Four Special Topics

Interaction Terms
Standardized Regression
Decomposing Regression Effects
Measurement Error
Interaction Terms

What Happens When Different Models Apply in Different Situations?
Regression of Blank Ballots (1996) on Median Rent (1990)

<table>
<thead>
<tr>
<th></th>
<th>Coeff.</th>
<th>s.e.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.36</td>
<td>0.48</td>
</tr>
<tr>
<td>Slope</td>
<td>-0.65</td>
<td>0.088</td>
</tr>
<tr>
<td>N</td>
<td>663</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.077</td>
<td></td>
</tr>
</tbody>
</table>

![Graph showing the regression of blank ballots on median rent with data points and fitted line.](image)
Regression of Blank Ballots (1996) on Median Rent (1990), By Ballot Type

Scanning

- Blank96
- Medianre
- Fitted values
- Coeff. 0.040  s.e. 0.48
- Slope -0.74  s.e. 0.088
- N 491
- $R^2$ 0.10

Electronic

- Blank96
- Medianre
- Fitted values
- Coeff. -2.83  s.e. 0.82
- Slope -0.14  s.e. 0.15
- N 172
- $R^2$ 0.005
What to do?

• Run two separate regressions
  – Advantage: conceptually simple
  – Disadvantage: hypothesis testing cumbersome

• Interaction terms
  – Advantage: hypothesis testing facilitated
  – Disadvantage: conceptually complex
Interaction terms generally

\[ y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon \]

Rewriting,

\[ y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \]
Interaction terms in the voting machine example

• Define $S_c = 1$ if the county uses optical scanning, 0 otherwise

• Run this regression:

$$\text{blankpct}_c = \beta_0 + \beta_1 \times \text{rent}_c + \beta_2 \times S_c + \beta_3 \times S_c \times \text{rent}_c + \varepsilon_c$$

Note that if $S_c = 0$ (i.e., electronic county), we have

$$\text{blankpct}_c = \beta_0 + \beta_1 \times \text{rent}_c + \varepsilon_c$$

If $S_c = 1$ (i.e., scanned county), we have

$$\text{blankpct}_c = \beta_0 + \beta_1 \times \text{rent}_c + \beta_2 + \beta_3 \times \text{rent}_c + \varepsilon_c \text{ or}$$

$$\text{blankpct}_c = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{rent}_c + \varepsilon_c$$
Doing this in **STATA**

```plaintext
. gen scan=ve96_cod=="5"
. gen s=scan
. gen scanrent=scan*rent
. reg blank rent scan scanrent
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 663</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>48.597571</td>
<td>3</td>
<td>16.1991903</td>
<td>F( 3, 659) = 32.71</td>
</tr>
<tr>
<td>Residual</td>
<td>326.36878</td>
<td>659</td>
<td>.495248527</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>374.966351</td>
<td>662</td>
<td>.566414427</td>
<td>R-squared = 0.1296</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.1256</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 0.70374</td>
</tr>
</tbody>
</table>

| blank | Coef.  | Std. Err. | t    | P>|t|   | [95% Conf. Interval] |
|-------|--------|-----------|------|------|---------------------|
| rent  | -.14256| .1686598  | -0.85| 0.398| -.4737353 .1886153  |
| scan  | 2.866141| 1.051092  | 2.73 | 0.007| .8022468 4.930035  |
| scanrent | -.6017711| .196494  | -3.06| 0.002| -.987601 -.2159413 |
| _cons | -2.826596| .8975356 | -3.15| 0.002| -4.58897 -1.064222 |
```
Standardized Regression

Comparing (Standardized) Apples with (Standardized) Oranges
Which “matters” more in determining vote outcomes, popularity or the economy?

```
. reg vote drdi Gallup
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>0.038942217</td>
<td>2</td>
<td>0.019471109</td>
</tr>
<tr>
<td>Residual</td>
<td>0.009732889</td>
<td>10</td>
<td>0.000973289</td>
</tr>
<tr>
<td>Total</td>
<td>0.048675106</td>
<td>12</td>
<td>0.004056259</td>
</tr>
</tbody>
</table>

Number of obs = 13
F( 2, 10) = 20.01
Prob > F = 0.0003
R-squared = 0.8000
Adj R-squared = 0.7601
Root MSE = 0.0312

| vote  | Coef. | Std. Err. | t    | P>|t|   | [95% Conf. Interval] |
|-------|-------|-----------|------|-------|---------------------|
| drdi  | 1.908849 | 0.545243  | 3.50 | 0.006 | 0.6939719 to 3.123726 |
| Gallup| 0.2554055 | 0.07231   | 3.53 | 0.005 | 0.0942888 to 0.4165223 |
| _cons | 0.3422054 | 0.0350065 | 9.78 | 0.000 | 0.2642061 to 0.4202047 |
Solutions I

• Normalize into percentages
  – Take logs of everything
  – Advantage: elegant
  – Disadvantages:
    • Not always appropriate transform
    • Zero, negative numbers

\[ \ln(y) = \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \epsilon \]

Calculate \( \frac{\partial y}{\partial x_1} \) and \( \frac{\partial y}{\partial x_2} \) and rearrange terms:

\( \beta_1 = \frac{\partial y / y}{\partial x_1 / x_1}, \beta_2 = \frac{\partial y / y}{\partial x_2 / x_2} \)
Solutions II

- Transform the variables into unit deviates (i.e., mean 0, s.d. 1)
  - Subtract each variable from its mean and divide by its standard deviation, i.e.:

\[
Z_{i,j} = \frac{(Z_{i,j} - \bar{Z}_i)}{s_{Z_i}}
\]
Doing this in STATA

```
. reg vote drdi gallup, beta

Source |       SS   df       MS              Number of obs =      13
-------------+------------------------------ F(  2,    10) =   20.01
Model |  .038942217     2  .019471109   Prob > F      =  0.0003
Residual |  .009732889    10  .000973289         R-squared     =  0.8000
          |                     Adj R-squared =  0.7601
          |                           Root MSE =   .0312
-------------+------------------------------
Total |  .048675106    12  .004056259

------------------------------------------------------------------------------
vote | Coef.  Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
drdi |   1.908849    .545243     3.50   0.006                  .535644
gallup |   .2554055     .07231     3.53   0.005                 .5404141
_cons |   .3422054   .0350065     9.78   0.000           .
------------------------------------------------------------------------------
```
Decomposing Regression Effects

Direct and Indirect Effects
Recall the OLS solution

If

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

then

\begin{align*}
\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)}
\end{align*}
Rearrange the first line

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

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

(Overall association b/t \(X_1\) and \(Y\)) =

(Direct effect of \(X_1\) on \(Y\)) +

(Direct effect of \(X_2\) on \(Y\)) \times (Bivariate effect of \(X_1\) on \(X_2\))
Graphically

\[
\frac{\text{cov}(X_1, X_2)}{\text{var}(X_1)} = \gamma_{1,2}
\]

(Overall association b/t \(X_1\) and \(Y\)) =
(Direct effect of \(X_1\) on \(Y\)) +
(Direct effect of \(X_2\) on \(Y\)) \times (Bivariate effect of \(X_1\) on \(X_2\))
Decomposing the effects of popularity and the economy on the vote

<table>
<thead>
<tr>
<th>Effect</th>
<th>Bivariate</th>
<th>Direct</th>
<th>Indirect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gallup</td>
<td>0.35</td>
<td>0.26</td>
<td>0.097</td>
</tr>
<tr>
<td></td>
<td>(100%)</td>
<td>(74%)</td>
<td>(26%)</td>
</tr>
<tr>
<td>Economy</td>
<td>2.64</td>
<td>1.91</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td>(100%)</td>
<td>(72%)</td>
<td>(28%)</td>
</tr>
</tbody>
</table>
Measurement Error

What Happens When You Can’t Measure Things Perfectly?
Suppose we measure $x$ with error?

Instead of observing $x$, we observe $x' = x + e$

($e$ is random with mean $\bar{e}$ and variance $\nu_e$)

∴ instead of doing the regression

$y = \alpha + \beta x + \varepsilon,$

we do the regression

$y = \alpha + \beta' x' + \varepsilon.$

What is the relationship between $\beta$ and $\beta'$?
Answer

\[
\beta' = \frac{\text{cov}(x, y)}{\text{var}(x) + \text{var}(e)}
\]
Errors in Independent Variables: The Picture
Suppose we measure $y$ with error $\varepsilon$.

Instead of observing $y$, we observe $y' = y + e$ (where $e$ is random with mean $\bar{e}$ and variance $\nu_e$).

\[ y = \alpha + \beta x + \varepsilon, \]

we do the regression

\[ y' = \alpha + \beta' x + \varepsilon. \]

What is the relationship between $\beta$ and $\beta'$?
The answer

\[ \beta' = \frac{\text{cov}(x, y)}{\text{var}(x)} = \beta \]

But…

• Standard errors and s.e.r. inflated
• R^2 deflated
Errors in Dependent Variables: The Picture