Addressing Alternative Explanations: Multiple Regression

17.871
Spring 2012
Did Clinton hurt Gore example

- Did Clinton hurt Gore in the 2000 election?
  - Treatment is not liking Bill Clinton
Bivariate regression of Gore thermometer on Clinton thermometer
Did Clinton hurt Gore example

- What alternative explanations would you need to address?
- Nonrandom selection into the treatment group (disliking Clinton) from many sources
- Let’s address one source: party identification
- How could we do this?
  - Matching: compare Democrats who like or don’t like Clinton; do the same for Republicans and independents
  - Multivariate regression: control for partisanship statistically
    - Also called multiple regression, Ordinary Least Squares (OLS)
    - Presentation below is intuitive
Democratic picture

Gore thermometer

Clinton thermometer
Independent picture

Gore thermometer vs. Clinton thermometer
Republican picture

Gore thermometer vs. Clinton thermometer
Combined data picture

Gore thermometer

Clinton thermometer
Combined data picture with regression: bias!
Combined data picture with “true” regression lines overlaid
Tempting yet wrong normalizations

Subtract the Gore therm. from the avg. Gore therm. score

Subtract the Clinton therm. from the avg. Clinton therm. score
3D Relationship
3D Linear Relationship
3D Relationship: Clinton
3D Relationship: party
The Linear Relationship between Three Variables

\[ Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \epsilon_i \]
The method of least squares
(again)

Pick $\beta_0$, $\beta_1$, and $\beta_2$ to minimize

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \text{ or }$$

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i - \beta_2 X_2)^2$$
The Slope Coefficients

\[ \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y - Y_i)(\bar{X}_1 - X_{1,i})}{\sum_{i=1}^{n} (\bar{X}_1 - X_{1,i})^2} - \hat{\beta}_2 \]

\[ \hat{\beta}_2 = \frac{\sum_{i=1}^{n} (Y - Y_i)(\bar{X}_2 - X_{1,i})}{\sum_{i=1}^{n} (\bar{X}_2 - X_{2,i})^2} - \hat{\beta}_1 \]

and

\[ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}_1 - \hat{\beta}_2 \bar{X}_2 \]

X\(_1\) is Clinton thermometer, X\(_2\) is PID, and Y is Gore thermometer
The Slope Coefficients More Simply

\[ \hat{\beta}_1 = \frac{\text{cov}(X_1, Y)}{\text{var}(X_1)} - \hat{\beta}_2 \frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)} \text{ and } \]

\[ \hat{\beta}_2 = \frac{\text{cov}(X_2, Y)}{\text{var}(X_2)} - \hat{\beta}_1 \frac{\text{cov}(X_1, X_2)}{\text{var}(X_2)} \]

X_1 is Clinton thermometer, X_2 is PID, and Y is Gore thermometer
The Matrix form

\[
\begin{array}{c|cccc}
  & 1 & x_{1,1} & x_{2,1} & \ldots & x_{k,1} \\
\hline
y_1 & 1 & x_{1,2} & x_{2,2} & \ldots & x_{k,2} \\
\hline
\vdots & 1 & \vdots & \vdots & \ldots & \vdots \\
\hline
y_n & 1 & x_{1,n} & x_{2,n} & \ldots & x_{k,n} \\
\end{array}
\]

\[
\beta = \left( X'X \right)^{-1} X'y
\]
Multivariate slope coefficients

Bivariate estimate:

\[ \hat{\beta}_1^B = \frac{\text{cov}(X_1, Y)}{\text{var}(X_1)} \] vs.

Multivariate estimate:

\[ \hat{\beta}_1^M = \frac{\text{cov}(X_1, Y)}{\text{var}(X_1)} - \hat{\beta}_2^M \frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)} \]

When does \( \hat{\beta}_1^B = \hat{\beta}_1^M \)? Obviously, when \( \hat{\beta}_2^M \frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)} = 0 \)

\( X_1 \) is Clinton thermometer, \( X_2 \) is PID, and \( Y \) is Gore thermometer

Clinton effect (on Gore) in bivariate \((B)\) regression

Are Gore and Party ID related?

Clinton effect (on Gore) in multivariate \((M)\) regression

Are Clinton and Party ID related?
## The Output

```
. reg gore clinton party3

Source | SS      | df | MS
-------------+---------+-----+---------
Model       | 629261.91| 2   | 314630.955
Residual    | 522964.934| 1742| 300.209492
-------------+---------+-----+---------
Total       | 1152226.84| 1744| 660.68053

Number of obs = 1745
F( 2, 1742) = 1048.04
Prob > F = 0.0000
R-squared = 0.5461
Adj R-squared = 0.5456
Root MSE = 17.327

|          | Coef.  | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------------|--------|-----------|-------|-----|---------------------|
| gore             |        |           |       |     |                     |
| clinton          | .5122875| .0175952 | 29.12 | 0.000 | .4777776 , .5467975 |
| party3           | 5.770523| .5594846 | 10.31 | 0.000 | 4.673191 , 6.867856 |
| _cons            | 28.6299 | 1.025472 | 27.92 | 0.000 | 26.61862 , 30.64119 |
```

**Interpretation of clinton effect:** *Holding constant party identification*, a one-point increase in the Clinton feeling thermometer is associated with a .51 increase in the Gore thermometer.
Separate regressions

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>23.1</td>
<td>55.9</td>
<td>28.6</td>
</tr>
<tr>
<td>Clinton</td>
<td>0.62</td>
<td>--</td>
<td>0.51</td>
</tr>
<tr>
<td>Party</td>
<td>--</td>
<td>15.7</td>
<td>5.8</td>
</tr>
</tbody>
</table>

\[
\hat{\beta}_1 = \frac{\text{cov}(X_1, Y)}{\text{var}(X_1)} - \hat{\beta}_2 \frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)} \quad \text{and} \\
\hat{\beta}_2 = \frac{\text{cov}(X_2, Y)}{\text{var}(X_2)} - \hat{\beta}_1 \frac{\text{cov}(X_1, X_2)}{\text{var}(X_2)}
\]
Why did the Clinton Coefficient change from 0.62 to 0.51

```
. corr gore clinton party, cov
(obs=1745)

<table>
<thead>
<tr>
<th></th>
<th>gore</th>
<th>clinton</th>
<th>party3</th>
</tr>
</thead>
<tbody>
<tr>
<td>gore</td>
<td>660.681</td>
<td>549.993</td>
<td>883.182</td>
</tr>
<tr>
<td>clinton</td>
<td>13.7008</td>
<td>16.905</td>
<td>.8735</td>
</tr>
<tr>
<td>party3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```
The Calculations

\[ \hat{\beta}_1^B = \frac{\text{cov}(gore, clinton)}{\text{var}(clinton)} = \frac{549.993}{883.182} = 0.6227 \]

\[ \hat{\beta}_1^M = \frac{\text{cov}(gore, clinton)}{\text{var}(clinton)} - \hat{\beta}_2^M \frac{\text{cov}(clinton, party)}{\text{var}(clinton)} \]

\[ = \frac{549.993}{883.182} - 5.7705 \frac{16.905}{883.182} \]

\[ = 0.6227 - 0.1105 \]

\[ = 0.5122 \]

```
corr gore clinton party, cov
(obs=1745)

<table>
<thead>
<tr>
<th></th>
<th>gore</th>
<th>clinton</th>
<th>party3</th>
</tr>
</thead>
<tbody>
<tr>
<td>gore</td>
<td>660.681</td>
<td></td>
<td></td>
</tr>
<tr>
<td>clinton</td>
<td>549.993</td>
<td>883.182</td>
<td></td>
</tr>
<tr>
<td>party3</td>
<td>13.7008</td>
<td>16.905</td>
<td>.8735</td>
</tr>
</tbody>
</table>
```
Another way of thinking about this

Rewrite

\[ \hat{\beta}_1^M = \frac{\text{cov}(\text{gore, clinton})}{\text{var}(\text{clinton})} - \hat{\beta}_2^M \frac{\text{cov}(\text{clinton, party})}{\text{var}(\text{clinton})} \]

as

\[ \frac{\text{cov}(\text{gore, clinton})}{\text{var}(\text{clinton})} = \hat{\beta}_1^M + \hat{\beta}_2^M \frac{\text{cov}(\text{clinton, party})}{\text{var}(\text{clinton})} \]

Total effect = Direct effect + indirect effect

The Total Effect of the Clinton thermometer on the Gore thermometer (.61) can be Broken down into a direct effect of .51, plus an indirect effect (though party) of .11
Drinking and Greek Life Example

- Why is there a correlation between living in a fraternity/sorority house and drinking?
  - Greek organizations often emphasize social gatherings that have alcohol. The effect is being in the Greek organization itself, not the house.
  - There’s something about the House environment itself.
Dependent variable: Times Drinking in Past 30 Days

C8. When did you last have a drink (that is more than just a few sips)?
- I have never had a drink → Skip to C22 (page 10)
- Not in the past year → Skip to C22 (page 10)
- More than 30 days ago, but in the past year → Skip to C17 (page 8)
- More than a week ago, but in the past 30 days → Go to C9
- Within the last week → Go to C9

C9. On how many occasions have you had a drink of alcohol in the past 30 days? (Choose one answer.)
- Did not drink in the last 30 days
- 1 to 2 occasions
- 3 to 5 occasions
- 6 to 9 occasions
- 10 to 19 occasions
- 20 to 39 occasions
- 40 or more occasions


© Harvard School of Public Health. All rights reserved. This content is excluded from our Creative Commons license. For more information, see [http://ocw.mit.edu/fairuse](http://ocw.mit.edu/fairuse).
. infix age 10-11 residence 16 greek 24 screen 102
timespast30 103 howmuchpast30 104 gpa 278-279 studying 281
timeshs 325 howmuchhs 326 socializing 283 stwgt_99 475-493
weight99 494-512 using da3818.dat,clear
(14138 observations read)

. recode timespast30 timeshs (1=0) (2=1.5) (3=4) (4=7.5)
(5=14.5) (6=29.5) (7=45)
(timespast30: 6571 changes made)
(timeshs: 10272 changes made)

. replace timespast30=0 if screen<=3
(4631 real changes made)
. tab timespast30

<table>
<thead>
<tr>
<th>timespast30</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>4,652</td>
<td>33.37</td>
<td>33.37</td>
</tr>
<tr>
<td>1.5</td>
<td>2,737</td>
<td>19.64</td>
<td>53.01</td>
</tr>
<tr>
<td>4</td>
<td>2,653</td>
<td>19.03</td>
<td>72.04</td>
</tr>
<tr>
<td>7.5</td>
<td>1,854</td>
<td>13.30</td>
<td>85.34</td>
</tr>
<tr>
<td>14.5</td>
<td>1,648</td>
<td>11.82</td>
<td>97.17</td>
</tr>
<tr>
<td>29.5</td>
<td>350</td>
<td>2.51</td>
<td>99.68</td>
</tr>
<tr>
<td>45</td>
<td>45</td>
<td>0.32</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 13,939 | 100.00 |
Key explanatory variables

- Live in fraternity/sorority house
  - Indicator variable (dummy variable)
  - Coded 1 if live in, 0 otherwise

- Member of fraternity/sorority
  - Indicator variable (dummy variable)
  - Coded 1 if member, 0 otherwise
Three Regressions

<table>
<thead>
<tr>
<th>Dependent variable: number of times drinking in past 30 days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Live in frat/sor house (indicator variable)</td>
</tr>
<tr>
<td>4.44 (0.35)</td>
</tr>
<tr>
<td>Member of frat/sor (indicator variable)</td>
</tr>
<tr>
<td>--- (---)</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>4.54 (0.56)</td>
</tr>
<tr>
<td>S.E.R.</td>
</tr>
<tr>
<td>6.49 (---)</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>13,876</td>
</tr>
</tbody>
</table>

What is the substantive interpretation of the coefficients?

Note: Standard errors in parentheses. Corr. Between living in frat/sor house and being a member of a Greek organization is .42
The Picture

\[ \hat{\beta}_2^M = 2.26 \]

\[ \hat{\beta}_1^M = 2.44 \]

Living in frat house \( x_2 \)

\( \gamma_{21} = 0.19 \)

Member of fraternity \( x_1 \)

Drinks per 30 days \( y \)
Accounting for the total effect

\[ \hat{\beta}_1^B = \hat{\beta}_1^M + \hat{\beta}_2^M \gamma_{21} \]

Total effect = Direct effect + indirect effect

\[ \gamma_{21} = 0.19 \]

\[ \beta_2^M = 2.26 \]

\[ \beta_1^M = 2.44 \]
Accounting for the effects of frat house living and Greek membership on drinking

<table>
<thead>
<tr>
<th>Effect</th>
<th>Total</th>
<th>Direct</th>
<th>Indirect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Member of Greek org.</td>
<td>2.88</td>
<td>2.44</td>
<td>0.44</td>
</tr>
<tr>
<td>Live in frat/sor. house</td>
<td>4.44</td>
<td>2.26</td>
<td>2.18</td>
</tr>
</tbody>
</table>

From bivariate regressions

From multiple regressions

From accounting identity: T=D+I

<table>
<thead>
<tr>
<th>Effect</th>
<th>Total</th>
<th>Direct (%)</th>
<th>Indirect (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Member of Greek org.</td>
<td>2.88</td>
<td>85%</td>
<td>15%</td>
</tr>
<tr>
<td>Live in frat/sor. house</td>
<td>4.44</td>
<td>51%</td>
<td>49%</td>
</tr>
</tbody>
</table>
17.871 Political Science Laboratory
Spring 2012

For information about citing these materials or our Terms of Use, visit http://ocw.mit.edu/terms.