Multiple regression: fitting the best fitting plane

\[ \hat{Y} = a + b_1 X_1 + b_2 X_2 \]

Where, 
- \( a \) = intercept (when \( X_1 = 0 \) and \( X_2 = 0 \))
- \( Y \) = income
- \( X_1 \) = education
- \( X_2 \) = % female in occupation

[unit of analysis = occupation]

---

Multiple regression: model

Educ \[\rightarrow\] Occupational income

% female \[\rightarrow\] Occupational income
Multiple regression: 3D model

Source: Agresti and Finlay, 1986, p. 317
Multiple regression: 
interpretation of b’s

- $b_1 =$ slope for education (net effect of education on income controlling for percent female; how much income in dollars for each year of education)
- $b_2 =$ slope for % female (net effect of percent female on income controlling for education; how much income in dollars for each percent female)

Coefficients called:
- Metric coefficient
- Net regression coefficient
- Partial regression coefficient
- Unstandardized regression coefficient

Interpretation:
- Net effect
- Independent effect
- Partial relationship
- Controlling for
Multiple regression: standardized regression coefficients

\[ \hat{Y} = a + b_1X_1 + b_2X_2 \]
\[ \hat{y} = B_1x_1 + B_2x_2 \]

where,

\[ B_{yx} = b_{yx} \left( \frac{s_x}{s_y} \right) \]

Creating standard scores:

\[ x_i = (X_i - \bar{X_i}) / s_x \]

Multiple regression: standardized regression coefficients, or “relative effects”

- B’s range from -1 to +1
- Interpretation: a one s.d. change in the independent variable produces a predicted change of “Beta” s.d.’s in the dependent variable, net of other variables
- More common interpretation: if \( B_1 > B_2 \) then education is more important in predicting income than is % female
- “considerably larger, more than twice as important”
Murdock data:

$Y =$ stratification, $X_1=$political integration, $X_2=$money exchange

<table>
<thead>
<tr>
<th>Society</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$Y$</th>
<th>Society</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$Y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>001</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>011</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>003</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>013</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>005</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>015</td>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>007</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>017</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>009</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>019</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

Calculating standardized and unstandardized coefficients
( computational formula)

$$r_{y_i|x_i} = \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n \sum X_i^2 - (\sum X_i)^2][n \sum Y_i^2 - (\sum Y_i)^2]}}$$
Calculating standardized and unstandardized coefficients (adapting for Y and X₂)

\[
r_{yX_2} = \frac{n \sum X_2 Y - (\sum X_2)(\sum Y)}{\sqrt{[n \sum X_2^2 - (\sum X_2)^2][n \sum Y^2 - (\sum Y)^2]}}
\]

Calculating standardized and unstandardized coefficients

\[
\begin{align*}
r_{yX_1} &= .865 \\
r_{yX_2} &= .620 \\
r_{x_1x_2} &= .482
\end{align*}
\]
Calculating standardized coefficients:

general formula

\[ B_{yx.z} = \frac{r_{yx} - r_{yz} r_{xz}}{1 - r^2_{xz}} \]

Calculating standardized coefficients:

adapt for \( B_{yx1.x2} \)

\[ B_{yx1.x2} = \frac{r_{yx1} - r_{yx2} r_{x1.x2}}{1 - r^2_{x1.x2}} \]
Calculating standardized coefficients

\[ B_{yx_1 \cdot x_2} = .737 \]
\[ B_{yx_2 \cdot x_1} = .265 \]
\[ \hat{y} = .737x_1 + .265x_2 \]

**Interpretation:** Political integration is substantially more important than money in determining level of stratification.

Calculating unstandardized coefficients

\[ b_{yx} = B_{yx} \left( \frac{s_y}{s_x} \right), \]

Using s.d. computational formula

\[ s_y = \sqrt{[\sum Y^2 / n] - [\sum Y / n]^2} \]
Calculating unstandardized coefficients

Standard deviations

\[
\begin{align*}
    s_y &= .9 \\
    s_{x_1} &= 1.37 \\
    s_{x_2} &= 1.33
\end{align*}
\]

\[
b_{yx_1 \cdot x_2} = B_{yx_1 \cdot x_2} \left( \frac{s_y}{s_{x_1}} \right) \\
= (.737)(.9 / 1.37) = .484
\]

\[
b_{yx_2 \cdot x_1} = B_{yx_2 \cdot x_1} \left( \frac{s_y}{s_{x_2}} \right) \\
= (.265)(.9 / 1.33) = .179
\]

Calculating intercept

\[
\bar{Y} = a + b_1 \bar{X}_1 + b_2 \bar{X}_2
\]

\[
a = \bar{Y} - b_1 \bar{X}_1 - b_2 \bar{X}_2
\]

\[
a = (1.3) - (.484)(2.1) - (.179)(1.2)
\]

\[
a = .069
\]
Prediction equation:  
Interpret!

\[
\hat{Y} = .069 + .484X_1 + .179X_2
\]

Calculating \( R^2 \)

\[
R^2_{y.x_1x_2} = B_{y|x_1} r_{y|x_1} + B_{y|x_2} r_{y|x_2}
\]
\[
= (.737)(.865) + (.265)(.620)
\]
\[
= .802
\]

Interpretation: 80 percent of the variation in stratification is explained by political integration and money
Standardized vs. unstandardized coefficients

- Use standardized to compare variables within equations
- Use unstandardized to compare same variable across equations

Unstandardized coefficients
Caveats

1) Don’t interpret regression lines beyond where you have data

2) Report to three significant nonzero digits (retain larger # of digits in intermediate calculations)

3) Multicollinearity: problem with high correlations