Addressing Alternative Explanations: Multiple Regression

17.871

Spring 2013
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
Democratic picture

Gore thermometer

Clinton thermometer
Independent picture

Gore thermometer vs. Clinton thermometer
Republican picture

Gore thermometer

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

\[ Y_i = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i \]
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

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

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

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

\(X_1\) is Clinton thermometer, \(X_2\) is PID, and \(Y\) is Gore thermometer
The Slope Coefficients More Simply

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

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

\[ X_1 \text{ is Clinton thermometer, } X_2 \text{ is PID, and } Y \text{ is Gore thermometer} \]
The Matrix form

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

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

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
The Output

. reg gore clinton party3

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 1745</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>629261.91</td>
<td>2</td>
<td>314630.955</td>
<td>F( 2, 1742) = 1048.04</td>
</tr>
<tr>
<td>Residual</td>
<td>522964.934</td>
<td>1742</td>
<td>300.209492</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>1152226.84</td>
<td>1744</td>
<td>660.68053</td>
<td>R-squared = 0.5461</td>
</tr>
</tbody>
</table>

|           | 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>

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

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

\[ \text{Total effect} = \text{Direct effect} + \text{indirect effect} \]

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
Graphical way of thinking about this

Total effect

$\hat{\beta}_1^B$

Clinton feeling thermometer $\rightarrow$ Gore feeling thermometer
Graphical way of thinking about this

Total effect

Clinton feeling thermometer $\rightarrow \hat{\beta}_1^B$ Gore feeling thermometer

Can be broken down into:

Direct effect

Clinton feeling thermometer $\rightarrow \hat{\beta}_1^M$ Gore feeling thermometer

$\frac{\text{cov}(gore, clinton)}{\text{var}(clinton)} = \gamma_{2,1}$

Party ID $\rightarrow \hat{\beta}_2^M$
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
. 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>
<tr>
<td>Total</td>
<td>13,939</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
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>--- 2.88 (0.16)</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 6.44 6.44</td>
</tr>
<tr>
<td>R²</td>
</tr>
<tr>
<td>.011 .023 .025</td>
</tr>
<tr>
<td>N</td>
</tr>
<tr>
<td>13,876 13,876 13,876</td>
</tr>
</tbody>
</table>

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

What is the substantive interpretation of the coefficients?
The Picture

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

\[ \gamma_{21} = 0.19 \]

Living in frat house \( x_2 \)

Member of fraternity \( x_1 \)

Drinks per 30 days \( Y \)

Remember that:

\[ \hat{\beta}_1^B = 2.88 \]
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

- Living in frat house $X_2$
- Member of fraternity $X_1$
- Drinks per 30 days $Y$

\[ \hat{\beta}_2^M = 2.26 \]
\[ \hat{\beta}_1^M = 2.44 \]
\[ \gamma_{21} = 0.19 \]
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 (85%)</td>
<td>0.44 (15%)</td>
</tr>
<tr>
<td>Live in frat/sor. house</td>
<td>4.44</td>
<td>2.26 (51%)</td>
<td>2.18 (49%)</td>
</tr>
</tbody>
</table>

From bivariate regressions

From multiple regressions

From accounting identity: T=D+I
Return to the state legislative example
. list state ry after10 in 1/10

<table>
<thead>
<tr>
<th>state</th>
<th>ry</th>
<th>after10</th>
</tr>
</thead>
<tbody>
<tr>
<td>West Virginia</td>
<td>25.17959</td>
<td>1</td>
</tr>
<tr>
<td>Rhode Island</td>
<td>23.48404</td>
<td>0</td>
</tr>
<tr>
<td>Kentucky</td>
<td>23.12402</td>
<td>0</td>
</tr>
<tr>
<td>Arkansas</td>
<td>18.00081</td>
<td>1</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>16.55731</td>
<td>1</td>
</tr>
<tr>
<td>Nevada</td>
<td>10.97821</td>
<td>0</td>
</tr>
<tr>
<td>Mississippi</td>
<td>7.828268</td>
<td>0</td>
</tr>
<tr>
<td>Louisiana</td>
<td>7.805305</td>
<td>0</td>
</tr>
<tr>
<td>Hawaii</td>
<td>6.73896</td>
<td>1</td>
</tr>
<tr>
<td>Utah</td>
<td>5.545974</td>
<td>-1</td>
</tr>
</tbody>
</table>

41. Georgia        | -9.231257 | -1 |
42. North Dakota   | -9.460065 | -1 |
43. Indiana        | -9.967912 | -1 |
44. Florida        | -10.72338 | -1 |
45. South Dakota   | -10.92215 | -1 |
46. North Carolina | -11.10709 | 0  |
47. Ohio           | -11.19955 | -1 |
48. Wisconsin      | -14.06958 | -1 |
49. Virginia       | -19.61035 | 0  |
<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obama vote</td>
<td>1.43 (0.14)</td>
<td>---</td>
<td>1.09 (0.14)</td>
</tr>
<tr>
<td>Dem. state</td>
<td>---</td>
<td>16.33 (2.33)</td>
<td>8.25 (1.82)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-23.25 (6.81)</td>
<td>50.51 (1.85)</td>
<td>-4.93 (7.01)</td>
</tr>
<tr>
<td>N</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td>S.E.R.</td>
<td>9.79</td>
<td>12.62</td>
<td>8.23</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.71</td>
<td>.51</td>
<td>.80</td>
</tr>
</tbody>
</table>
Implications of for model-building

• Q: When do you decide whether to “control for” another variable?
  – A1: When excluding another variable(s) would lead to a biased estimate of the effect you are interested in
    • The omitted variable is correlated with the independent variable of interest and
    • The omitted variable is also related (statistically) to the dependent variable.*
  – A2: When theory or the question tell you to (e.g. the role of race in American politics)
  – A3: to deal with efficiency (covered after spring break)

*If you don’t do this, you commit omitted variables bias
Standardized regression

- Used to try and judge which variables are “more important” in a multiple regression
- Other standardizations are possible (e.g., putting all variables into a 0,1 interval)
- Less informative than regressing on raw values or the 0,1 interval, but is useful to know about because it is so common.
The idea

• Transform every variable according to the following formula:

\[ \text{newvar} = \frac{\text{oldvar} - \text{oldvar}}{\sigma_{\text{oldvar}}} \]

• Do the regression on these “z-scores”
  – The intercept drops away
  – In bivariate regression, the standardized coefficient is equal to the correlation coefficient
  – The coefficients are sometimes called “BETA” coefficients (very confusingly)
Example: Knowledge of party control of Congress

• Who knows which party controls the House?
• Variables:
  – `know_reps`: $= 1$ if the R knows the House is controlled by the Reps, 0 otherwise
  – `lfaminc` = (recoded) family income, in thousands
  – `educ` = education completed, in years (recoded from categorical variable)
Summary statistics

```
.summ know_reps lf educ_new [aw=v102]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Weight</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>know_reps</td>
<td>842</td>
<td>842.842355</td>
<td>.6431059</td>
<td>.4793679</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>lfaminc</td>
<td>878</td>
<td>896.217452</td>
<td>4.066185</td>
<td>1.069957</td>
<td>1.609438</td>
<td>6.907755</td>
</tr>
<tr>
<td>educ_new</td>
<td>1000</td>
<td>1000</td>
<td>13.27775</td>
<td>2.483685</td>
<td>8</td>
<td>18</td>
</tr>
</tbody>
</table>
```
Comparison of regular regression and standardized regression

```
. reg know_reps lf educ_new [aw=v102],beta
(sum of wgt is 7.5562e+02)
```

```
Source |       SS       df       MS              Number of obs =     734
-------------+------------------------------ F(  2,   731) =   31.39
Model |  13.2004642     2  6.60023208           Prob > F      =  0.0000
Residual |  153.72410   731  .210292886           R-squared     =  0.0791
-------------+------------------------------ Adj R-squared =  0.0766
Total |  166.924564   733  .227727918           Root MSE      =  .45858

------------------------------------------------------------------------------
know_reps |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
lfaminc |    .046186   .0158845     2.91   0.004                   .10494
educ_new |   .0469317   .0069827     6.72   0.000                 .2425739
_cons |  -.1633859   .1046256    -1.56   0.119                        .
------------------------------------------------------------------------------
```

Number of obs = 734
F( 2, 731) = 31.39
Prob > F = 0.0000
R-squared = 0.0791
Adj R-squared = 0.0766
Root MSE = 0.45858

| Variable   | Coef.  | Std. Err. | t     | P>|t|  | Beta   |
|------------|--------|-----------|-------|------|--------|
| know_reps  |        |           |       |      |        |
| lfaminc    | 0.0462 | 0.0159   | 2.91  | 0.004| 0.1049 |
| educ_new   | 0.0469 | 0.0069   | 6.72  | 0.000| 0.2426 |
| _cons      | -0.1634| 0.1046   | -1.56 | 0.119|        |
Comparison of 0-1 normalization and standardized regression

```
. reg know_reps lfaminc01 educ_new01 [aw=v102],beta
(sum of wgt is 7.5562e+02)

Source |       SS       df  MS
-------------+------------------------------ F(  2,   731) =   31.39
Model |  13.2004641     2  6.60023203           Prob > F      =  0.0000
Residual |    153.7241   731  .210292886           R-squared     =  0.0791
-------------+------------------------------ Adj R-squared =  0.0766
Total |  166.924564   733  .227727918           Root MSE      =  .45858

know_reps |      Coef.   Std. Err.      t    P>|t|    Beta
-------------+---------------------------------------------
lfaminc01 |   .2447083   .0841609     2.91   0.004   .10494
educ_new01 |   .4693169   .0698272     6.72   0.000   .2425739
_cons |   .2864012   .0516328     5.55   0.000
```
(Multi)collinearity

• Collinearity is when two or more independent variables are highly correlated
  – More precisely: when one independent variable is a linear combination of the other independent variables

• Effects:
  – Coefficients of the affected variables may will be unstable
  – Standard errors (for after spring break) will be inflated
  – BUT, the overall predictive power of the regression will not be affected
Simplest example: Sex

• Say I have a variable named **female**, coded 1 if the respondent is female, 0 if the respondent is male

• A second variable named male coded 1 if the respondent is male, 0 if the respondent is female is **collinear** with **female** because
  
  \[ \text{male} = 1 - \text{female} \]
Without the redundant variable

```
. reg approve female [aw=v102]
(sum of wgt is  9.5361e+02)

Source |       SS       df    MS               Number of obs =     977
-------------+------------------------------ F(  1,   975) =    5.70
Model |  8.42953813     1  8.42953813           Prob > F      =  0.0172
Residual |  1442.49523   975  1.47948229           R-squared     =  0.0058
-------------+------------------------------ Adj R-squared =  0.0048
Total |  1450.92477   976  1.48660325           Root MSE      =  1.2163

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
female |   .1858039   .0778409     2.39   0.017     .0330489    .3385589
_cons |   2.292257   .0555348    41.28   0.000     2.183275    2.401238
------------------------------------------------------------------------------
```
With the redundant variable

```
. reg approve female male [aw=v102]
(sum of wgt is 9.5361e+02)
note: male omitted because of collinearity
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8.42953813</td>
<td>1</td>
<td>8.42953813</td>
</tr>
<tr>
<td>Residual</td>
<td>1442.49523</td>
<td>975</td>
<td>1.47948229</td>
</tr>
<tr>
<td>Total</td>
<td>1450.92477</td>
<td>976</td>
<td>1.48660325</td>
</tr>
</tbody>
</table>

Number of obs =    977
F(  1,   975) =    5.70
Prob > F =    0.0172
R-squared =    0.0058
Adj R-squared =    0.0048
Root MSE =    1.2163

| approve_ob~a | Coef. | Std. Err. | t    | P>|t| | [95% Conf. Interval] |
|--------------|-------|-----------|------|------|----------------------|
|              | female| .1858039  | .0778409 | 2.39 | 0.017    | .0330489   | .3385589 |
|              | male  | 0 (omitted)|       |      |          |           |         |
|              | _cons | 2.292257  | .0555348 | 41.28 | 0.000    | 2.183275   | 2.401238 |

```
With a fake collinear version of female

. gen female2=female

. replace female2=0 if female==1&_n<=50
(15 real changes made)

. tab female female2

|        female2 |         0          1 |     Total  
-----------+----------------------+----------
0 |       461          0 |       461  
1 |        15        524 |       539  
-----------+----------------------+----------
Total |       476        524 |     1,000  

. corr female female2
(oss=1000)

|         female female2  
-------------+------------------
female |   1.0000  
female2 |   0.9703   1.0000  

With a fake collinear version of female

```
. reg approve_o female female2 [aw=v102]
(sum of wgt is 9.5361e+02)

Number of obs = 977
F(  2,   974) = 6.41
Prob > F      = 0.0017
R-squared     = 0.0130
Adj R-squared = 0.0110
Root MSE      = 1.2126

Source |       SS       df  MS              Number of obs = 977
-------------+------------------------------ F(  2,   974) = 6.41
Model | 18.8520935     2  9.42604675           Prob > F      = 0.0017
Residual | 1432.07268   974  1.47030049           R-squared     = 0.0130
-------------+------------------------------ Adj R-squared = 0.0110
Total | 1450.92477   976  1.48660325           Root MSE      = 1.2126

______________________________
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
______________________________
female |    1.03613   .3286674     3.15   0.002     .3911523    1.681108
female2 |   -.874974   .3286329    -2.66   0.008    -1.519884   -.2300639
_cons |   2.292257   .0553622    41.40   0.000     2.183614      2.4009
______________________________
```
Multiple dummy variables

• Dummy variables are how we deal with categorical independent variables
  – Race (white, black, Hispanic, Asian-Amer., other)
  – Alliance (NATO, Warsaw Pact, other)
  – Religion (Christian, Jewish, Muslim, Hindu, other)

• To code dummy variables, create a new variable for each category
  – In the regression, exclude one of the variables (usually the most numerous) --- **remember the lecture on collinearity**
Example: predicting Obama approval by race (white is omitted category)

```
. gen white=race==1
. gen black=race==2
. gen hisp=race==3
. gen other_race=1-white-black-hisp
. reg approve_o black hisp other_race [aw=v102]
(sum of wgt is 9.5361e+02)
```

```
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Source</td>
<td>SS</td>
<td>df</td>
<td>MS</td>
</tr>
<tr>
<td>----------</td>
<td>------------</td>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>Model</td>
<td>135.231871</td>
<td>3</td>
<td>45.0772903</td>
</tr>
<tr>
<td>Residual</td>
<td>1315.6929</td>
<td>973</td>
<td>1.35220236</td>
</tr>
<tr>
<td>Total</td>
<td>1450.92477</td>
<td>976</td>
<td>1.48660325</td>
</tr>
</tbody>
</table>

Number of obs = 977
F( 3, 973) = 33.34
Prob > F = 0.0000
R-squared = 0.0932
Adj R-squared = 0.0904
Root MSE = 1.1628

| approve_o~a | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|------------|---------|-----------|-------|-----|-----------------------|
| black      | 1.097946 | .1125408  | 9.76  | 0.000 | .8770956 1.318797    |
| hisp       | .4846481 | .1577768  | 3.07  | 0.002 | .1750261 .79427     |
| other_race | .0429154 | .2131171  | 0.20  | 0.840 | -.3753068 .4611375  |
| _cons      | 2.216973 | .0420862  | 52.68 | 0.000 | 2.134383 2.299563    |
```
Example: predicting Obama approval by race (forgetting omitted category)

```
. reg approve_o white black hisp other_race [aw=v102]
(sum of wgt is 9.5361e+02)
note: other_race omitted because of collinearity
```

```
Source |       SS       df MS              Number of obs =     977
-------------+------------------------------ F(  3,   973) =   33.34
Model |  135.231871     3  45.0772903           Prob > F      =  0.0000
Residual |   1315.6929  973  1.35220236           R-squared     =  0.0932
-------------+------------------------------ Adj R-squared =  0.0904
Total |  1450.92477  976  1.48660325           Root MSE      =  1.1628

------------------------------------------------------------------------------
approve_ob~a |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  white |  -.0429154   .2131171    -0.20   0.840    -.4611375    .3753068
  black |   1.055031    .233542     4.52   0.000     .5967269    1.513335
  hisp |   .4417327   .2583988     1.71   0.088    -.0653504    .9488157
  other_race |          0  (omitted)
   _cons |   2.259888   .2089202    10.82   0.000     1.849902    2.669875
------------------------------------------------------------------------------
```
Interaction Terms

• Sometimes we think that one set of regression coefficients apply to one population and another set to another population

• Example: evaluations of Barak Obama
  – Assume ideology and party ID is the baseline factor leading to approval of Obama
  – We know that African-Americans approve of Obama more than whites, but is it:
    • African-Americans simply like Obama better, controlling for Ideology and party ID or
    • African-Americans weigh ideology and party ID differently?
```plaintext
. reg approve_obama liberal01 dem01 african_am [aw=v102]
(sum of wgt is 7.5675e+02)

Source       | SS       | df  | MS              
-------------|----------|-----|-----------------
Model        | 638.646662| 3   | 212.882221      
Residual     | 495.553833| 777 | .637778421      
Total        | 1134.20049| 780 | 1.4541032       

Number of obs = 781
F(  3,   777) = 333.79
Prob > F      = 0.0000
R-squared     = 0.5631
Adj R-squared = 0.5614
Root MSE      = .79861

approve_ob~a | Coef.     | Std. Err. | t     | P>|t| | [95% Conf. Interval] 
---------------|-----------|-----------|-------|------|------------------------
liberal01     | 1.662104  | .1287268  | 12.91 | 0.000| 1.40941 1.914797     
dem01         | 1.222213  | .0929022  | 13.16 | 0.000| 1.039844 1.404582    
african_am    | .5646431  | .0907161  | 6.22  | 0.000| .3865655 .7427207   
_cons         | .8615467  | .0569603  | 15.13 | 0.000| .7497324 .9733611   
```
. reg approve_obama liberal01 dem01 african_am [aw=v102] if african_am==1
(sum of wgt is 1.0086e+02)
note: african_am omitted because of collinearity

| approve_obama | Coef.   | Std. Err. |   t    | P>|t| | [95% Conf. Interval] |
|---------------|---------|-----------|--------|-----|----------------------|
| liberal01     | 0.69666 | 0.31922   | 2.18   | 0.031 | 0.0634095 - 1.329911 |
| dem01         | 1.20565 | 0.34771   | 3.47   | 0.001 | 0.5158776 - 1.895426 |
| african_am    | 0       | (omitted) |        |      |                      |
| _cons         | 1.88605 | 0.31818   | 5.93   | 0.000 | 1.254866 - 2.517224  |

. reg approve_obama liberal01 dem01 african_am [aw=v102] if african_am==0
(sum of wgt is 6.5588e+02)
note: african_am omitted because of collinearity

| approve_obama | Coef.   | Std. Err. |   t    | P>|t| | [95% Conf. Interval] |
|---------------|---------|-----------|--------|-----|----------------------|
| liberal01     | 1.93479 | 0.14099   | 13.72  | 0.000 | 1.657948 - 2.21165   |
| dem01         | 1.11863 | 0.09697   | 11.54  | 0.000 | 0.9282298 - 1.309027 |
| african_am    | 0       | (omitted) |        |      |                      |
| _cons         | 0.78904 | 0.05722   | 13.79  | 0.000 | 0.6766921 - 0.9013826 |
• We can do this comparison all in one regression

• $\text{obama}\_\text{approve} = b_0 + b_1*\text{african}_\text{am} + b_2*\text{liberal01} + b_3*\text{dem01} + b_4*\text{african}_\text{am}\times\text{liberal01} + b_5*\text{african}_\text{am}\times\text{dem01}$
Note, if the R is white, african_am = 0, therefore:

\[
\text{obama\_approve} = b_0 + b_1 \times \text{african\_am} + b_2 \times \text{liberal01} + b_3 \times \text{dem01} + b_4 \times \text{african\_am} \times \text{liberal01} + b_5 \times \text{african\_am} \times \text{dem01}
\]

BECOMES

\[
\text{obama\_approve} = b_0 + b_2 \times \text{liberal01} + b_3 \times \text{dem01} +
\]
• Note, if the R is black, african_am = 1, therefore:

• obama_Approve = b₀ + b₁*african_am + b₂*liberal01 + b₃*dem01 + b₄*african_am*liberal01 + b₅*african_am*dem01

• BECOMES

• obama_Approve = (b₀ + b₁) + (b₂+b₄)liberal01 + (b₃+ b₅)dem01
. gen african_amXliberal01=african_am*liberal01
(166 missing values generated)

. gen african_amXdem01=african_am*dem01
(180 missing values generated)

. reg approve_obama african_am liberal01 dem01 african_amXliberal01 african_amXdem01
[aw=v102]
(sum of wgt is 7.5675e+02)

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 781</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>649.513362</td>
<td>5</td>
<td>129.902672</td>
<td>F(  5,   775) =  207.71</td>
</tr>
<tr>
<td>Residual</td>
<td>484.687133</td>
<td>775</td>
<td>.625402752</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1134.20049</td>
<td>780</td>
<td>1.4541032</td>
<td>R-squared = 0.5727</td>
</tr>
</tbody>
</table>

| approve_obama | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------------|-------|-----------|-------|------|---------------------|
| african_am    | 1.097008 | .2732159 | 4.02  | 0.000 | .5606767 - 1.633339 |
| liberal01     | 1.934799 | .1456907 | 13.28 | 0.000 | 1.648803 - 2.220794 |
| dem01         | 1.118628 | .1001958 | 11.16 | 0.000 | .9219412 - 1.315316 |
| african_amXliberal01 | -1.238138 | .3047051 | -4.06 | 0.000 | -1.836283 - .639993 |
| african_amXdem01 | .0870236 | .3082445 | 0.28  | 0.778 | -.5180694 - .6921166 |
| _cons         | .7890373 | .0591208 | 13.35 | 0.000 | .6729814 - .9050932 |
A more parsimonious model excludes the party interaction

```
.reg approve_obama african_am liberal01 dem01 african_amXliberal01 [aw=v102]
(sum of wgt is 7.5675e+02)

Source |       SS       df  MS
-------------+------------------------------ F(  4,   776) =  259.93
Model | 649.463515     4 162.365879           Prob > F      =  0.0000
Residual | 484.73698   776  .624661057           R-squared     =  0.5726
-------------+------------------------------ Adj R-squared =  0.5704
Total | 1134.20049   780  1.4541032           Root MSE      =  .79036

approve_obama |       Coef.   Std. Err.      t    P>|t|      [95% Conf. Interval]
---------------------+--------------------------------------------------
african_am |   1.157717   .1684415     6.87   0.000     .8270618    1.488372
liberal01 |   1.926111   .1423196    13.53   0.000     1.646734    2.205488
dem01 |   1.127823   .0946985    11.91   0.000     .9419277    1.313719
african_amXliberal01 |  -1.213316    .291572    -4.16   0.000     -1.78568   -.6409531
_cons |   .7884288   .0590465    13.35   0.000     .6725191    .9043386

.predict py if e(sample)
(option xb assumed; fitted values)
(219 missing values generated)
```
Predicted approval level

Liberal scale (0,1)

White

African-American

Republicans

Independents

Democrats