Data: National Youth Longitudinal Study

Causal relationship

Types of multivariate relationships

Multiple Regression

◆ Model
◆ Interpretation
◆ Analysis of variance

Reading: Chapters 10 & 11 in Agresti & Finlay
A Data Set

- National Youth Longitudinal Study.

- The Design:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>13</td>
<td>14</td>
<td>16</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>14</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>15</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>16</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>20</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>21</td>
</tr>
</tbody>
</table>
Variables in NYLS

- Youth’s attitude toward deviant behavior: Are each of the nine behaviors wrong?
  
  - Cheat on tests at school
  - Damage or destroy property
  - Hit someone
  - Break into a car
  - Sell hard drugs
  - Use Marijuana
  - Use Alcohol
  - Steal less than $5
  - Steal more than $50

- Response options and scoring:
  
  1 = not wrong
  2 = a little bit wrong
  3 = wrong
  4 = very wrong
Variables in NYLS

- **Peer**: How many of your close friends engage in (each of) the deviant behaviors. The response options

  5 = all of them
  4 = most of them
  3 = some of them
  2 = very few
  1 = none

- **Parent’s attitude**: Approve of the behaviors (excluding selling hard drugs). The response options were

  1 = strongly approve
  2 = approve
  3 = neither
  4 = disapprove
  5 = strongly disapprove
Variables in NYLS

Only using data from 1976 and \( n = 1408 \) (i.e., only cases with no missing data).

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>youth</td>
<td>31.20</td>
<td>4.14</td>
<td>14.00</td>
<td>36.00</td>
</tr>
<tr>
<td>parent</td>
<td>33.07</td>
<td>2.30</td>
<td>22.00</td>
<td>35.00</td>
</tr>
<tr>
<td>peer</td>
<td>15.20</td>
<td>5.33</td>
<td>9.00</td>
<td>40.00</td>
</tr>
<tr>
<td>age</td>
<td>13.84</td>
<td>1.92</td>
<td>11.00</td>
<td>17.00</td>
</tr>
<tr>
<td>gpa</td>
<td>3.75</td>
<td>0.81</td>
<td>1.00</td>
<td>5.00</td>
</tr>
<tr>
<td>gender</td>
<td></td>
<td></td>
<td>48% boys, 52% girls</td>
<td></td>
</tr>
</tbody>
</table>
Correlations between variables

All are significant at $\alpha = .05$.

<table>
<thead>
<tr>
<th></th>
<th>youth</th>
<th>parent</th>
<th>peer</th>
<th>age</th>
<th>gpa</th>
</tr>
</thead>
<tbody>
<tr>
<td>youth</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>parent</td>
<td>.17</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>peer</td>
<td>-.60</td>
<td>-.06</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>-.36</td>
<td>-.08</td>
<td>.32</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>gpa</td>
<td>.17</td>
<td>.06</td>
<td>-.23</td>
<td>-.10</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Causal Relationships

e.g.: Does having friends who engage in deviant behaviors cause kids to think the behaviors are not wrong?

- Determining causality is difficult.
- At a minimum you must show
  - An association exists between variables.
  - Proper time ordering.
  - Eliminate alternative explanations
- Control Variables (Statistical Control). e.g.,
  Does a relationship between youth attitude and Peer’s actions exist if we hold parent’s attitude constant?
Multivariate Relationship: Type I

Relationship exists

Spurious Relationship

If we control for Peer the relationships between Youth and GPA disappears
A Data Set

Causal Relationships

Types of Multivariate Relationships

- Multivariate Relationship: Type I
- Multivariate Relationship II
- Multivariate Relationship III
- Multivariate Relationship IV
- Multivariate Relationship V
- Summary (Adapted from Agresti & Finlay)

Multiple Regression

Interpreting $b_j$

ANOVA, $R^2$, & Sums of Squares

---

### Multi-variate Relationship II

**Relationship exists:**

- GPA
- Youth

**Chain relationship:**

- GPA
- Peer
- Youth

If control for Peer, the relationship between GPA and Youth disappears.

---

**GPA and Youth disappears.**
Multivariate Relationship III

Interaction of Age & Peer on Youth

Relationship between Peer and Youth varies with Age.
A Data Set

Causal Relationships

Types of Multivariate Relationships
- Multivariate Relationship: Type I
- Multivariate Relationship II
- Multivariate Relationship III
- Multivariate Relationship IV
- Multivariate Relationship V
- Summary (Adapted from Agresti & Finlay)

Multiple Regression

Interpreting $b_j$

ANOVA, $R^2$, & Sums of Squares

Multivariate Relationship IV

Multiple “Causes”

Parent

Youth

Peer

Relationship between Peer & Youth does not change with Parent.

Relationship between Parent & Youth does not change with Peer.
Multivariate Relationship V

Direct and Indirect Effects
- Direct effect of Peer on Youth.
- Direct effect of Age on Youth.
- Indirect effect of Age on Youth.

Relationship between Peer and Youth changes when control for Age in model, but it does not disappear.
<table>
<thead>
<tr>
<th>Multivariate Relationship</th>
<th>Relationship between $X_1$ and $Y$ disappears</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1 \rightarrow X_2 \rightarrow Y$</td>
<td>Chain relationship; $X_2$ intervenes; $X_1$ indirectly “causes” $Y$</td>
</tr>
<tr>
<td>$X_2$</td>
<td>Interaction</td>
</tr>
<tr>
<td>$X_1 \rightarrow Y$</td>
<td>Relationship between $X_1$ and $Y$ varies according to value of $X_2$</td>
</tr>
<tr>
<td>$X_1 \rightarrow Y$</td>
<td>Multiple “causes”</td>
</tr>
<tr>
<td>$X_2 \rightarrow Y$</td>
<td>Relationship between $X_1$ and $Y$ does not change</td>
</tr>
<tr>
<td>Both Direct and indirect effects</td>
<td>Relationship between $X_1$ and $Y$ changes, but does not disappear</td>
</tr>
</tbody>
</table>

Summary (Adapted from Agresti & Finlay)
Multiple Regression:

For two explanatory variables (i.e. X’s)…

- The basic model:

  \[ Y_i = \mu_{Y|X_1, X_2} + \epsilon_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon_i \]

  where \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \).

- \( E(Y|X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2 \)

- For 2 explanatory variables, the regression “surface” is a plane and predicted values are points on the plane.
Multiple Regression: 2 $X$’s

\[ \hat{Y} = \alpha + \beta_1 X_1 \rightarrow \]

\[ \hat{Y} = \alpha + \beta_2 X_2 \]

\[ \hat{Y} = \alpha + \beta_1 X_1 + \beta_2 X_2 \]
Estimation

- Least square estimates, find $a, b_1,$ and $b_2$ to minimize

$$
\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (a + b_1 x_{1i} + b_2 x_{2i}))^2
$$

- The LSE for “slope” parameters are

$$
b_1 = \frac{\frac{sy_{x1}}{s_{x1}^2} - r_{yx2} r_{x1x2} \frac{sy}{s_{x2}}} {1 - r_{x1x2}^2}
$$

$$
b_2 = \frac{\frac{sy_{x2}}{s_{x2}^2} - r_{yx1} r_{x1x2} \frac{sy}{s_{x1}}} {1 - r_{x1x2}^2}
$$

“partial regression coefficients”
Interpreting $b_j$

Compare regression coefficients from

- **Simple linear regression (i.e., $Y = a + bX$),**

  \[
  b = r_{yx} \frac{s_y}{s_x} = \frac{\text{cov}(y, x)}{\text{var}(X)} = \frac{s_{yx}}{s_x^2}
  \]

- **Multiple regression (with predictor variables $x_1$ & $x_2$),**

  \[
  b_1 = \frac{s_{yx_1}}{s_{x_1}^2} - r_{yx_2} r_{x_1 x_2} \frac{s_y}{s_{x_2}}
  \]

  \[
  b_1 = \frac{s_{yx_1}}{s_{x_1}^2} - r_{x_1 x_2}^2 \frac{s_y}{s_{x_2}}
  \]

  If $r_{x_1 x_2} = 0$, then

  \[
  b_1 = \frac{s_{yx_1}}{s_{x_1}^2}
  \]

- **NYLS: Youth $= age + peer$**
- **NYLS: Youth $= age + peer$**
- **NYLS: Youth $= age + peer$**
- **NYLS: Youth $= age + peer$**
- **NYLS: Another Model**
- **Multiple "Causes"**
- **Youth $= parent + age + peer$**
- **Interaction**
- **Youth $= parent + age + peer + (age)(peer)$**
- **Youth $= parent + age + peer + (age)(peer)$**
- **Interaction**
Interpreting $b_j$

$$r_{yx_1|x_2} = \frac{r_{yx_1} - r_{yx_2}r_{x_1x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1x_2}^2)}}$$

- $r_{yx_1|x_2}$ is the partial correlation between $Y$ and $X_1$ controlling for $X_2$.
- Proportional to standardized partial regression coefficient.
- Correlation between the residuals from predicting $Y$ from $X_2$ and the residuals from predicting $X_1$ from $X_2$,

$$Y_i = a_1 + b_1 X_{2i} + e_{yi} \quad \text{and} \quad X_{1i} = a_2 + b_2 X_{2i} + e_{x_1i}$$

$$r_{yx_1} = \text{correlation between } \hat{e}_y \text{ and } \hat{e}_{x_1}$$
NYLS: Youth = age + peer

Model: \( Y_i = \alpha + \beta_1 (age)_i + \beta_2 (peer)_i + \epsilon_i \longrightarrow R^2 = .39 \)

| Parameter | Estimate | Std Error | t Value | Pr > |t| |
|-----------|----------|-----------|---------|-------|---|
| Intercept | 43.44    | 0.63      | 69.34   | <.0001|   |
| age       | -0.43    | 0.05      | -9.07   | <.0001|   |
| peer      | -0.42    | 0.02      | -24.47  | <.0001|   |

\[ \hat{Y}_i = 43.44 - 0.43(age)_i - 0.42(peer)_i \]

Interpretation:
A Data Set
Causal Relationships
Types of Multivariate Relationships
Multiple Regression

Interpreting $b_j$
- $b_j$ for NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer

Interaction
- Youth = parent + age + peer
- Youth = parent + age + peer + (age)(peer)
- Youth = parent + age + peer + (age)(peer)

ANOVA, $R^2$, & Sums of Squares

Fitted Surface in 3-Dimensions

NYLS: Youth = age + peer
A Data Set Causal Relationships Types of Multivariate Relationships Multiple Regression Interpreting $b_j$

NYLS: Youth = age + peer

$Y = a + b_1 \cdot \text{age} + b_2 \cdot \text{peer}$

Fitted Surface in 3-Dimensions

ANOVÁ, $R^2$, & Sums of Squares
NYLS: Youth = age + peer

Y = a + b1*age + b2*peer

Youth by Peer

| age | 11 | 12 | 13 | 14 | 15 | 16 | 17 |

Youth by Age

<table>
<thead>
<tr>
<th>peer_quartile</th>
<th>25t</th>
<th>50t</th>
<th>75t</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max</td>
<td>Min</td>
<td></td>
</tr>
</tbody>
</table>

How Often Peer Engages in Deviant Behavior

Age of Youth

ANOVA, $R^2$, & Sums of Squares
NYLS: Another Model

Simple Relationship

Model: \( Y_i = \alpha + \beta_1(gpa) + \epsilon \rightarrow R^2 = .03 \)

| Parameter | Estimate | Std Error | t Value | Pr > |t|
|-----------|----------|-----------|---------|------|
| Intercept | 27.90    | 0.51      | 54.23   | <.0001|
| gpa       | 0.88     | 0.13      | 6.56    | <.0001|

Suprious (or Chain)?

Model: \( Y_i = \alpha + \beta_2(peer) + \beta_1(gpa) + \epsilon \rightarrow R^2 = .36 \)

| Parameter | Estimate | Std Error | t Value | Pr > |t|
|-----------|----------|-----------|---------|------|
| Intercept | 37.54    | 0.55      | 68.23   | <.0001|
| peer      | -0.46    | 0.02      | -26.94  | <.0001|
| gpa       | 0.17     | 0.11      | 1.51    | 0.1302|
# Multiple “Causes”

<table>
<thead>
<tr>
<th></th>
<th>estimate</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>38.10</td>
<td>27.35</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>parent</td>
<td>0.15</td>
<td>4.29</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>age</td>
<td>−0.41</td>
<td>−8.77</td>
<td>&lt; .01</td>
</tr>
<tr>
<td>peer</td>
<td>−0.41</td>
<td>−24.51</td>
<td>&lt; .01</td>
</tr>
</tbody>
</table>

\[
\hat{Y}_i = 38.10 + .15\text{parent}_i - .41\text{age}_i - .41\text{peer}_i
\]
Youth = parent + age + peer

Youth by Age (age = 13.85)

Youth by Peer (parent = 33.1)

Youth by Peer (peer = 15.2)
Youth = parent + age + peer

---

**A Data Set**

**Causal Relationships**

**Types of Multivariate Relationships**

**Multiple Regression**

Interpreting $b_j$

- Interpreting $b_j$
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Another Model
- Multiple "Causes"
- Youth = parent + age + peer
- Youth = parent + age + peer
- Interaction
- Youth = parent + age + peer
  + (age)(peer)
- Youth = parent + age + peer
  + (age)(peer)
- Interaction

ANOVA, $R^2$, & Sums of Squares

---

Multivariate Relationships and Multiple Linear Regression
## Interaction

### Multiple “Causes”

<table>
<thead>
<tr>
<th></th>
<th>estimate</th>
<th>$t$</th>
<th>$p$</th>
<th></th>
<th>estimate</th>
<th>$t$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>38.10</td>
<td>27.35</td>
<td>&lt;.01</td>
<td>31.83</td>
<td>13.68</td>
<td>&lt;.01</td>
<td></td>
</tr>
<tr>
<td>parent</td>
<td>0.15</td>
<td>4.29</td>
<td>&lt;.01</td>
<td>0.15</td>
<td>4.22</td>
<td>&lt;.01</td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>-0.41</td>
<td>-8.77</td>
<td>&lt;.01</td>
<td>0.04</td>
<td>0.30</td>
<td>&lt;.77</td>
<td></td>
</tr>
<tr>
<td>peer</td>
<td>-0.41</td>
<td>-24.51</td>
<td>&lt;.01</td>
<td>0.03</td>
<td>0.19</td>
<td>&lt;.85</td>
<td></td>
</tr>
<tr>
<td>age*peer</td>
<td></td>
<td></td>
<td></td>
<td>-0.03</td>
<td>-3.36</td>
<td>&lt;.01</td>
<td></td>
</tr>
</tbody>
</table>

$$\hat{Y}_i = [31.83 + 0.04(age)_i] + 0.15(parent)_i + [0.03 - 0.03(age)_i](peer)_i$$

$$\hat{Y}_i = [31.83 + 0.03(peer)_i] + 0.15(parent)_i + [0.04 - 0.03(peer)_i](age)_i$$

### Indirect and Direct effects?
A Data Set

Causal Relationships

Types of Multivariate Relationships

Multiple Regression

Interpreting $b_j$

- Interpreting $b_j$
- Interpreting $b_j$
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Another Model
- Multiple "Causes"
- Youth = parent + age + peer
- Youth = parent + age + peer
- Interaction
- Youth = parent + age + peer + (age)(peer)
- Youth = parent + age + peer + (age)(peer)
- Interaction

ANOVA, $R^2$, & Sums of Squares

Multivariate Relationships and Multiple

Youth = parent + age + peer + (age)(peer)
A Data Set

Causal Relationships

Types of Multivariate Relationships

Multiple Regression

Interpreting $b_j$

- Interpreting $b_j$
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Youth = age + peer
- NYLS: Another Model
- Multiple "Causes"
- Youth = parent + age + peer
- Youth = parent + age + peer
- Interaction
- Youth = parent + age + peer + (age)(peer)
- Youth = parent + age + peer + (age)(peer)
- Interaction

ANOVA, $R^2$, & Sums of Squares

Multivariate Relationships and Multiple Regression Slide 30 of 37

Youth = parent + age + peer + (age)(peer)

$Y = a + b_1 \cdot \text{parent} + b_2 \cdot \text{age} + b_3 \cdot \text{peer} + b_4 \cdot \text{age} \cdot \text{peer}$

Youth by Peer (Parent = 33.31)

Youth by Age (Parent = 33.31)
Interaction

Why did the results change so much when we added the interaction?

“Multicollinearity”

<table>
<thead>
<tr>
<th></th>
<th>parent</th>
<th>peer</th>
<th>age</th>
<th>peer X age</th>
</tr>
</thead>
<tbody>
<tr>
<td>parent</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>peer</td>
<td>−.05</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>−.09</td>
<td>0.32</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>peer X age</td>
<td>−.07</td>
<td>0.93</td>
<td>0.59</td>
<td>1.00</td>
</tr>
</tbody>
</table>
ANOVA, $R^2$, & Sums of Squares

- Doing multiple regression in SAS (PROC GLM).

- ANOVA: model vs error

- $R^2 = \text{proportion of variance accounted for by the model}$

  \[
  R^2 = \frac{SS_{\text{total}} - SS_{\text{error}}}{SS_{\text{total}}} = \frac{SS_{\text{model}}}{SS_{\text{total}}} = r(Y_i, \hat{Y}_i)^2
  \]

- Sums of Squares for effects
  - Type III (standard regression interpretation)
  - Type I (hierarchical regression)
There are 4 possibilities:

- **Type I** \(SS\).
- **Type II** \(SS\) (briefly when we talk about factorial ANOVA).
- **Type III** \(SS\).
- **Type IV** \(SS\) (won’t talk about this one).
Type III Sums of Squares

**Regression Strategy:** Compute each sums of squares adjusting for everything else in the model.

- $SS_{X_1 | X_2, X_3} = t$.
- $SS_{X_2 | X_1, X_3} = z$.
- $SS_{X_3 | X_1, X_2} = x$.

- Can make decisions for the hypothesis regardless of order, etc.
- This goes along with interpretation we’ve talked about.
**Type I Sums of Squares**

**Hierarchical Method**

- First compute $SS_{X_1}$ (for example) in the usual way
  
  $SS_{X_1} = t + u + v + w$

- Next compute $SS_{X_2}$ adjusting for factor $X_1$,
  
  $SS_{X_2|X_1} = z + y$

- Compute $SS_{X_3}$ adjusting for factors $X_1$ and $X_2$
  
  $SS_{X_3|X_1,X_2} = x$

- Order matters!
Qualitative Explanatory Variables?

Recall the NELS example:

Two Different Schools

(sch_id:24725) Math = 34.39382 + 5.592664*(Homework)

(sch_id:62821) Math = 59.11331 + 1.133278*(Homework)
**NELS model with interaction**

Define

\[ \text{school} = \begin{cases} 
0 & \text{if school is } #24725 \\
1 & \text{if school is } #62821 
\end{cases} \]

\[ R^2 = 0.7564 \]

| Variable         | DF | Parameter Estimate | Standard Error | t Value | Pr>|t| |
|------------------|----|--------------------|----------------|---------|-------|
| Intercept        | 1  | 34.39382           | 1.75119        | 19.64   | < .0001 |
| homework         | 1  | 5.59266            | 0.79780        | 7.01    | < .0001 |
| school           | 1  | 24.71949           | 2.29151        | 10.79   | < .0001 |
| homework*school  | 1  | -4.45939           | 0.89502        | -4.98   | < .0001 |

Fitted model:

School #24725 : \( \hat{\text{math}}_i = 34.39 + 5.59(\text{homework})_i \)

School #62821 : \( \hat{\text{math}}_i = (34.39 + 24.72) + (5.59 - 4.46)(\text{homework})_i \)

\[ = 59.11 + 1.13(\text{homework})_i \]