomization function.
# restricted randomization often generates unequal probabilities of assignment.
# if so, inverse probability weighting is required.

perms <- genperms.custom(numiter=10000,randfun=randfun)
probs <- genprob(perms)
weights <- (1/probs) *D + (1/(1-probs))*(1-D)
var.weights.treat <- var(weights[D==1])
var.weights.control <- var(weights[D==0])
```

The variance of the weights is 4×10^{-4} in the treatment condition and 6×10^{-4} in the control condition. Indeed, units do have different probabilities of assignments as a result of the restriction scheme, but the differences are small.

b) Use randomization inference to test the sharp null hypothesis that D_i has no effect on Y_i by regressing Y_i on D_i and comparing the estimate to the sampling distribution under the null hypothesis. Make sure that your sampling distribution includes only random allocations that satisfy the restriction mentioned above. Be sure to weight units by inverse probability weights as produced by the random allocation procedure. Estimate the ATE, calculate the p-value, and interpret the results.

```
ate <- estate(Y,D,prob=probs)
Ys <- genouts(Y,D,ate=0)
distout <- gendist(Ys,perms,prob=probs)
p.value <- mean(abs(distout) > abs(ate))
ate

## [1] 10.73

p.value

## [1] 0.0054
```

The IPW estimate of the ATE is 10.73, which is close to the unweighted estimate above. Using a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject, we find a p-value of 0.005, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some effect.

c) Use randomization inference to test the sharp null hypothesis that D_i has no effect on Y_i by regressing Y_i on D_i and X_i and comparing the estimate to the sampling distribution under the null hypothesis. Estimate the ATE, calculate the p-value, and interpret the results.


```
perms <- genperms.custom(numiter=10000,randfun=randfun)
probs <- genprob(perms)
ate_cov <- estate(Y,D,X,prob=probs)
Ys <- genouts(Y,D,ate=0)
distout_cov <- gendist(Ys,perms,X,prob=probs)
p.value_cov <- mean(abs(distout_cov) > abs(ate_cov))
ate_cov

##     Z
## 5.346

p.value_cov

## [1] 0.0017
```

The IPW estimate of the ATE is 5.35, which is close to the unweighted estimate above. We again use a two-tailed test in order to evaluate the null hypothesis that the treatment has no effect for any subject. We find a p-value of 0.002, which leads us to reject the null hypothesis in favor of the alternative hypothesis that the treatment has some effect.

d) Compare the sampling distributions under the null hypothesis in parts (a) and (b) to the sampling distributions obtained in exercises 4(d) and 4(e), which assumed that the randomization was unrestricted.

```r
## Sampling Distributions from 4(d) and 4(e)
perms_complete_RA <- genperms(D,maxiter=10000)

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 137846528820 to perform exact estimation.

probs_complete_RA <- genprobexact(D)

ate_complete_RA <- estate(Y,D,prob=probs_complete_RA)
Ys_complete_RA <- genouts(Y,D,ate=ate_complete_RA)
distout_complete_RA <- gendist(Ys_complete_RA,perms_complete_RA,
                               prob=probs_complete_RA)
se_complete_RA <- sd(distout_complete_RA)
se_complete_RA

## [1] 4.601

ate_cov_complete_RA <- estate(Y,D,X,prob=probs_complete_RA)
Ys_cov_complete_RA <- genouts(Y,D,ate=ate_cov_complete_RA)
distout_cov_complete_RA <- gendist(Ys_cov_complete_RA,perms_complete_RA,X,
                                   prob=probs_complete_RA)
se_cov_complete_RA <- sd(distout_cov_complete_RA)
se_cov_complete_RA

## [1] 1.593

## Sampling Distributions from 5(a) and 5(b)
perms_restricted_RA <- genperms.custom(numiter=10000,randfun=randfun)
probs_restricted_RA <- genprob(perms_restricted_RA)

ate_restricted_RA <- estate(Y,D,prob=probs_restricted_RA)
Ys_restricted_RA <- genouts(Y,D,ate=ate_restricted_RA)
distout_restricted_RA <- gendist(Ys_restricted_RA,perms_restricted_RA,
                                 prob=probs_restricted_RA)
se_restricted_RA <- sd(distout_restricted_RA)
se_restricted_RA

## [1] 4.199

ate_cov_restricted_RA <- estate(Y,D,X,prob=probs_restricted_RA)
Ys_cov_restricted_RA <- genouts(Y,D,ate=ate_cov_restricted_RA)
distout_cov_restricted_RA <- gendist(Ys_cov_restricted_RA,perms_restricted_RA,X,
                                     prob=probs_restricted_RA)
se_cov_restricted_RA <- sd(distout_cov_restricted_RA)
se_cov_restricted_RA

## [1] 1.607
```

Table 3: Summary of Estimated Standard Errors

	Without Covariates	With Covariates
Complete Random Assignment	4.601	1.593
Restricted Random Assignment	4.199	1.607

Without covariates and assuming complete randomization, we obtain a standard error of 4.601. Under restricted randomization, the standard error declines to 4.199. Including a covariate and assuming complete randomization, we obtain a standard error of 1.593. Under restricted randomization, the standard error remains essentially unchanged at 1.607. Restricted randomization is akin to blocking, in that it rules out random allocations that result in imbalance; however, its advantages in terms of precision are limited when the researcher controls for a strongly prognostic covariate, which achieves most of the precision gains associated with blocking.

Question 6

One way to practice your experimental design skills is to undertake a mock randomization of an existing non-experimental dataset. In this exercise, the existing dataset is treated as though it were a baseline data collection effort that an experimental researcher gathered in preparation for a random intervention. The actual data in question come from a panel study of Russian villagers. Villagers from randomly selected rural areas of Russia were interviewed in 1995 and re-interviewed in 1996 and 1997. Our attention focuses on the 462 respondents who were interviewed in all three waves and provided answers to questions about their income, church membership, and evaluation of national conditions (i.e., how well are things going in Russia?). Imagine that an experimental intervention occurred after the 1996 survey and that national evaluations in the 1997 survey were the experimental outcome of interest. The dataset provided at isps.research.yale.due/FEDAI contains the following pre-treatment covariates that may be used for blocking: sex, church membership, social class, and evaluations of national conditions in 1995 and 1996. As you design your experiment, imagine that "post-intervention" evaluations of national conditions in 1997 were unknown. [10pts]

a) One way to develop a sense of which variables are likely to predict post-intervention evaluations of national conditions in 1997 is to regress evaluations of national conditions in 1996 on sex, church membership, social class, and evaluations in 1995. Which of these variables seem to most strongly predict evaluations of national conditions in 1996? What is the R^2 from this regression?

```
russia <- within(russia,{
  female <- as.numeric(sexresp6 == "woman")
  class <- relevel(group6,ref="very poor")
  church_member <- as.numeric(memberc6=="yes")
  id <- 1:nrow(russia)
  class_verypoor <- as.numeric(class=="very poor")
  class_poor <- as.numeric(class=="poor")
  class_middle <- as.numeric(class=="middle")
  class_morethanmiddle <- as.numeric(class=="more than middle")
})
```

```
fit <- lm(index96 ~ index95 + female + church_member + class, data=russia)
summary(fit)$r.squared

## [1] 0.3937

fit.nolag <- lm(index96 ~ female + church_member + class, data=russia)
summary(fit.nolag)$r.squared

## [1] 0.02828
```

The regression treats "index95" as a continuous variable and all others as categorical. the R-squared is 0.394, which implies that the regressors predict about 40% of the variance in "index96". The strongest predictor is 95, the lagged dependent variable. Had we omitted this variable from the model, the R-squared would have fallen to 0.028.

b) Suppose you were to design a block random assignment in order to predict evaluations in 1997. Use the R package `blockTools` (for example code, see isps.research.yale.due/FEDAI) to perform a block random assignment, blocking on sex, church membership, social class, and evaluations in 1996. Decide for yourself how many subjects to include in each block. Compare the treatment and control groups to verify that blocking produced groups that have the same profile of sex, church membership, social class, and evaluations in 1996.


```
block.out <- block(data = russia, n.tr = 2,
                   id.vars = "id",algorithm="randGreedy",
                    block.vars = c("female", "church_member",
                                   "index96","class_verypoor",
                                   "class_poor", "class_middle"))
assign.out <- assignment(block.out)

# extracting the treatment assignment from blockTools takes some work
# The commands below check to see which ID numbers appear on the
# list of assign.out's assignment to Treatment 1

russia$Z_blocked <- as.numeric(russia$id %in%
                         as.numeric(as.character(
                           unlist(assign.out$assg[[1]]["Treatment 1"]))))

arm::display(lm(Z_blocked ~ female + church_member + class + index96,
                data=russia))

## lm(formula = Z_blocked ~ female + church_member + class + index96,
##     data = russia)
##                         coef.est coef.se
## (Intercept)              0.53     0.17
## female                   0.00     0.06
## church_member            0.01     0.08
## classpoor               -0.04     0.16
```

```
## classmiddle            -0.05     0.16
## classmore than middle -0.05     0.22
## index96                 0.00     0.01
## ---
## n = 462, k = 7
## residual sd = 0.50, R-Squared = 0.00
```

Using the package `blockTools`, we created blocks of size 2 based on gender, church membership, evaluations in 1996, and social class. The package also conducts complete random assignment – with some work, this assignment can be extracted. Regressing this treatment assignment on the set of pretreatment covariates reveals that the groups are well balanced.

c) Suppose you wanted to assess how well your blocking design performed in terms of increasing the precision with which treatment effects are estimated. Of course, there was no actual treatment in this case, but imagine that shortly after the survey in 1996, a treatment were administered to a randomly selected treatment group. (Here is an instance in which the sharp null hypothesis of no effect is known to be true!) The outcome from this imaginary experiment is evaluations of national conditions in 1997. Compare the sampling distribution of the estimated treatment effect (which should be centered on zero) under balanced complete random assignment to the sampling distribution of the estimated treatment effect under block random assignment.
Answer:
See below

d) Calculate the sampling distribution of the estimated treatment effect under balanced complete random assignment using regression to control for the variables that would have otherwise been used to form blocks. Compare the resulting distribution to the sampling distribution of the estimated treatment effect under block random assignment. Does blocking produce an appreciable gain in precision over what is achieved by covariate adjustment?

```
sims <- 10000
results <- matrix(NA,sims,3)
colnames(results) <- c("complete","adjusted","blocked")
N <- nrow(russia)

for(i in 1:sims) {
    # Complete RA, with and without adjustment
    russia$Z_complete <- ifelse(1:N %in% sample(N, N/2), 1, 0)
    results[i,1] <- lm(index97 ~ Z_complete, data=russia)$coefficients[2]
    results[i,2] <- lm(index97 ~ Z_complete + female + church_member + class + index96,
                       data=russia)$coefficients[2]

    # Blocked RA, without adjustment
    assign.out <- assignment(block.out)
    russia$Z_blocked <- as.numeric(russia$id %in%
                         as.numeric(as.character(
                           unlist(assign.out$assg[[1]]["Treatment 1"]))))
    results[i,3] <- lm(index97 ~ Z_blocked, data=russia)$coefficients[2]
    }
```

```r
# use apply() to extract means and SDs for each column (2 refers to columns)
results_table <- rbind(apply(results,2,mean),apply(results,2,sd))
rownames(results_table) <- c("Average Estimate", "Standard Error")
results_forxtable <- xtable(results_table,caption="Comparison of 3 estimators")
```

```r
print.xtable(results_forxtable,caption.placement="top",table.placement="H")
```

Table 4: Comparison of 3 estimators

	complete	adjusted	blocked
Average Estimate	0.00	0.00	-0.00
Standard Error	0.17	0.13	0.13

The table above shows a comparison of three estimators of the ATE: difference-in-means under complete random assignment, OLS with covariate adjustment under complete random assignment, and difference-in-means under blocked random assignment. All three estimators are centered on the true ATE of zero. The least precise method is complete random assignment with the difference-in-means estimator, which produces a standard error of 0.169. The most precise approach is blocking, which produces a standard error of 0.131. Slightly inferior to blocking is covariate adjustment, which produces a standard error of 0.133. Blocking's slight superiority stems from the fact that, under blocking, there is no incidental correlation between the covariates and random assignments and therefore no "collinearity penalty."

Question 7

Researchers may be concerned about using block randomization when they are unsure whether the variable used to form the blocks actually predicts the outcome. Consider the case in which blocks are formed randomly – in other words, the variable used to form the blocks has no prognostic value whatsoever. Below is a schedule of potential outcomes for four observations. [10pts]

Table 5: Question 7 Table

Subject	Y(0)	Y(1)
A	1	2
B	0	3
C	2	2
D	5	5

a) Suppose you were to use complete random assignment such that $m = 2$ units are assigned to treatment. What is the sampling variance of the difference-in-means estimator across all six possible random assignments?

The average estimated ATE is 1.0, which is the true ATE. The variance of the estimated ATEs over all 6 possible randomizations is 2.833.

Table 6: Question 7a table

Treated Units	$\bar{Y(1)}$	$\bar{Y(0)}$	$\widehat{ATE}$
A and B	2.5	3.5	-1
A and C	2	2.5	-0.5
A and D	3.5	1	2.5
B and C	2.5	3	-0.5
B and D	4	1.5	2.5
C and D	3.5	0.5	3

b) Suppose you were to form blocks by randomly pairing the observations. Within each pair, you randomly allocate one subject to treatment and the other to control so that $m = 2$ units are assigned to treatment.There are three possible blocking schemes; for each blocking scheme, there are four possible random assignments. What is the sampling variance of the difference-in-means estimator across all twelve possible random assignments?

Table 7: Question 7b table

	Treated Units	$\bar{Y(1)}$	$\bar{Y(0)}$	$\widehat{ATE}$
AB and CD blocked	A,C	2	2.5	-0.5
	A,D	3.5	1	2.5
	B,D	4	1.5	2.5
	B,C	2.5	3	-0.5
AC and BD blocked	A,B	2.5	3.5	-1
	A,D	3.5	1	2.5
	C,B	2.5	3	-0.5
	C,D	3.5	0.5	3
AD and BC blocked	A,B	2.5	3.5	-1
	A,C	2	2.5	-0.5
	D,B	4	1.5	2.5
	D,C	3.5	0.5	3

Across the 12 possible random assignments, the variance of the estimated ATE is again 2.833. Notice that every estimate in the previous table appears in this table twice.

c) From this example, what do you infer about the risks of blocking on a non-prognostic covariate?
Answer:
There is no risk of increasing variance with a useless blocking variable; at worst, the variable will be random noise, in which case the sampling variance will be the same as a design without blocking.

Question 8

Sometimes researchers randomly assign subjects from lists that are later discovered to have duplicate entries. Suppose, for example, that a fund-raising experiment randomly assigns 500 of 1,000 names to a treatment that consists of an invitation to contribute to a charitable cause. However, it is later discovered that 600 names appear once and 200 names appear twice. Before the invitations are mailed, duplicate invitations are discarded, so that no one receives more than one invitation. [10pts]

a) What is the probability of assignment to the treatment group among those whose names appeared once in the original list? What is the probability of assignment to the treatment group among those whose names appeared twice in the original list?
Answer:
The probability of being assigned to treatment if your name appears once is 0.5. The probability of being assigned to treatment if your name is a duplicate is $0.5 + (0.5)(0.5) = 0.75$, where the first term is the probability you were assigned to treatment the first time your name came up and the second term is the probability you were assigned to control the first time multiplied by the probability you were assigned to treatment the second time.

b) Of the 800 unique names in the original list, how many would you expect to be assigned to treatment and control?
Answer:
Of the 600 unique names that appear once, 300 are, in expectation, allocated to treatment. Of the 200 unique names that appear twice, 150 are, in expectation, allocated to treatment. Thus, we expect a total of 450 unique names in the treatment group.

c) What estimation procedure should one use in order to obtain unbiased estimates of the ATE?
Answer:
One should analyze the experiment as though it were randomized in two blocks: the names that appear once and the names that appear twice. Use an estimator like equation (4.11).

Question 9

Gerber and Green conducted a mobilization experiment in which calls from a large commercial phone bank urged voters in Iowa and Michigan to vote in the November 2002 election.[2] The randomization was conducted within four blocks: uncompetitive congressional districts in Iowa, competitive congressional districts in Iowa, uncompetitive congressional districts in Michigan, and competitive congressional districts in Michigan. Table 4.3 presents results only for one-voter households in order to sidestep the complications of cluster assignment. [10pts]

a) Within each of the four blocks, what was the apparent effect of being called by a phone bank on voter turnout?
Answer:
From the "Estimated ATE" Row: Block 1: .0096, Block 2: -.0078, Block 3: -.0136, Block 4: .0083. Substantively, these results suggest that calls encouraging voter turnout had effects ranging from -1.4 percentage points to +1.0 percentage point.

[2] Gerber and Green 2005.

b) When all of the subjects in this experiment are combined (see the rightmost column of the table), turnout seems substantially higher in the treatment group than the control group. Explain why this comparison gives a biased estimate of the ATE.
Answer:
This estimator is biased because individuals in each stratum had different propensities to enter into treatment. The uncompetitive Michigan block has the lowest rate of treatment and also has the lowest rate of voting in the control group. Overall, blocks with higher rates of treatment tend to have higher rates of voting in the control group, which accounts for the upward bias.

c) Using the weighted estimator described in Chapter 3, show the calculations used to generate an unbiased estimate of the overall ATE.

```
ests <- c(.00964, -.007829, -.01362,.008271)
shareoftotalN <- c(0.049487, 0.1520981, 0.626616, 0.171799)
overall_ate <-sum(ests*shareoftotalN)
overall_ate

## [1] -0.007827
```

d) When analyzing block randomized experiments, researchers frequently use regression to estimate the ATE by regressing the outcome on the treatment and indicator variables for each of the blocks (omitting one block if the regression includes an intercept.) This regression estimator places extra weight on blocks that allocate approximately half of the subjects to the treatment condition (i.e., $P_j = 0.5$) because these blocks tend to estimate the within-block ATE with less sampling variability. Compare the four OLS weights to the weights W_j used in part (c).
Answer:
The weights used in part (c) are based on the share of the subject pool that is in each block. This weighting scheme places a great deal of weight on the relatively large Michigan block. By contrast, the OLS weights are a blend of the number of subjects in each block and each block's balance between treatment and control allocations. Because the blocks do not differ very much in terms of their allocation rates, the OLS weights tend to be similar across blocks.

e) Regression provides an easy way to calculate the weighted estimate of the ATE in part (c) above. For each treatment subject i, compute the proportion of subjects in the same block who were assigned to the treatment group. For control subjects, compute the proportion of subjects in the same block who were assigned to the control group. Call this variable q_i. Regress outcomes on treatment, weighting each observation by $1/q_i$, and show that this type of weighted regression produces the same estimate as weighting the estimated ATEs for each block.

```
Y <- phones$vote02
block <- phones$strata
Z <- phones$treat2

## Proportion of subjects in each block assigned to treatment
block.pr <- tapply(Z, block, mean)

q <- rep(NA, length(Y))
```

```r
for(i in 1:4){
  q[block==i] <- block.pr[i]*Z[block==i] + (1-block.pr[i])*(1-Z[block==i])
}

fit <- lm(Y ~ Z, weights=1/q)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ Z, weights = 1/q)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -4.051 -0.469 -0.469  0.537  4.786
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.466198   0.000727  641.31  < 2e-16 ***
## Z           -0.007828   0.001028   -7.61  2.7e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.705 on 940713 degrees of freedom
## Multiple R-squared:  6.16e-05,Adjusted R-squared:  6.06e-05
## F-statistic:   58 on 1 and 940713 DF,  p-value: 2.65e-14
```

The coefficient on the treatment indicator is -0.0078, which is the same as was found in part c.

Question 10

The 2003 Kansas City voter mobilization experiment described in Chapter 3 is a cluster randomized design in which 28 precincts comprising 9,712 voters were randomly assigned to treatment and control.[3] The study contains a wealth of covariates: the registrar recorded whether each voter participated in elections dating back to 1996. The dataset may be obtained at isps.research.yale.edu/FEDAI. [10pts]

a) Test the balance of the treatment and control groups by looking at whether past turnout predicts treatment assignment. Regress treatment assignment on the entire set of past votes, and calculate the F-statistic. Use randomization inference to test the null hypothesis that none of the past turnout variables predict treatment assignment. Remember that to simulate the distribution of the F-statistic, you must generate 1,000 random cluster assignments and calculate the F-statistic for each simulated assignment. Judging from the p-value of this test, what does the F-statistic seem to suggest about whether subjects in the treatment and control groups have comparable background characteristics?

[3] Arceneaux 2005.

```
Z <-  kansas$treatmen
Y <- kansas$vote03
clust <- kansas$unit
covs <- as.matrix(kansas[,2:21])  # covariates are past voter turnout

probs <- genprobexact(Z,clustvar=clust)  # subjects are clustered by precinct
perms <- genperms(Z,maxiter=1000,clustvar=clust)    # clustered assignment

## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 40116600 to perform exact estimation.

numiter <- ncol(perms)

Fstat <- summary(lm(Z~covs))$fstatistic[1]   # F-statistic from actual data

Fstatstore <- rep(NA,numiter)
for (i in 1:numiter) {
  Fstatstore[i] <- summary(lm(perms[,i]~covs))$fstatistic[1]
        }

p.value <- mean(Fstatstore >= Fstat)
p.value

## [1] 0.936
```

Using randomization inference, we recover a p-value of 0.936; we therefore cannot reject the null hypothesis of random assignment.

b) Regress turnout in 2003 (after the treatment was administered) on the experimental assignment and the full set of covariates. Interpret the estimated ATE. Use randomization inference to test the sharp null hypothesis that experimental assignment had no effect on any subject's decision to vote.

```
ate <- estate(Y,Z,X=covs,prob=probs)
Ys <- genouts(Y,Z,ate=0)
distout <- gendist(Ys,perms,X=covs,prob=probs)
p.value.onetailed <- mean(distout >= ate)

ate

##       Z
## 0.05596

p.value.onetailed

## [1] 0.005
```

The estimate of the treatment effect is 0.056, implying that treatment increased turnout by 5.6 percentage points. This finding is statistically significant. Under the sharp null, estimates as large or larger only occur 0.5% of the time.

c) When analyzing cluster randomized experiments with clusters of varying size, one concern is that difference-in-means estimation is prone to bias. This concern also applies to regression. In order to sidestep this problem, researchers may choose to use the difference-in-totals estimator in equation (3.24) to estimate the ATE. Estimate the ATE using this estimator.

```
ateHT <- estate(Y,Z,prob=probs,HT=TRUE)
ateHT


## [1] 0.05395
```

The difference-in-totals estimate of the treatment effect is that treatment increased turnout by 5.4 percentage points.

d) Use randomization inference to test the sharp null hypothesis that treatment assignment had no effect, using the difference-in-totals estimator.

```
distoutHT <- gendist(Ys,perms,prob=probs,HT=TRUE)
p.value.onesidedHT <- mean(distoutHT >= ateHT)
p.value.onesidedHT


## [1] 0.198
```

Estimates generated under the sharp null equaled or exceeded the observed difference-in-totals 19.8% of the time, meaning we cannot reject the null.

e) The difference-in-totals estimator can generate imprecise estimates, but its precision can be improved by incorporating information about covariates. Create a new outcome variable that is the difference between a subject's turnout (1 = vote, 0 = abstain) and the average rate of turnout in all past elections. Now, using this "differenced" outcome variable, estimate the ATE using the difference-in-totals estimator, and test the sharp null hypothesis of no effect.

```
Y_diff <- Y - rowMeans(covs)

ateHT2 <- estate(Y_diff,Z,prob=probs,HT=TRUE)  # difference-in-differenced totals
Ys <- genouts(Y_diff,Z,ate=0)
distoutHT2 <- gendist(Ys,perms,prob=probs,HT=TRUE)
p.value.onesidedHT2 <- mean(distoutHT2 >= ateHT2)

ateHT2


## [1] 0.04874


p.value.onesidedHT2


## [1] 0.012
```

Using the differenced outcome variable tightened our estimates – the p-value under the sharp null is now 0.012, meaning we can reject the sharp null of no effect for any unit.