Chapter learning objectives

• Compute and interpret partial correlation coefficients
• Find and interpret the least-squares multiple regression equation with partial slopes
• Find and interpret standardized partial slopes or beta-weights ($b^*$)
• Calculate and interpret the coefficient of multiple determination ($R^2$)
• Explain the limitations of partial and regression analysis
Multiple regression

• Discuss ordinary least squares (OLS) multiple regressions
  – OLS: linear regression
  – Multiple: at least two independent variables

• Disentangle and examine the separate effects of the independent variables

• Use all of the independent variables to predict $Y$

• Assess the combined effects of the independent variables on $Y$
Partial correlation

• Partial correlation measures the correlation between X and Y, controlling for Z

• Comparing the bivariate (zero-order) correlation to the partial (first-order) correlation
  – Allows us to determine if the relationship between X and Y is direct, spurious, or intervening
  – Interaction cannot be determined with partial correlations
Formula for partial correlation

• Formula for partial correlation coefficient for X and Y, controlling for Z

\[ r_{yx.z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2} \sqrt{1 - r_{xz}^2}} \]

• We must first calculate the zero-order coefficients between all possible pairs of variables (Y and X, Y and Z, X and Z) before solving this formula
Example

- Husbands’ hours of housework per week ($Y$)
- Number of children ($X$)
- Husbands’ years of education ($Z$)

Scores on Three Variables for 12 Dual-Wage-Earner Families

<table>
<thead>
<tr>
<th>Family</th>
<th>Husband's Housework ($Y$)</th>
<th>Number of Children ($X$)</th>
<th>Husband's Years of Education ($Z$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>E</td>
<td>3</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>F</td>
<td>1</td>
<td>2</td>
<td>16</td>
</tr>
<tr>
<td>G</td>
<td>5</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>H</td>
<td>0</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>I</td>
<td>6</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>J</td>
<td>3</td>
<td>4</td>
<td>12</td>
</tr>
<tr>
<td>K</td>
<td>7</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>L</td>
<td>4</td>
<td>5</td>
<td>16</td>
</tr>
</tbody>
</table>

Correlation matrix

• The bivariate (zero-order) correlation between husbands’ housework and number of children is +0.50

  – This indicates a positive relationship

<table>
<thead>
<tr>
<th>Zero-Order Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Husband’s Housework (Y)</strong></td>
</tr>
<tr>
<td>--------------------------</td>
</tr>
<tr>
<td><strong>Husband’s Housework (Y)</strong></td>
</tr>
<tr>
<td><strong>Number of Children (X)</strong></td>
</tr>
<tr>
<td><strong>Husband’s Years of Education (Z)</strong></td>
</tr>
</tbody>
</table>
First-order correlation

Calculate the partial (first-order) correlation between husbands’ housework $(Y)$ and number of children $(X)$, controlling for husbands’ years of education $(Z)$

\[
 r_{y|x,z} = \frac{r_{yx} - (r_{yz})(r_{xz})}{\sqrt{1 - r_{yz}^2} \sqrt{1 - r_{xz}^2}}
\]

\[
 r_{y|x,z} = \frac{(0.50) - (-0.30)(-0.47)}{\sqrt{1 - (-0.30)^2} \sqrt{1 - (-0.47)^2}}
\]

\[
 r_{y|x,z} = 0.43
\]
Interpretation

• Comparing the bivariate correlation (+0.50) to the partial correlation (+0.43) finds little change.

• The relationship between number of children and husbands’ housework has not changed, controlling for husbands’ education.

• Therefore, we have evidence of a direct relationship.
Bivariate & multiple regressions

• Bivariate regression equation

\[ Y = a + bX = \beta_0 + \beta_1X \]

− \( a = \beta_0 \) = \( Y \) intercept
− \( b = \beta_1 \) = slope

• Multivariate regression equation

\[ Y = a + b_1X_1 + b_2X_2 = \beta_0 + \beta_1X_1 + \beta_2X_2 \]

− \( b_1 = \beta_1 \) = partial slope of the linear relationship between the first independent variable and \( Y \)
− \( b_2 = \beta_1 \) = partial slope of the linear relationship between the second independent variable and \( Y \)
Multiple regression

\[ Y = a + b_1 X_1 + b_2 X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \]

- \( a = \beta_0 \) = the \( Y \) intercept, where the regression line crosses the \( Y \) axis
- \( b_1 = \beta_1 \) = partial slope for \( X_1 \) on \( Y \)
  - \( \beta_1 \) indicates the change in \( Y \) for one unit change in \( X_1 \), controlling for \( X_2 \)
- \( b_2 = \beta_2 \) = partial slope for \( X_2 \) on \( Y \)
  - \( \beta_2 \) indicates the change in \( Y \) for one unit change in \( X_2 \), controlling for \( X_1 \)
Partial slopes

• The partial slopes indicate the effect of each independent variable on Y

• While controlling for the effect of the other independent variables

• This control is called *ceteris paribus*
  – Other things equal
  – Other things held constant
  – All other things being equal
Ceteris paribus

Income = \beta_0 + \beta_1 \text{education} + \beta_2 \text{experience} + u
Ceteris paribus

Income = \( \beta_0 + \beta_1 \text{education} + \beta_2 \text{experience} + u \)

These two effects will result on \( \beta_1 \)
Ceteris paribus

\[ \text{Income} = \beta_0 + \beta_1 \text{education} + \beta_2 \text{experience} + u \]
Ceteris paribus

\[ Income = \beta_0 + \beta_1 \text{education} + \beta_2 \text{experience} + u \]

These two effects will result on $\beta_1$

These two effects will result on $\beta_2$
Interpretation of partial slopes

- The partial slopes show the effects of the $X$’s in their original units

- These values can be used to predict scores on $Y$

- Partial slopes must be computed before computing the $Y$ intercept ($\beta_0$)
Formulas of partial slopes

\[ b_1 = \beta_1 = \left( \frac{s_y}{s_1} \right) \left( \frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2} \right) \]

\[ b_2 = \beta_2 = \left( \frac{s_y}{s_2} \right) \left( \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \right) \]

- \( b_1 = \beta_1 = \) partial slope of \( X_1 \) on \( Y \)
- \( b_2 = \beta_2 = \) partial slope of \( X_2 \) on \( Y \)
- \( s_y = \) standard deviation of \( Y \)
- \( s_1 = \) standard deviation of the first independent variable (\( X_1 \))
- \( s_2 = \) standard deviation of the second independent variable (\( X_2 \))
- \( r_{y1} = \) bivariate correlation between \( Y \) and \( X_1 \)
- \( r_{y2} = \) bivariate correlation between \( Y \) and \( X_2 \)
- \( r_{12} = \) bivariate correlation between \( X_1 \) and \( X_2 \)
Formula of constant

- Once $b_1$ ($\beta_1$) and $b_2$ ($\beta_2$) have been calculated, use those values to calculate the Y intercept

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

\[ \beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2 \]
Example

• Using information below, calculate the slopes

<table>
<thead>
<tr>
<th>Husband’s Housework</th>
<th>Number of Children</th>
<th>Husband’s Education</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\overline{Y} = 3.3$</td>
<td>$\overline{X}_1 = 2.7$</td>
<td>$\overline{X}_2 = 13.7$</td>
</tr>
<tr>
<td>$s_y = 2.1$</td>
<td>$s_1 = 1.5$</td>
<td>$s_2 = 2.6$</td>
</tr>
</tbody>
</table>

Zero-Order Correlations

$r_{y1} = 0.50$
$r_{y2} = -0.30$
$r_{12} = -0.47$

Result and interpretation of $b_1$

\[
b_1 = \beta_1 = \left(\frac{s_y}{s_1}\right) \left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)
\]

\[
b_1 = \beta_1 = \left(\frac{2.1}{1.5}\right) \left(\frac{0.50 - (-0.30)(-0.47)}{1 - (-0.47)^2}\right) = 0.65
\]

- As the number of children in a dual-career household increases by one, the husband’s hours of housework per week increases on average by 0.65 hours (about 39 minutes), controlling for husband’s education.
Result and interpretation of $b_2$

$$b_2 = \beta_2 = \left( \frac{S_y}{s_2} \right) \left( \frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2} \right)$$

$$b_2 = \beta_2 = \left( \frac{2.1}{2.6} \right) \left( \frac{-0.30 - (0.50)(-0.47)}{1 - (-0.47)^2} \right) = -0.07$$

• As the husband’s years of education increases by one year, the number of hours of housework per week decreases on average by 0.07 (about 4 minutes), controlling for the number of children
Result and interpretation of $a$

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

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - \beta_2 \bar{X}_2$$

$$a = \beta_0 = 3.3 - (0.65)(2.7) - (-0.07)13.7$$

$$a = \beta_0 = 2.5$$

- With zero children in the family and a husband with zero years of education, that husband is predicted to complete 2.5 hours of housework per week on average
Final regression equation

• In this example, this is the final regression equation

\[ Y = a + b_1 X_1 + b_2 X_2 \]

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \]

\[ Y = 2.5 + (0.65)X_1 + (-0.07)X_2 \]

\[ Y = 2.5 + 0.65X_1 - 0.07X_2 \]
Prediction

• Use the regression equation to predict a husband’s hours of housework per week when he has 11 years of schooling and the family has 4 children

\[ Y' = 2.5 + 0.65X_1 - 0.07X_2 \]

\[ Y' = 2.5 + (0.65)(4) + (-0.07)(11) \]

\[ Y' = 4.3 \]

• Under these conditions, we would predict 4.3 hours of housework per week
Standardized coefficients ($b^*$)

- Partial slopes ($b_1=\beta_1 ; b_2=\beta_2$) are in the original units of the independent variables
  - This makes assessing relative effects of independent variables difficult when they have different units
  - It is easier to compare if we standardize to a common unit by converting to $Z$ scores

- Compute beta-weights ($b^*$) to compare relative effects of the independent variables
  - Amount of change in the standardized scores of $Y$ for a one-unit change in the standardized scores of each independent variable
    - While controlling for the effects of all other independent variables
  - They show the amount of change in standard deviations in $Y$ for a change of one standard deviation in each $X$
Formulas

• Formulas for standardized coefficients

\[ b_1^* = b_1 \left( \frac{s_1}{s_y} \right) = \beta_1^* = \beta_1 \left( \frac{s_1}{s_y} \right) \]

\[ b_2^* = b_2 \left( \frac{s_2}{s_y} \right) = \beta_2^* = \beta_2 \left( \frac{s_2}{s_y} \right) \]
Example

• Which independent variable, number of children ($X_1$) or husband’s education ($X_2$), has the stronger effect on husband’s housework in dual-career families?

$$b_1^* = b_1 \left( \frac{s_1}{s_y} \right) = (0.65) \left( \frac{1.5}{2.1} \right) = 0.46$$

$$b_2^* = b_2 \left( \frac{s_2}{s_y} \right) = (-0.07) \left( \frac{2.6}{2.1} \right) = -0.09$$

- The standardized coefficient for number of children (0.46) is greater in absolute value than the standardized coefficient for husband’s education (−0.09)

- Therefore, number of children has a stronger effect on husband’s housework
Standardized coefficients

• Standardized regression equation

\[ Z_y = a_z + b_1^*Z_1 + b_2^*Z_2 \]

– where \( Z \) indicates that all scores have been standardized to the normal curve

• The Y intercept will always equal zero once the equation is standardized

\[ Z_y = b_1^*Z_1 + b_2^*Z_2 \]

• For the previous example

\[ Z_y = (0.46)Z_1 + (-0.09)Z_2 \]
Multiple correlation

• The coefficient of multiple determination ($R^2$) measures how much of $Y$ is explained by all of the $X$'s combined

• $R^2$ measures the percentage of the variation in $Y$ that is explained by all of the independent variables combined

• The coefficient of multiple determination is an indicator of the strength of the entire regression equation

$$R^2 = r_{y1}^2 + r_{y2.1}^2 (1 - r_{y1}^2)$$

- $R^2 =$ coefficient of multiple determination
- $r_{y1}^2 =$ zero-order correlation between $Y$ and $X_1$
- $r_{y2.1}^2 =$ partial correlation of $Y$ and $X_2$, while controlling for $X_1$
Partial correlation of \( Y \) and \( X_2 \)

- Before estimating \( R^2 \), we need to estimate the partial correlation of \( Y \) and \( X_2 \) (\( r_{y2.1} \))

\[
r_{y2.1} = \frac{r_{y2} - (r_{y1})(r_{12})}{\sqrt{1 - r_{y1}^2} \sqrt{1 - r_{12}^2}}
\]

- We need three correlations
  - Between \( X_1 \) and \( Y \): 0.50
  - Between \( X_2 \) and \( Y \): -0.30
  - Between \( X_1 \) and \( X_2 \): -0.47

\[
r_{y2.1} = \frac{(-0.30) - (0.50)(-0.47)}{\sqrt{1 - (0.50)^2} \sqrt{1 - (-0.47)^2}}
\]

\[
r_{y2.1} = -0.08
\]
Result and interpretation

• For this example, $R^2$ will tell us how much of husband’s housework is explained by the combined effects of the number of children ($X_1$) and husband’s education ($X_2$)

$$R^2 = r_{y1}^2 + r_{y2.1}^2 (1 - r_{y1}^2)$$

$$R^2 = (0.50)^2 + (-0.08)^2 (1 - 0.50^2)$$

$$R^2 = 0.255$$

• Number of children and husband’s education explain 25.5% of the variation in husband’s housework
Normality assumption

• OLS regressions require normal distribution for its interval-ratio-level variables

• We can analyze histograms to determine if variables have a normal distribution
Histogram of income

Source: 2016 General Social Survey.
**Income = F(age, education)**

***Repondent's income by age and years of schooling

```
. svy: reg conrinc age educ if year==2016
(running regress on estimation sample)
```

Survey: Linear regression

|               | Coef.     | Std. Err. | t     | P>|t|  | [95% Conf. Interval] |
|---------------|-----------|-----------|-------|------|----------------------|
| conrinc       |           |           |       |      |                      |
| age           | 435.2034  | 46.70745  | 9.32  | 0.000| 341.9222 - 528.4846  |
| educ          | 4209.814  | 449.8191  | 9.36  | 0.000| 3311.464 - 5108.165  |
| _cons         | -43706    | 6198.866  | -7.05 | 0.000| -56085.99 - 31326.01 |

Source: 2016 General Social Survey.
Standardized coefficients

- Standardized regression coefficients
- (i.e., standardized partial slopes, beta-weights)
- It does not allow the use of GSS complex survey design

```
reg conrinc age educ if year==2016, beta
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 1,626</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2.5113e+11</td>
<td>2</td>
<td>1.2557e+11</td>
<td>F(2, 1623) = 109.52</td>
</tr>
<tr>
<td>Residual</td>
<td>1.8608e+12</td>
<td>1,623</td>
<td>1.1465e+09</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>2.1120e+12</td>
<td>1,625</td>
<td>1.2997e+09</td>
<td>R-squared = 0.1189</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.1178</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 33861</td>
</tr>
</tbody>
</table>

| conrinc   | Coef.       | Std. Err. | t     | P>|t| | Beta       |
|-----------|-------------|-----------|-------|------|------------|
| age       | 350.4332    | 59.38046  | 5.90  | 0.000| .1376138   |
| educ      | 3891.937    | 292.0824  | 13.32 | 0.000| .3107139   |
| _cons     | -35944.9    | 4885.588  | -7.36 | 0.000| .          |

Source: 2016 General Social Survey.
Power transformation

• Lawrence Hamilton ("Regression with Graphics", 1992, p.18–19)

\[
\begin{align*}
    Y^3 & \rightarrow q = 3 \\
    Y^2 & \rightarrow q = 2 \\
    Y^1 & \rightarrow q = 1 \\
    Y^{0.5} & \rightarrow q = 0.5 \\
    \log(Y) & \rightarrow q = 0 \\
    -(Y^{-0.5}) & \rightarrow q = -0.5 \\
    -(Y^{-1}) & \rightarrow q = -1
\end{align*}
\]

• $q > 1$: reduce concentration on the right (reduce negative skew)
• $q = 1$: original data
• $q < 1$: reduce concentration on the left (reduce positive skew)
• $\log(x+1)$ may be applied when $x = 0$. If distribution of $\log(x+1)$ is normal, it is called lognormal distribution
Histogram of log of income

Source: 2016 General Social Survey.
Interpretation of coefficients
(with continuous independent variables)

• With the logarithm of the dependent variable
  – Coefficients are interpreted as percentage changes

• If coefficient of $X_1$ equals 0.12
  – $\exp(\beta_1)$ times
    • $X_1$ increases by one unit, $Y$ increases on average **1.13 times**, controlling for other independent variables
  – $100*[\exp(\beta_1)–1]$ percent
    • $X_1$ increases by one unit, $Y$ increases on average by **13%**, controlling for other independent variables

• If coefficient has a small magnitude: $-0.3<\beta<0.3$
  – $100*\beta$ percent
    • $X_1$ increases by one unit, $Y$ increases on average approximately by **12%**, controlling for other independents
log income = F(age, education)

. ***Log of respondent's income by age and years of schooling
. svy: reg ln inconrinc age educ if year==2016
(running regress on estimation sample)

Survey: Linear regression

Number of strata = 65
Number of PSUs = 130
Number of obs = 1,626
Population size = 1,688.1407
Design df = 65
F(2,64) = 99.50
Prob > F = 0.0000
R-squared = 0.1327

| ln inconrinc | Coef. | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|--------------|-------|-----------|-------|------|----------------------|
| age          | 0.0157926 | 0.0021115 | 7.48  | 0.000 | 0.0115756 - 0.0200096 |
| educ         | 0.1229175 | 0.0109476 | 11.23 | 0.000 | 0.1010536 - 0.1447814 |
| _cons        | 7.506136 | 0.1702918 | 44.08 | 0.000 | 7.16604 - 7.846233    |

Source: 2016 General Social Survey.
Interpretation of example (with continuous independent variables)

• Coefficient for **age** equals 0.016
  
  – \( \exp(\beta_1) \) times
    
    • When age increases by one unit, income increases on average by **1.0161 times**, controlling for education
  
  – 100*\([\exp(\beta_1)−1]\) percent
    
    • When age increases by one unit, income increases on average by **1.61%**, controlling for education
  
  – 100*\(\beta_1\) percent
    
    • When age increases by one unit, income increases on average approximately by **1.6%**, controlling for education
Standardized coefficients

. ***Standardized regression coefficients
. ****(i.e., standardized partial slopes, beta-weights)
. ***It does not allow the use of GSS complex survey design
. reg lnconrinc age educ if year==2016, beta

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs</th>
<th>F(2, 1623)</th>
<th>Prob &gt; F</th>
<th>R-squared</th>
<th>Adj R-squared</th>
<th>Root MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>240.448673</td>
<td>2</td>
<td>120.224336</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>1926.88361</td>
<td>1,623</td>
<td>1.18723574</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2167.33229</td>
<td>1,625</td>
<td>1.33374294</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| lnconrinc  | Coef.    | Std. Err. | t     | P>|t|  | Beta       |
|------------|----------|-----------|-------|------|------------|
| age        | .0118659 | .0019108  | 6.21  | 0.000 | .1454577   |
| educ       | .1179159 | .009399  | 12.55 | 0.000 | .2938647   |
| _cons      | 7.752344 | .1572141 | 49.31 | 0.000 |         |
Dummy variables

- Many variables that are important in social life are nominal-level variables
  - They cannot be included in a regression equation or correlational analysis (e.g., sex, race/ethnicity)
- We can create dummy variables
  - Two categories, one coded as 0 and the other as 1

<table>
<thead>
<tr>
<th>Sex</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Race/ethnicity</th>
<th>White</th>
<th>Black</th>
<th>Hispanic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Age in interval-ratio level

• Age does not have a normal distribution

• Generate age group variable (categorical)
  – 18–24; 25–34; 35–49; 50–64; 65+

Source: 2016 General Social Survey.
Age in ordinal level

- Age has five categories

\[
\text{. egen agegr = cut(age), at(18,25,35,50,65,90)}
\]

(22 missing values generated)

\[
\text{. table agegr, contents(min age max age count age)}
\]

<table>
<thead>
<tr>
<th>agegr</th>
<th>min(age)</th>
<th>max(age)</th>
<th>N(age)</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>18</td>
<td>24</td>
<td>671</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>34</td>
<td>1,452</td>
</tr>
<tr>
<td>35</td>
<td>35</td>
<td>49</td>
<td>2,116</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>64</td>
<td>2,022</td>
</tr>
<tr>
<td>65</td>
<td>65</td>
<td>89</td>
<td>1,440</td>
</tr>
</tbody>
</table>

- Generate dummy variables for age...

Source: 2016 General Social Survey.
Dummies for age

• Generate dummy variables for age group

<table>
<thead>
<tr>
<th>Age group</th>
<th>Age 18–24</th>
<th>Age 25–35</th>
<th>Age 35–49</th>
<th>Age 50–64</th>
<th>Age 65–89</th>
</tr>
</thead>
<tbody>
<tr>
<td>18–24</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25–34</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>35–49</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>50–64</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>65–89</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

• Use the category with largest sample size as the reference (35–49)

Source: 2016 General Social Survey.
Education in interval-ratio level

- Education does not have a normal distribution

- Use the categorical education variable (degree)
  - Less than high school; high school; junior college; bachelor; graduate

Source: 2016 General Social Survey.
Education in ordinal level

- Education has five categories

```
. tab degree

<table>
<thead>
<tr>
<th>rs highest degree</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>lt high school</td>
<td>997</td>
<td>12.92</td>
<td>12.92</td>
</tr>
<tr>
<td>high school</td>
<td>3,897</td>
<td>50.52</td>
<td>63.44</td>
</tr>
<tr>
<td>junior college</td>
<td>585</td>
<td>7.58</td>
<td>71.03</td>
</tr>
<tr>
<td>bachelor</td>
<td>1,418</td>
<td>18.38</td>
<td>89.41</td>
</tr>
<tr>
<td>graduate</td>
<td>817</td>
<td>10.59</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>7,714</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
```

- Generate dummy variables for education...

Source: 2016 General Social Survey.
### Dummies for education

#### Generate dummy variables for education group

<table>
<thead>
<tr>
<th>Education group</th>
<th>&lt;HS</th>
<th>HS</th>
<th>JC</th>
<th>BA</th>
<th>GR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than high school</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>High school</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Junior college</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Bachelor</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Graduate</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

#### Use the category with largest sample size as the reference (HS)

<table>
<thead>
<tr>
<th>rs highest degree</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>lt high school</td>
<td>997</td>
<td>12.92</td>
<td>12.92</td>
</tr>
<tr>
<td>high school</td>
<td>3,897</td>
<td>50.52</td>
<td>63.44</td>
</tr>
<tr>
<td>junior college</td>
<td>585</td>
<td>7.58</td>
<td>71.03</td>
</tr>
<tr>
<td>bachelor</td>
<td>1,418</td>
<td>18.38</td>
<td>89.41</td>
</tr>
<tr>
<td>graduate</td>
<td>817</td>
<td>10.59</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>7,714</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>
log income = F(age, education)

.svy: reg lnconrinc agegr1 agegr2 agegr4 agegr5 educgr1 educgr3 educgr4 educgr5 if year==2016
(running regress on estimation sample)

Survey: Linear regression

<table>
<thead>
<tr>
<th></th>
<th>Linearized</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef.</td>
<td>Std. Err.</td>
<td>t</td>
<td>P&gt;</td>
<td>t</td>
<td></td>
</tr>
<tr>
<td>lnconrinc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>agegr1</td>
<td>-1.166963</td>
<td>.1220959</td>
<td>-9.56</td>
<td>0.000</td>
<td>-1.410805</td>
<td>-.9231207</td>
</tr>
<tr>
<td>agegr2</td>
<td>-.3345438</td>
<td>.0736023</td>
<td>-4.55</td>
<td>0.000</td>
<td>-.4815379</td>
<td>-.1875498</td>
</tr>
<tr>
<td>agegr4</td>
<td>-.0050007</td>
<td>.0638917</td>
<td>-0.08</td>
<td>0.938</td>
<td>-.1326013</td>
<td>.1225999</td>
</tr>
<tr>
<td>agegr5</td>
<td>-.4155278</td>
<td>.096474</td>
<td>-4.31</td>
<td>0.000</td>
<td>-.6081997</td>
<td>-.2228559</td>
</tr>
<tr>
<td>educgr1</td>
<td>-.4276264</td>
<td>.1163403</td>
<td>-3.68</td>
<td>0.000</td>
<td>-.6599739</td>
<td>-.1952789</td>
</tr>
<tr>
<td>educgr3</td>
<td>.2367316</td>
<td>.0940649</td>
<td>2.52</td>
<td>0.014</td>
<td>.0488711</td>
<td>.4245921</td>
</tr>
<tr>
<td>educgr4</td>
<td>.4559903</td>
<td>.0843136</td>
<td>5.41</td>
<td>0.000</td>
<td>.2876045</td>
<td>.6243761</td>
</tr>
<tr>
<td>educgr5</td>
<td>.8516728</td>
<td>.0920326</td>
<td>9.25</td>
<td>0.000</td>
<td>.667871</td>
<td>1.035475</td>
</tr>
<tr>
<td>_cons</td>
<td>9.949482</td>
<td>.0471336</td>
<td>211.09</td>
<td>0.000</td>
<td>9.855349</td>
<td>10.04361</td>
</tr>
</tbody>
</table>

R-squared         = 0.1982
Prob > F          = 0.0000
F(  8,   58)      = 53.49
Design df         = 65
Number of obs     = 1,626
Population size   = 1,688.1407
Number of strata  = 65
Number of PSUs    = 130

Source: 2016 General Social Survey.
Interpretation of example
(with dummies as independent variables)

• High school is reference category for education
• Coefficient for junior college equals 0.237
  - $\exp(\beta_1)$ times
    • People with junior college degree have on average earnings 1.27 times higher than earnings of high school graduates, controlling for the other independent variables
  - 100*$[\exp(\beta_1)-1]$ percent
    • People with junior college degree have on average earnings 27% higher than earnings of high school graduates, controlling for the other independent variables
  - 100*$\beta_1$ percent
    • People with junior college degree have on average earnings approximately 23.7% higher than earnings of high school graduates, controlling for the other independent variables
Interpretation of example
(with dummies as independent variables)

• 35–49 age group is reference category for age
• Coefficient for 18–24 age group equals –1.167
  – \( \exp(\beta_1) \) times
    • People between 18 and 24 years of age have on average earnings 0.31 times the earnings of people between 35 and 49 years of age, controlling for the other independent variables
  – 100*[\exp(\beta_1)–1] percent
    • People between 18 and 24 years of age have on average earnings 69% lower than earnings of people between 35 and 49 years of age, controlling for the other independent variables
  – 100*\( \beta_1 \) percent: result is not good because the magnitude is high
    • People between 18 and 24 years of age have on average earnings approximately 117% lower than high school graduates, controlling for the other independent variables
Standardized coefficients

```
. reg lnconrinc agegr1 agegr2 agegr4 agegr5 educgr1 educgr3 educgr4 educgr5 if year==2016, beta

Source | SS       | df | MS        | Number of obs = 1,626
--------|----------|----|-----------|-------------------------
Model   | 395.582999 | 8  | 49.4478749 | F(8, 1617) = 45.13
Residual| 1771.74929  | 1,617 | 1.09570148 | Prob > F = 0.0000
Total   | 2167.33229  | 1,625 | 1.33374294 | R-squared = 0.1825
         |           |    |           | Adj R-squared = 0.1785
         |           |    |           | Root MSE = 1.0468

| lnconrinc | Coef. | Std. Err. | t    | P>|t| | Beta |
|-----------|-------|-----------|------|------|------|
| agegr1    | -1.133145 | .1077317 | -10.52 | 0.000 | -.2566052 |
| agegr2    | -.3035091 | .072592 | -4.18 | 0.000 | -.1091483 |
| agegr4    | .0151364 | .0656439 | 0.23 | 0.818 | .0061022 |
| agegr5    | -.4761732 | .103716 | -4.59 | 0.000 | -.1110716 |
| educgr1   | -.4255014 | .1019697 | -4.17 | 0.000 | -.0970916 |
| educgr3   | .2160742 | .0977414 | 2.21 | 0.027 | .0516399 |
| educgr4   | .5121465 | .0675632 | 7.58 | 0.000 | .1828828 |
| educgr5   | .7994411 | .0810967 | 9.86 | 0.000 | .2359353 |
| _cons     | 9.927145 | .0541188 | 183.43 | 0.000 | . |
```

Source: 2016 General Social Survey.
Table 1. Coefficients and standard errors estimated with ordinary least squares models for the logarithm of respondent’s income as the dependent variable, U.S. adult population, 2004, 2010, and 2016

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>2004 Coefficients</th>
<th>Standardized coefficients</th>
<th>2010 Coefficients</th>
<th>Standardized coefficients</th>
<th>2016 Coefficients</th>
<th>Standardized coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>10.030***</td>
<td>(0.063)</td>
<td>9.919***</td>
<td>(0.090)</td>
<td>9.949***</td>
<td>(0.047)</td>
</tr>
<tr>
<td>Age groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18–24</td>
<td>-1.114***</td>
<td>(-0.104)</td>
<td>-1.438***</td>
<td>(0.188)</td>
<td>-1.167***</td>
<td>(0.122)</td>
</tr>
<tr>
<td>25–34</td>
<td>-0.306***</td>
<td>(0.074)</td>
<td>-0.406***</td>
<td>(0.102)</td>
<td>-0.335***</td>
<td>(0.074)</td>
</tr>
<tr>
<td>35–49</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
</tr>
<tr>
<td>50–64</td>
<td>0.132*</td>
<td>(0.068)</td>
<td>0.043</td>
<td>(0.092)</td>
<td>-0.005</td>
<td>(0.064)</td>
</tr>
<tr>
<td>65+</td>
<td>-0.596***</td>
<td>(0.165)</td>
<td>-0.720***</td>
<td>(0.175)</td>
<td>-0.416***</td>
<td>(0.097)</td>
</tr>
<tr>
<td>Education groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than high school</td>
<td>-0.410***</td>
<td>(0.117)</td>
<td>-0.477***</td>
<td>(0.125)</td>
<td>-0.428***</td>
<td>(0.116)</td>
</tr>
<tr>
<td>High school</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
<td>ref.</td>
</tr>
<tr>
<td>Junior college</td>
<td>0.276***</td>
<td>(0.097)</td>
<td>0.142</td>
<td>(0.122)</td>
<td>0.237**</td>
<td>(0.094)</td>
</tr>
<tr>
<td>Bachelor</td>
<td>0.620***</td>
<td>(0.062)</td>
<td>0.579***</td>
<td>(0.099)</td>
<td>0.456***</td>
<td>(0.084)</td>
</tr>
<tr>
<td>Graduate</td>
<td>0.785***</td>
<td>(0.097)</td>
<td>0.983***</td>
<td>(0.088)</td>
<td>0.852***</td>
<td>(0.092)</td>
</tr>
<tr>
<td>R²</td>
<td>0.242</td>
<td>0.222</td>
<td>0.288</td>
<td>0.272</td>
<td>0.198</td>
<td>0.183</td>
</tr>
<tr>
<td>Number of observations</td>
<td>1,685</td>
<td>1,685</td>
<td>1,201</td>
<td>1,201</td>
<td>1,626</td>
<td>1,626</td>
</tr>
</tbody>
</table>

Note: Coefficients and standard errors were generated with the complex survey design of the General Social Survey. The standardized coefficients were generated without the complex survey design. Standard errors are reported in parentheses. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Limitations

• Multiple regression and correlation are among the most powerful techniques available to researchers
  – But powerful techniques have high demands

• These techniques require
  – Every variable is measured at the interval-ratio level
  – Each independent variable has a linear relationship with the dependent variable
  – Independent variables do not interact with each other
  – Independent variables are uncorrelated with each other
  – When these requirements are violated (as they often are), these techniques will produce biased and/or inefficient estimates
  – There are more advanced techniques available to researchers that can correct for violations of these requirements