Handout 7

Student von Student III

In this handout, we will learn how to test difference-in-means using the function `lm`. We will include some more practice using loops.

Topics and Concepts Covered

- Fitting a linear model to estimate a difference-in-means
- Interpreting the coefficients
- Assessing statistical significance
- Running a loop with the command `for`
- Estimating a difference-in-means using `lm(y ~ x)`
- Using `summary` to return the output from `lm`
- Extracting coefficients, standard errors, t-values, and p-values from `lm`
- Using `rowMeans` and `colMeans` to return the means of the rows or columns of a matrix
- Creating a PDF figure using `pdf`
- Running a file with the command `source`
- Loading an R data file (ending in `.Rdata`) with `load`

Before beginning this handout, Do not forget to make a new folder for this assignment and set your working directory!

Using a Linear Model to Test Differences-in-Means

Gartner (2008) explored the relationship between gender and support for the Iraq War. We can state this as a two-sample problem: what is the difference in support for the Iraq War between males and females? Of course, there is going to be *some* difference between these two means, so what we really want to know is whether the effect is *statistically significant*.

Statistical significance means, informally, that we can differentiate the observed difference from chance error. A bit more formally, an effect is statistically significant at the 5% level if the probability of observing the observed difference in means if, in truth, there were no effect is 5% or less. While the logic may seem, at first, a bit serpentine, a few examples can help illustrate how the concepts are used in practice.

We will use the command `lm` to estimate the difference-in-means between two groups. `lm` stands for "linear model" which we will be discussing quite a bit. The linear model allows a means to test the difference-in-means between two groups, but you will see that it can also be used for multivariate regression.

Recall that a linear regression is a model for the outcome Y, of the form

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

In this model, we are assuming that our observed value Y_i is equal to some expected value that depends on the value of X_i plus chance error (ϵ_i). So, the expected value of Y_i if $X_i = 0$ is β_0 and the expected value of Y_i if $X_i = 1$ is $\beta_0 + \beta_1$. Therefore, the effect of X_i is β_1.

The linear model returns fitted values of the form where $\widehat{\beta}_1$ is the estimate of the effect of X_i on Y_i. In other words, $\widehat{\beta}_1$ is the estimate of the difference-in-means for Y_i between the groups $X_i = 1$ and $X_i = 0$

In R, we use the `lm` function to fit a linear model. The `lm` function takes a formula of the form `Y ~ X`, where Y is the outcome variable (dependent variable) and X is the treatment (independent variable). Let's start with loading in the data. The data contains the variable `female` which is a 1 if the respondent is female, and

a 0 otherwise, and a variable *mistake* which takes a value of 1 if the respondent felt that the Iraq War was a mistake, and a 0 otherwise.

First, load and summarize the data:

```
load("data/iraq.RData")
head(iraq)
```

```
  female mistake
1      1       0
2      1       0
3      0       0
4      0       0
5      0       1
6      1       0
```

```
summary(iraq)
```

```
     female          mistake
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000
 Median :1.0000   Median :0.0000
 Mean   :0.5711   Mean   :0.1685
 3rd Qu.:1.0000   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :1.0000
```

To assess the difference-in-means, we are going to fit a linear model, and look at it:

```
lm1 <- lm(iraq$mistake ~ iraq$female)
lm1
```

```

Call:
lm(formula = iraq$mistake ~ iraq$female)

Coefficients:
(Intercept)  iraq$female
    0.19388     -0.04445
```

```
mean(iraq$mistake[iraq$female == 0]) # Intercept
```

```
[1] 0.1938776
```

```
mean(iraq$mistake[iraq$female == 1]) -
  mean(iraq$mistake[iraq$female == 0]) # Coefficient
```

```
[1] -0.04445226
```

We can interpret the first coefficient as the mean for males (`female == 0`), and the second coefficient as the difference-in-means between females and males. In other words, women were approximately 4.44 percentage less likely to think the Iraq War was a mistake.

Is the effect statistically significant? Let's use `summary` to return the details from the `lm` object:

```
summary(lm1)
```

```

Call:
lm(formula = iraq$mistake ~ iraq$female)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.1939 -0.1939 -0.1494 -0.1494  0.8506

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.19388    0.02675   7.248 1.82e-12 ***
iraq$female -0.04445    0.03539  -1.256     0.21
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3745 on 455 degrees of freedom
Multiple R-squared:  0.003455,  Adjusted R-squared:  0.001264
F-statistic: 1.577 on 1 and 455 DF,  p-value: 0.2098
```

We see that it is not. The p-value, in the last column, is approximately 0.21, which is not below the standard 5% threshold. In other words, we cannot reject the null hypothesis that the difference between males and females is zero.

`coef` allows you to access to the coefficients, standard errors, t-statistics, and p-values of the regression:

```
summary(lm1)$coef
```

```
               Estimate Std. Error   t value     Pr(>|t|)
(Intercept)  0.19387755 0.02674816  7.248259 1.823736e-12
iraq$female -0.04445226 0.03539417 -1.255921 2.097893e-01
```

Then, if we wanted the coefficient and the p-value for the effect due to being female, we would use

```
# One coefficient
summary(lm1)$coef[2,1]
```

```
[1] -0.04445226
```

```
# p-value
summary(lm1)$coef[2,4]
```

```
[1] 0.2097893
```

```
# 95% Confidence interval
mean.iraq <- summary(lm1)$coef[2,1] # Difference in means
se.iraq <- summary(lm1)$coef[2,2] # Standard error
c(mean.iraq - 1.96 * se.iraq, mean.iraq + 1.96 * se.iraq)
```

```
[1] -0.1138248  0.0249203
```

A Second Example: Growth During Wartime

We turn next to the relationship between civil war and GDP. A broad range of literature in International Relations and Comparative Politics explores the relationship between civil war and economic growth. The literature has argued that the incidence of civil war is often associated with low levels of development and economic growth.

We rely on the `growth.RData` data set. The data set contains the following variables:

- `year` - year of observation
- `country` - country
- `war` - binary indicator of civil war (1 if the country experienced civil war in the observed year, 0 else)

- `gdppc` - GDP per capita for the year of observation
- `gdppc.lag` - GDP per capita from the previous year
- `growth.rate` - GDP per capita growth rate, in percentage points

First, we read in the data.

```
load("data/growth.RData")
head(growth)
```

```
  year     country war gdppc gdppc.lag growth.rate
2 1961 AFGHANISTAN   0 0.477     0.454        5.07
3 1962 AFGHANISTAN   0 0.504     0.477        5.66
4 1963 AFGHANISTAN   0 0.551     0.504        9.33
5 1964 AFGHANISTAN   0 0.562     0.551        2.00
6 1965 AFGHANISTAN   0 0.586     0.562        4.27
7 1966 AFGHANISTAN   0 0.620     0.586        5.80
```

```
summary(growth)
```

```
      year           country          war             gdppc        
 Min.   :1960   ARGENTINA:  39   Min.   :0.0000   Min.   : 0.133  
 1st Qu.:1971   AUSTRALIA:  39   1st Qu.:0.0000   1st Qu.: 0.978  
 Median :1981   AUSTRIA  :  39   Median :0.0000   Median : 2.214  
 Mean   :1981   BELGIUM  :  39   Mean   :0.1498   Mean   : 3.966  
 3rd Qu.:1990   BENIN    :  39   3rd Qu.:0.0000   3rd Qu.: 5.112  
 Max.   :1999   BOLIVIA  :  39   Max.   :1.0000   Max.   :51.989  
                (Other)  :4820                                    
   gdppc.lag       growth.rate      
 Min.   : 0.133   Min.   :-60.550  
 1st Qu.: 0.968   1st Qu.: -0.820  
 Median : 2.170   Median :  2.110  
 Mean   : 3.910   Mean   :  1.849  
 3rd Qu.: 4.988   3rd Qu.:  4.820  
 Max.   :53.901   Max.   :119.590  
```

To begin, we wish to consider the association between civil war and levels of economic development. As such, we wish to compare the GDP per capita between countries that experienced a war versus those that did not. We would expect those countries that experience war to have levels of economic development (i.e. measured as GDP per capita) that are significantly lower than countries that do not experience war. Due to the skew in the data, we must log GDP per capita.

```
growth$log.growth <- log(growth$gdppc) # Log data
lm2 <- lm(growth$log.growth ~ growth$war) # Estimate diff in means
summary(lm2)
```

```

Call:
lm(formula = growth$log.growth ~ growth$war)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.95805 -0.82910 -0.01321  0.79311  3.01039 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.94064    0.01556   60.45   <2e-16 ***
```

```
growth$war  -0.67658    0.04020  -16.83   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.02 on 5052 degrees of freedom
Multiple R-squared:  0.05308,   Adjusted R-squared:  0.0529
F-statistic: 283.2 on 1 and 5052 DF,  p-value: < 2.2e-16
```

We observe that growth rates are lower by about -0.67 in countries at war versus countries not at war. The result is statistically significant, since the p-value on the difference-in-means is much lower than 0.05. Therefore, we can reject the null hypothesis that the difference in growth between countries at war and not at war is zero.

That said, is the effect causal? As you may have guessed, no. There are any number of other possible confounders and omitted variables. We will start discussing how to handle them in the next handout.

Problems

Recall Gerber, Green, and Larimer (2008), the social pressure experiment which we discussed in class. In this experiment, researchers administered sent one of four different postcards to Michigan voters in 2006, as well as maintaining a control group that received no such postcard. The outcome was whether the respondent voted or not. We will be considering the treatment of whether the individual received a postcard alerting her to her neighbor's voting behavior and letting her know that her neighbors will be made aware of her behavior (1) versus the control condition of receiving no postcard.

In the `data` folder, you will find the file `social.txt`.

Question 1

Was the treatment effect positive or negative? Was it statistically significant?

Question 2

Construct a 95% and confidence interval. What is the confidence interval telling us?

Question 3

Can we interpret the effect in a causal fashion? Why or why not?

Question 4

(Practice with a Loop, do what you can with this one, and then we will discuss it in precept)

We're going to continue with some data from the New Haven experiment (Gerber and Green 2000) which we discussed in the last handout. You will find in the data folder a file called `GerberGreenSubset.txt`, which contains the following variables:

Name	Description
`voted98`	1 if they voted in the 1998 election, 0 otherwise (dependent variable)
`mailgrp`	1 if they were sent mail encouraging them to vote; 0 otherwise
`ward`	Voting ward of household (2, 3, . . . , 30)

Name	Description
`appeal`	Type of appeal (1, 2, 3)
`mail`	Number of mailings sent (0,1,2,3)

You notice that mailings have no significant effect on turnout. Yet, pretend you are a less-than-scrupulous purveyor of campaign mailings. You get your hand on Gerber and Green's data, and do the following:

For every possible combination of number of mailings (1, 2, 3), appeal (1, 2, 3), and ward (2, 3,..., 30) calculate the treatment effect of receiving the mailings and its p-value. (Hint: this is a triple loop.)

What percentage of your p-values are below 0.05?

What percentage would you expect by chance?

Being less-than-scrupulous, you assert that you have found significant results - lots of them! Why should your customers be wary of the claim that mailings have an impact?

Explain why this process is termed 'data-fishing' and why it is frowned upon.

References

Gartner, Scott S. 2008. "The Multiple Effects of Casualties on Public Support for War: An Experimental Approach." *American Political Science Review* 102 (01): 95–106. doi:10.1017/S0003055408080027.

Gerber, Alan S., and Donald P. Green. 2000. "The Effects of Canvassing, Telephone Calls, and Direct Mail on Voter Turnout: A Field Experiment." *American Political Science Review* 94 (03): 653–63. doi:10.2307/2585837.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. "Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment." *American Political Science Review* 102 (01): 33–48. doi:10.1017/S000305540808009X.