Hypothesis testing: Social Pressure on Voter Turnout (replication)

Brianna Hayes

2023-12-16

> Alan S. Gerber, Donald P. Green, and Christopher W. Larimer (2008) "Social pressure and voter turnout: Evidence from a large-scale field experiment." *American Political Science Review*, vol. 102, no. 1, pp. 33–48

```
setwd('/Users/briannahayes/Documents/R')
social <- read.csv('social.csv', stringsAsFactors=T)
social <- subset(social, social$messages=='Neighbors' | social$messages=='Control')
social$messages <- str_replace(social$messages, 'Neighbors', 'Mailer')
```

1. **Hypotheses**

Null hypothesis: Social pressure through mailers had no affect on voter turnout in the 2006 primary election.

Alternative hypothesis: Social pressure through mailers increased voter turnout in the 2006 primary election.

Independent variable: messages, whether the voter was encouraged to vote through mailers (Control, Mailer)

Dependent variable: primary2006, whether the voter turned out in the 2006 primary election (1=voted, 0=abstained)

2. **Difference in means between the control and test group**

```
## Mean voter turnout in the control group
cont_avg <- mean(social$primary2006[social$messages=='Control'])

## Mean voter turnout in the test group
mail_avg <- mean(social$primary2006[social$messages=='Mailer'])

## Difference in means
dim <- mail_avg-cont_avg
dim
```

```
## [1] 0.08130991
```

3. **Standard error of difference in means between the test and control group, reported with the 95% confidence interval of the estimate**

```r
## Standard error of difference in means
dim_se <- sqrt((mail_avg*(1-mail_avg))/
                 length(social$primary2006[social$messages=='Mailer']) +
               (cont_avg*(1-cont_avg))/
                 length(social$primary2006[social$messages=='Control']))
dim_se
```

```
## [1] 0.002691722
```

```r
## Confidence interval
ci <- 1.96*dim_se
ci
```

```
## [1] 0.005275775
```

```r
## High and low bounds of the confidence interval
ci_low <- dim-ci
ci_high <- dim+ci

ci_low
```

```
## [1] 0.07603414
```

```r
ci_high
```

```
## [1] 0.08658569
```

The difference in means standard error is calculated by finding the square root of the sum of squared errors between the treatment and control groups. Here, that value was calculated to be approximately .003. The confidence interval bounds under a 95% confidence are calculated as the difference in means plus/minus the difference in means standard error times 1.96 (approximately how many standard deviations include 95% of values). The confidence interval was found to be approximately .005. With a difference in means of .081, this makes the lower bound .076 and upper bound .087.

4. **p-value of the observed difference in means where $\alpha = 0.05$, and the implications it has for the null hypothesis**

```r
## Calculating the z-score
zscore <- (dim-0)/dim_se

## Calculating the two-sided p-value under the standard normal
pval <- 2*pnorm(zscore, lower.tail=F)
pval
```

```
## [1] 1.893805e-200
```

This p-value indicates that we can reject the null hypothesis. Assuming alpha is .05, less than 5% of values must be as or more extreme than our observed value to be able to reject the null hypothesis. With a p-value of 1.894 e-200, we are most certainly able to reject the null hypothesis and are inclined to accept the alternative. Thus, we are inclined to believe that social pressure does have an effect on voter turnout, as opposed to no effect.

Is the Effectiveness of the Experiment Confounded by Gender?

1. **Hypotheses**

Null hypothesis: There is no difference in the effectiveness of the experiment based on gender.

Alternative hypothesis: There is a difference in the effectiveness of the experiment based on gender.

2. **Using regression to calculate a point estimate**

```
fit1 <- lm(primary2006 ~ sex + messages + sex:messages, data=social)
fit1
```

```
##
## Call:
## lm(formula = primary2006 ~ sex + messages + sex:messages, data = social)
##
## Coefficients:
##            (Intercept)                  sexmale          messagesMailer
##              0.2904558               0.0123389               0.0808995
## sexmale:messagesMailer
##              0.0008487
```

The point estimate we are interested in is the sexmale:messagesNeighbors term. The value of the coefficient represents the difference in estimated proportion of voter turnout between men and women who received social pressure. Thus, our regression estimates that men who received social pressure had a voter turnout proportion of about .001 greater than women.

3. **Constructing a confidence interval on the difference of effectiveness between men and women**

```
## Regression summary accessing standard error of the coefficients
sum_fit1 <- summary(fit1)
sum_fit1
```

```
##
## Call:
## lm(formula = primary2006 ~ sex + messages + sex:messages, data = social)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3845 -0.3028 -0.2905  0.6286  0.7095
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             0.2904558  0.0014941 194.399  < 2e-16 ***
## sexmale                 0.0123389  0.0021108   5.846 5.05e-09 ***
## messagesMailer          0.0808995  0.0036583  22.114  < 2e-16 ***
## sexmale:messagesMailer 0.0008487  0.0051730   0.164     0.87
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4615 on 229440 degrees of freedom
## Multiple R-squared:  0.00447,    Adjusted R-squared:  0.004457
## F-statistic: 343.4 on 3 and 229440 DF,  p-value: < 2.2e-16
```

```
## Confidence interval
ci2 <- 1.96*sum_fit1$coefficients[8]
ci2
```

```
## [1] 0.01013908
```

```
## High and low bounds of the confidence interval
ci2_low <- sum_fit1$coefficients[4] - ci2
ci2_high <- sum_fit1$coefficients[4] + ci2

ci2_low
```

```
## [1] -0.00929041
```

```
ci2_high
```

```
## [1] 0.01098775
```

Again, the term of interest to estimate the difference in effectiveness between men and women is sex-male:messagesNeighbors. The standard error is calculated through the regression, so we can simply pull it from the regression summary. The standard error of the term is approximately .005, which we will use to calculate the confidence interval. The confidence interval bounds under a 95% confidence are calculated as the point estimate plus/minus the standard error times 1.96. The confidence interval was found to be approximately .01, making the lower bound approximately -.009 and upper bound .011.

4. **p-value of the observed difference in means where $\alpha = 0.05$, and the implications it has for the null hypothesis**

```
## Calculating the z-score
zscore2 <- (sum_fit1$coefficients[4]-0)/sum_fit1$coefficients[8]

## Calculating the two-sided p-value under the standard normal
pval2 <- 2*pnorm(zscore2, 0, 1, lower.tail=F)
pval2
```

```
## [1] 0.8696855
```

This p-value indicates that we cannot reject the null hypothesis. Assuming alpha is .05, less than 5% of values must be as or more extreme than our observed value of .008 to cause statistical significance. With a p-value of approximately .87, we certainly have more than 5% of values as or more extreme than the point estimate. Thus, we cannot reject the null hypothesis and are inclined to believe that sex is not a significant influence in voter turnout.

Considering Age as a Factor

1. **Plotting turnout by voter age in the control condition**

```r
## Creating a voter age variable in a new data frame with dplyr
df1 <- social %>%
  filter(messages=='Control') %>%
  group_by(yearofbirth) %>%
  summarize(avg_2006 = mean(primary2006)) %>%
  mutate(age = 2006 - yearofbirth)

## Plotting the relationship with ggplot
plt1 <- ggplot() +
  geom_point(aes(x=df1$age, y=df1$avg_2006)) +
  xlab('Voter age in 2006') +
  ylab('Avergae voter turnout') +
  ggtitle('Average voter turnout for those not recieving mailers, by age')
plt1
```

2. **Regressing voter turnout on age, interpreting the age coefficient as an observed value**

```r
## Creating a voter age variable in the original data set
social$age <- 2006 - social$yearofbirth

## Regression model
fit2 <- lm(primary2006 ~ age, data=social)
fit2
```

```
##
## Call:
## lm(formula = primary2006 ~ age, data = social)
##
## Coefficients:
## (Intercept)          age
```

```
##    0.105581     0.004107
```

When the 2006 voter turnout is run on age, the coefficient is calculated to be approximately .004. The intercept term serves as our Beta 0. The age coefficient serves as our Beta 1, or our observed value.

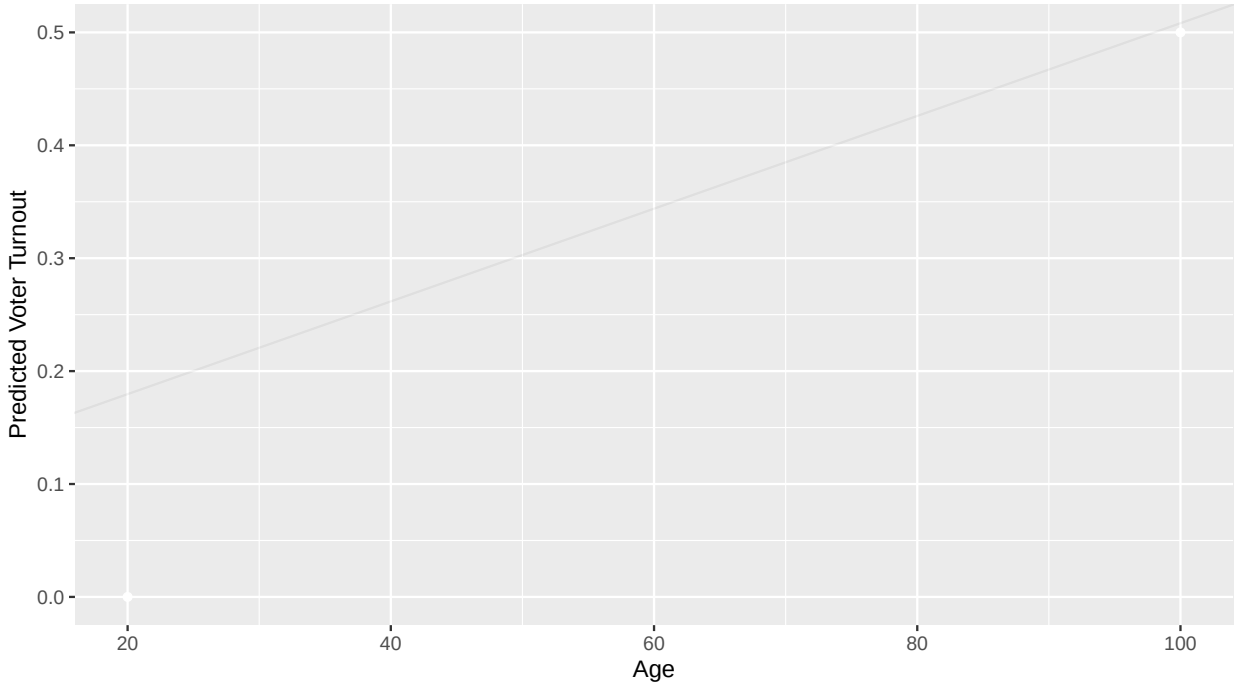
3. **Generating 100 bootstrap age coefficients**

```
coefs <- c()
for (i in 1:100) {
  row_nums <- sample(1:nrow(social), nrow(social), replace=T)
  bootstrap <- social[row_nums,]
  fit <- lm(primary2006 ~ age, data=bootstrap)
  coefs <- append(coefs, fit2$coefficients[2])
}
coefs
```

```
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666 0.004106666
##         age         age         age         age
## 0.004106666 0.004106666 0.004106666 0.004106666
```

4. **Plotting the bootstrap coefficients to evaluate chance error**

```
plt2 <- ggplot() +
  geom_point(aes(x = c(20, 100), y = c(0,.5)), col='white') +
  geom_abline(intercept = 0.097473257, slope = coefs, alpha=.05) +
  labs(x='Age', y = 'Predicted Voter Turnout')
plt2
```



Based on the bootstrap sample, it appears that the relationship observed between voter turnout and age is likely not due to chance sampling. If it were due to chance sampling, the estimated lines of best fit would not all have similar slopes, As we can see from the graph, the bootstrapped coefficients create regressions that well reflect our observed slope value. One of the assumptions from ordinary least squares is that unbiased Beta 1 estimators will be centered and close togehter, which is what can be deduced from the plot.