Section 2 : Regression Adjustment on Experimental Data

Andrew Bertoli

31 January 2014

Roadmap

1. OLS with Experimental Data
2. Fixed Effects
3. Voting Data

OLS with Experimental Data

Two reasons to prefer experiments

1. More realistic assumptions
2. Less opportunities for dishonest research

OLS with Experimental Data

Two reasons to prefer experiments

1. **More realistic assumptions**
2. Less opportunities for dishonest research

OLS with Experimental Data

Assumptions for an Experiment

1. $\{Y_{it}, Y_{ic}\} \perp\!\!\!\perp T_i$
2. SUTVA or non-interference

OLS with Experimental Data

Running an experiment is like randomly sampling from all the Y_{it}'s and Y_{ic}'s.

Treatment
(Y_{it}'s)

Control
(Y_{ic}'s)

OLS with Experimental Data

Estimator

$$\hat{\tau} = ave(Y_i : T_i = 1) - ave(Y_i : T_i = 0)$$

$$\widehat{SE}(\hat{\tau}) = \sqrt{\frac{\hat{\sigma}_T^2}{m} + \frac{\hat{\sigma}_C^2}{n-m}}$$

where m is the number of units in the treatment group. This measure corrects for heteroscedasticity.

Remember that we estimate $\hat{\sigma}_T^2$ and $\hat{\sigma}_C^2$ using the standard method of calculating the sample variance:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$$

OLS with Experimental Data

Motivating Example

OLS with Experimental Data

Motivating Example

OLS with Experimental Data

Motivating Example

OLS with Experimental Data

Motivating Example

OLS with Experimental Data

We believe that feeding the bears will prevent them from looting the camp sites.

OLS with Experimental Data

Bear	Type	Treat	Camp Sites Looted
1	Black	0	22
2	Brown	1	10
3	Black	0	27
4	Polar	0	26
5	Polar	1	13
6	Brown	1	14
7	Black	0	33
8	Black	1	16
9	Polar	1	10
10	Brown	0	25
11	Brown	0	26
12	Polar	1	12

OLS with Experimental Data

Estimator

$\hat{\tau} = ave(Y_i : T_i = 1) - ave(Y_i : T_i = 0)$

$\hat{\tau} = \frac{10+13+14+16+10+12}{6} - \frac{22+27+26+33+25+26}{6} = -14$

$\hat{\sigma}^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$

$\hat{\sigma}^2_T = \frac{1}{6-1}[(10 - 12.5)^2 + (13 - 12.5)^2 + (14 - 12.5)^2 + (16 - 12.5)^2 + (10 - 12.5)^2 + (12 - 12.5)^2] = 5.5$

$\hat{\sigma}^2_C = \frac{1}{6-1}[(22 - 26.5)^2 + (27 - 26.5)^2 + (26 - 26.5)^2 + (33 - 26.5)^2 + (25 - 26.5)^2 + (26 - 26.5)^2] = 13.1$

$\widehat{SE}(\hat{\tau}) = \sqrt{\frac{\hat{\sigma}^2_T}{m} + \frac{\hat{\sigma}^2_C}{n-m}} = \sqrt{\frac{5.5}{6} + \frac{13.1}{6}} \approx 1.76$

OLS with Experimental Data

```
    Welch Two Sample t-test

data:  data[data$Treat == 1, ]$CampSiteRaids and data[data$Treat == 0, ]$CampSiteRaids
t = -7.9515, df = 8.569, p-value = 3.088e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.013664  -9.986336
sample estimates:
mean of x mean of y 
     12.5      26.5 
```

OLS with Experimental Data

Assumptions

1. Treatment assignment was random
2. Each bear has exactly two potential outcomes

OLS with Experimental Data

Assumptions for Regression

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$
2. All independent and control variables are fixed (no measurement error)
3. There is no deterministic linear relationship between the X variables (no collinearity)
4. $E[\epsilon_i] = 0$ for all i
5. $\epsilon_i \sim N(0, \sigma^2)$ for all i

Assumptions 1-4 are necessary for $\hat{\beta}$ to be unbiased. Assumption 5 is required for the standard errors, p-values, and confidence intervals to be correct.

OLS with Experimental Data

Pop Quiz (continued)

Question: You run an experiment and find the estimated treatment effect $\hat{\tau}$. You then fit the data to the model

$$y = \alpha + \beta x + \epsilon$$

and estimate β using OLS. What is the relationship between $\hat{\beta}$ and $\hat{\tau}$?

OLS with Experimental Data

Important Relationship: $\hat{\beta} = \hat{\tau}$

$$\hat{\tau} = ave(Y_i : X_i = 1) - ave(Y_i : X_i = 0)$$

$$\hat{\tau} = \frac{\sum X_i Y_i}{\sum X_i} - \frac{\sum (1-X_i) Y_i}{\sum (1-X_i)}$$

$$\hat{\tau} = \frac{ave(XY)}{ave(X)} - \frac{ave(Y) - ave(XY)}{1 - ave(X)}$$

$$\hat{\tau} = \frac{ave(XY) - ave(X)ave(Y)}{ave(X)(1 - ave(X))}$$

$$\hat{\tau} = \frac{ave(XY) - ave(X)ave(Y)}{p(1-p)}$$

where p is the proportion of subjects assigned to the treatment group. So

$$\hat{\tau} = \frac{Cov(X,Y)}{Var(X)}$$

$$\hat{\tau} = \hat{\beta}$$

OLS with Experimental Data

```
> summary(lm(CampSiteRaids~Treat,data))

Call:
lm(formula = CampSiteRaids ~ Treat, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
 -4.50  -1.75  -0.50   0.75   6.50 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   26.500      1.245  21.285 1.17e-09 ***
Treat        -14.000      1.761  -7.951 1.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.05 on 10 degrees of freedom
Multiple R-squared:  0.8634,	Adjusted R-squared:  0.8498 
F-statistic: 63.23 on 1 and 10 DF,  p-value: 1.242e-05
```

OLS with Experimental Data

Proving the unbiasedness of $\hat{\beta}$ using the regression framework.

1. There are m treatment units and $n-m$ control units.
2. Fix $\alpha = \sum_{i=1}^{n} Y_{ic} = \bar{Y}_c$ and $\beta = \sum_{i=1}^{n} Y_{it} - \alpha = \bar{Y}_t - \alpha$.
3. For each unit, let $E_{i1} = Y_{ic} - \alpha$ and $E_{i2} = Y_{it} - (\alpha + \beta)$

(Note that $\bar{E}_1 = \bar{E}_2 = 0$)

4. Define n new units, where $x = 1$ for m of these units and $x = 0$ for the rest. The initial y-values for these units will be $\alpha + \beta$ and α.

5. Randomly take m draws from the E_{i2}'s without replacement and add them to the m points where $x = 1$. For the remaining i's, take the E_{i1} values and add them to the $n-m$ units at $x = 0$.

(Thus, we have redefined our procedure from sampling from (Y_{it}, Y_{ic}) to (E_{i1},E_{i2}))

OLS with Experimental Data

We want to show that $E[\hat{\beta}] = \beta$

1. $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon$

2. All independent and control variables are fixed (no measurement error)

3. There is no deterministic linear relationship between the X variables (no collinearity)

4. $E[\epsilon_i] = 0$ for all i

5. $\epsilon_i \sim N(0, \sigma^2)$ for all i

Assumptions 1-4 are necessary for $\hat{\beta}$ to be unbiased. Assumption 5 is required for the standard errors, p-values, and confidence intervals to be correct.

OLS with Experimental Data

Two Important Issues

1. The errors are not i.i.d. $N(0, \sigma^2)$, so the estimated standard error is not guaranteed to be accurate.

2. We haven't controlled for anything. When we do, we will lose this simple argument for unbiasedness.

OLS with Experimental Data

```
> summary(lm(CampSiteRaids~Treat,data))

Call:
lm(formula = CampSiteRaids ~ Treat, data = data)

Residuals:
   Min     1Q Median     3Q    Max 
 -4.50  -1.75  -0.50   0.75   6.50 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   26.500      1.245  21.285 1.17e-09 ***
Treat        -14.000      1.761  -7.951 1.24e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.05 on 10 degrees of freedom
Multiple R-squared:  0.8634,	Adjusted R-squared:  0.8498 
F-statistic: 63.23 on 1 and 10 DF,  p-value: 1.242e-05
```

OLS with Experimental Data

```
        Welch Two Sample t-test

data:  data[data$Treat == 1, ]$CampSiteRaids and data[data$Treat == 0, ]$CampSiteRaids
t = -7.9515, df = 8.569, p-value = 3.088e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.013664  -9.986336
sample estimates:
mean of x mean of y
     12.5      26.5
```

OLS with Experimental Data

The normal OLS standard error is the [2,2] element of the variance-covariance matrix

$\hat{\mathbf{cov}} = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$

where $\hat{\sigma}^2 = \frac{1}{n-p}\sum_i e_i^2$

To get the robust standard errors, we will use the

$\hat{\mathbf{cov}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$

where $\mathbf{\Sigma} = E[ee']$

There is some flexibility in estimating $\mathbf{\Sigma}$.

OLS with Experimental Data

The following code will give you the results using the default robust standard errors in R, which use the formula $E[e_i^2] = \frac{\hat{e}_i^2}{1-h_i}$, where h_i is the ith element of $X(\frac{1}{n}X'X)^{-1}X'$. (see Davidson and MacKinnon (1993))

```
> require("sandwich")
> require("lmtest")
> model$newse<-vcovHC(model)
> coeftest(model,model$newse)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  26.5000     1.6186 16.3718 1.504e-08 ***
Treat       -14.0000     1.9287 -7.2587 2.731e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OLS with Experimental Data

The robust standard error procedure implemented by default in Stata use the formula $E[e_i^2] = \frac{n}{n-p}\hat{e}_i^2$. This is a degrees of freedom correction. To get the results using this procedure, add the argument: type="HC1".

```
> model$newse<-vcovHC(model,type="HC1")
> coeftest(model,model$newse)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  26.5000     1.4776 17.9344 6.213e-09 ***
Treat       -14.0000     1.7607 -7.9515 1.242e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OLS with Experimental Data

Now say that we think that polar bears are more aggressive, so we add in a contra for polar bears.

```
> data$Polar=as.numeric(data$Type=="Polar")
>
> model=lm(CampSiteRaids~Treat+Polar,data)
> model$newse<-vcovHC(model,type="HC1")
> coeftest(model,model$newse)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  26.7143     1.6793 15.9083 6.763e-08 ***
Treat       -13.5714     1.8134 -7.4840 3.755e-05 ***
Polar        -1.2857     1.3933 -0.9228    0.3802
```

OLS with Experimental Data

Our estimator will be guaranteed to be unbiased if we control for the interaction term as well.

```
> model=lm(CampSiteRaids~Treat+Polar+Treat*Polar,data)
> model$newse<-vcovHC(model,type="HC1")
> coeftest(model,model$newse)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  26.6000     1.9779 13.4488 8.955e-07 ***
Treat       -13.2667     2.6501 -5.0061  0.001045 **
Polar        -0.6000     1.9779 -0.3034  0.769358
Treat:Polar  -1.0667     2.7930 -0.3819  0.712476
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

OLS with Experimental Data

Controlling for the interaction term makes this equivalent to post-stratification and reweighing.

```
> # Now look at just the treatment effect for polar bears
>
> model=lm(CampSiteRaids~Treat,data[data$Type=="Polar",])
> model$newse<-vcovHC(model,type="HC1")
> coeftest(model,model$newse)

t test of coefficients:

               Estimate  Std. Error     t value Pr(>|t|)
(Intercept)  2.6000e+01  6.2804e-16  4.1399e+16  < 2e-16 ***
Treat       -1.4333e+01  1.0184e+00 -1.4075e+01  0.00501 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fixed Effects

Basic Idea

Help eliminate omitted variable bias by focusing on the differences in outcome for specific subsets of units or changes in the outcome for individual units over time.

Fixed Effects

```
> data[order(data$Type),]
   Bear  Type Treat CampSiteRaids
1     1 Black     0            22
3     3 Black     0            27
7     7 Black     0            33
8     8 Black     1            16
2     2 Brown     1            10
6     6 Brown     1            14
10   10 Brown     0            25
11   11 Brown     0            26
4     4 Polar     0            26
5     5 Polar     1            13
9     9 Polar     1            10
12   12 Polar     1            12
```

Fixed Effects

Black Bear Camp Site Raid Mean: 24.5

Brown Bear Camp Site Raid Mean: 18.75

Polar Bear Camp Site Raid Mean: 15.25

Fixed Effects

Next Step: Replace each of the bears' outcomes with their deviations from the means on the previous slide.

```
> for(i in unique(data$Type)){
+
+ data[data$Type==i,]$CampSiteRaids=data[data$Type==i,]$CampSiteRaids-
mean(data[data$Type==i,]$CampSiteRaids)
+
+ }
```

Fixed Effects

```
> data[order(data$Type),]
   Bear  Type Treat CampSiteRaids
1     1 Black     0         -2.50
3     3 Black     0          2.50
7     7 Black     0          8.50
8     8 Black     1         -8.50
2     2 Brown     1         -8.75
6     6 Brown     1         -4.75
10   10 Brown     0          6.25
11   11 Brown     0          7.25
4     4 Polar     0         10.75
5     5 Polar     1         -2.25
9     9 Polar     1         -5.25
12   12 Polar     1         -3.25
```

Fixed Effects

Black Bear Treatment Mean: 0.25

Brown Bear Treatment Mean: 0.5

Polar Bear Treatment Mean: 0.75

Fixed Effects

Next Step: Replace each of the bears' treatment assignments with their deviations from the means of the treatment assignments on the previous slide.

```
> for(i in unique(data$Type)){
+
+ data[data$Type==i,]$Treat=data[data$Type==i,]$Treat-mean(data[data$Type==i,]
$Treat)
+
+ }
```

Fixed Effects

```
> data[order(data$Type),]
   Bear  Type Treat CampSiteRaids
1     1 Black -0.25         -2.50
3     3 Black -0.25          2.50
7     7 Black -0.25          8.50
8     8 Black  0.75         -8.50
2     2 Brown  0.50         -8.75
6     6 Brown  0.50         -4.75
10   10 Brown -0.50          6.25
11   11 Brown -0.50          7.25
4     4 Polar -0.75         10.75
5     5 Polar  0.25         -2.25
9     9 Polar  0.25         -5.25
12   12 Polar  0.25         -3.25
```

Fixed Effects

```
> model=lm(CampSiteRaids~Treat,data)
> model$newse<-vcovHC(model)
> coeftest(model,model$newse)

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  3.1489e-16  8.4043e-01  0.0000         1
Treat       -1.3100e+01  1.3577e+00 -9.6488 2.204e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fixed Effects

```
> model=lm(CampSiteRaids~Treat+Type,data.original)
> model$newse<-vcovHC(model)
> coeftest(model,model$newse)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  27.7750     3.0137  9.2163 1.556e-05 ***
Treat       -13.1000     1.7123 -7.6504 6.015e-05 ***
TypeBrown    -2.4750     2.7357 -0.9047    0.3921
TypePolar    -2.7000     2.5998 -1.0385    0.3294
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fixed Effects

```
> summary(lm(CampSiteRaids~Treat,data))

Call:
lm(formula = CampSiteRaids ~ Treat, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-5.7750 -1.0750  0.3625  1.1000  5.2250 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.149e-16  8.035e-01   0.000        1    
Treat       -1.310e+01  1.760e+00  -7.442 2.21e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.783 on 10 degrees of freedom
Multiple R-squared:  0.847,	Adjusted R-squared:  0.8317 
F-statistic: 55.38 on 1 and 10 DF,  p-value: 2.207e-05
```

Fixed Effects

```
> summary(lm(CampSiteRaids~Treat+Type,data.original))

Call:
lm(formula = CampSiteRaids ~ Treat + Type, data = data.original)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7750 -1.0750  0.3625  1.1000  5.2250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.775      1.632  17.020 1.44e-07 ***
Treat        -13.100      1.968  -6.656  0.00016 ***
TypeBrown     -2.475      2.255  -1.098  0.30430
TypePolar     -2.700      2.411  -1.120  0.29517
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.112 on 8 degrees of freedom
Multiple R-squared:  0.8862,    Adjusted R-squared:  0.8436
F-statistic: 20.77 on 3 and 8 DF,  p-value: 0.000393
```

Fixed Effects

So since we have 3 types of bears, the fixed effects procedure uses up 3-1=2 degrees of freedom.

We have to tell R that we are only have 8 degrees of freedom instead of 10.

Fixed Effects

As mentioned before, if you enter the argument: type="HC1", you get the default function from Stata, which adjusts for the degrees of freedom.

```
> omega=function(residuals, diaghat, df){return(n/(n-p)*residuals^2)}
```

So we want to fix this function for our bear example.

```
> omega=function(residuals, diaghat, df){return(12/(12-4)*residuals^2)}
```

Fixed Effects

```
> model=lm(CampSiteRaids~Treat,data)
> model$newse<-vcovHC(model,omega=omega)
> coeftest(model,model$newse,df=8)

t test of coefficients:

              Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  3.1489e-16  8.9835e-01  0.0000         1
Treat       -1.3100e+01  1.3657e+00 -9.5921 1.157e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fixed Effects

```
> model=lm(CampSiteRaids~Treat+Type,data.original)
> model$newse<-vcovHC(model,omega=omega)
> coeftest(model,model$newse,df=8)

t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  27.7750     2.6596 10.4435 6.134e-06 ***
Treat       -13.1000     1.3657 -9.5921 1.157e-05 ***
TypeBrown    -2.4750     2.3462 -1.0549    0.3223    
TypePolar    -2.7000     2.1746 -1.2416    0.2495    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fixed Effects

To get the estimator that is guaranteed to be unbiased, control for the interactions.

```
> data.original$Brown=as.numeric(data.original$Type=="Brown")
>
> omega=function(residuals, diaghat, df){return(12/(12-6)*residuals^2)}
>
> model=lm(CampSiteRaids~Treat+Type+Polar*Treat+Brown*Treat,data.original)
> model$newse<-vcovHC(model,omega=omega)
> coeftest(model,model$newse,df=6)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)  27.3333     3.6717  7.4443 0.0003027 ***
Treat       -11.3333     3.6717 -3.0867 0.0214793 *  
TypeBrown    -1.8333     3.7056 -0.4947 0.6383711    
TypePolar    -1.3333     3.6717 -0.3631 0.7289534    
Treat:Polar  -3.0000     3.8103 -0.7873 0.4610455    
Treat:Brown  -2.1667     4.2109 -0.5145 0.6252711    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Voting Data

TABLE 3. OLS Regression Estimates of the Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Model Specifications		
	(a)	(b)	(c)
Civic Duty Treatment (Robust cluster standard errors)	.018* (.003)	.018* (.003)	.018* (.003)
Hawthorne Treatment (Robust cluster standard errors)	.026* (.003)	.026* (.003)	.025* (.003)
Self-Treatment (Robust cluster standard errors)	.049* (.003)	.049* (.003)	.048* (.003)
Neighbors Treatment (Robust cluster standard errors)	.081* (.003)	.082* (.003)	.081* (.003)
N of individuals	344,084	344,084	344,084
Covariates**	No	No	Yes
Block-level fixed effects	No	Yes	Yes

Note: Blocks refer to clusters of neighboring voters within which random assignment occurred. Robust cluster standard errors account for the clustering of individuals within household, which was the unit of random assignment.

* $p < .001$.

** Covariates are dummy variables for voting in general elections in November 2002 and 2000, primary elections in August 2004, 2002, and 2000.

Also try running the test with all the interactions between the treatments and the covariates.