• Begin with:

• Types of multivariate relationships

• 1. spurious relationships
• 2. chain relationships
• 3. multiple causes
• 4. suppressor variables
• 5. interaction terms

• Move onto introduce:
  Linear regression
• Logistic regression..
• Spurious relationship:
• two variables are associated, yet there is absolutely no causal relationship between the two
• Spurious relationship:
  • two variables are associated, yet there is absolutely no causal relationship between the two

• Example: shoe size and mortality rate

• Initially, we hypothesize (foolishly)
  • shoe size $\rightarrow$ mortality rate

• Gender

Original association is spurious (no causality at all!!!)
1. Spuriousness

- In this second relationship (causal chain), $X_1$ has an "indirect" effect on $Y$ through an intervening variable $X_2$.
- $X_2$ had a "direct" effect on $Y$ in this chain relationship.

2. Chain relationships

- Unlike the spurious relationship, both $X_1$ and $X_2$ are relevant in the causal explanation.
• Note:
• For “spurious” relationships
• The initial relationship between $X_1$ and $Y$ should disappear or be seriously weakened (other hidden confounding variables might remain)

• Consumption of Timbits

Success as a hockey player

Typically, this will not be our concern
For “chain relationships”
Original relationship will likely be modified when we introduce control
Typically it will be weakened, but not necessarily, and it might disappear

Life style differences
  e.g. smoking

Mortality
3. Suppressor variables

- Originally no relationship documented (or perhaps a weak association)
- When we introduce control -> relationship appears or becomes stronger

- First consider the following bivariate relationship, income and age

  Note: Unique subsample (factory floor)
Contingency Table Relating Education and Income

<table>
<thead>
<tr>
<th>Education</th>
<th>Income</th>
<th></th>
<th></th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>High</td>
<td>250</td>
<td>50.0%</td>
<td>250</td>
</tr>
<tr>
<td>Low</td>
<td>High</td>
<td>250</td>
<td>50.0%</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td>Low</td>
<td>250</td>
<td>50.0%</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td></td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
</tbody>
</table>

No association is disclosed
Contingency Table Relating Education, Income and Age

<table>
<thead>
<tr>
<th>Income</th>
<th>Age = low</th>
<th>Age = High</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High income</td>
<td>Low income</td>
</tr>
<tr>
<td>High</td>
<td>125 35.7%</td>
<td>225 64.3%</td>
</tr>
<tr>
<td>Low</td>
<td>25 16.7%</td>
<td>125 83.3%</td>
</tr>
</tbody>
</table>

How could this happen???

Younger people earn less money than older people
Younger people tend to be better educated than older people

Consequently, it is only after we introduce a control for the effect of age, that the true relationship between income and education surfaces
• 3. Multiple Causes

- $X_1$  
- $X_2$  
- $X_3$  

- $Y$ earnings?

- We always work with this sort of model in sociology, as obviously social behavior is not explained by solely one variable
3. Multiple Causes

- $X_1 \rightarrow Y$
- $X_2 \rightarrow Y$
- $X_3$

We always work with this sort of model in sociology, as obviously social behavior is not explained by solely one variable.

Example:
- Age
- residence \rightarrow earnings
- ethnicity

Age, residence and ethnicity can all be understood as having “independent effects” of our dependent variable.
5. Interaction terms

When you enter a statistical control:
- The original bivariate association differs by category of the control variable

Example:
- Examining the relationship between education and income
- and how it differs by citizenship status
Contingency Table Relating Education, Income and Place of Birth

<table>
<thead>
<tr>
<th>Education</th>
<th>Foreign born</th>
<th></th>
<th></th>
<th>Canadian born</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>High income</td>
<td>Low income</td>
<td>Total</td>
<td>High income</td>
<td>Low income</td>
<td>Total</td>
</tr>
<tr>
<td>High</td>
<td>125</td>
<td>225</td>
<td>350</td>
<td>125</td>
<td>25</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>35.7%</td>
<td>64.3%</td>
<td></td>
<td>83.3%</td>
<td>16.7%</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>65</td>
<td>125</td>
<td>190</td>
<td>80</td>
<td>125</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>34.2%</td>
<td>65.8%</td>
<td></td>
<td>39.0%</td>
<td>61.0%</td>
<td></td>
</tr>
</tbody>
</table>

How do we summarize this relationship??

For foreign born -> no association between income and education
For Canadian born -> strong association exists between income and education

The effect of an independent variable on the dependent -> differs by category of the control variable
• In sum
• Through the introduction of control variables:
  • evidence for:
  • spuriousness
  • chain relationships
  • suppressor variables
  • statistical interaction

• Return to all this with regression procedures
Learning Objectives

• Understand when to use “linear” or “logistic” regression
• Understand the difference between simple and multiple regression, and how each are represented mathematically
• Recognize what is meant by “standardized slope” and “unstandardized slope” in simple and multiple regression.
• Recognize and interpret the significance test associated with the slope
• Be capable of interpreting “partial regression coefficients” in multiple regression (SPSS output with 2 or more independent variables)
Regression: Learning Objectives:

- Understand what is meant by linear regression – and when it is appropriate to use this procedure

In this course, we are learning about 2 types of regression:

1. OLS regression (ordinary least scared)
   - Also referred to as Linear regression

2. Logistic regression
   - Similar, but necessary when focusing on “dichotomous variables”..
• Simple regression - 1 independent variable
• Multiple regression – many independent variables
• Advantage over cross tabs..
• ->regression can handle many variables simultaneously..
• ->Can work with data in original form (e.g. interval/ratio)
• -> can document the relative strength of various associations
Regression analysis:

-> must specify dependent variable

independent variable

We decide upon an analysis of “illegal drug use” (as our dependent variable)

Regression procedures:
We can simultaneously consider the impact
Of many variables in explaining the dependent variable
Note: clarity of level of measurement is crucial in developing your models!!!
Nominal measurement N
Ordinal measurement O
Interval/ratio measurement I/R

“illicit drug use”..

“Jesse, have you used “illicit drugs over the last 12 months?” (yes, no) N

Assume we hope to understand whether men or women are more likely to use illicit drugs? Sex: Men/women N

Education?
“Jesse, how many years of schooling have you had? ____ I/R

Or:
“Jesse, what is your highest level of education?
Less than high school, high school, some college/university, college/univ grad O
• **Nominal** (can’t rank order persons and/or dichotomous)

• **Ordinal** (can rank order, but generally crude measurement)
  - e.g. Likert scale

• **Interval/Ratio** (rank order, precision with meaningful units)
  - eg. BMI, years of education, income (in dollars)

**IMPORTANT WHEN CONSIDERING YOUR DEPENDENT VARIABLE:**

• IF Interval/Ratio -> use Linear regression
• IF Ordinal with at least 6-7 categories, you can use Linear regression
• If Nominal or with less than 7 categories in your variable, you should use Logistic regression

• **IF UNCERTAIN!!! DOUBLE CHECK WITH ME**
• **Note: don’t worry about independent variables at this point..**
• We begin with Linear regression (Logistic regression not this lecture)
• We begin with Linear regression (Logistic regression not this lecture)

• What does linear regression do?
• Consider the following graph (unit of analysis CT: neighbourhood)

• It attempts to describe the nature of an association by fitting a straight line to represent it.
Dependent Variable $Y$

Independent Variable $X$

Positive association: As we move to CT’s with higher unemployment, we find higher property crime rates.

Any straight line can be represented by a simple equation

$$y = 2766.3 + 270.08x$$
• With Regression procedures

• Both “positive” and “negative” associations are possible

As the unemployment rate goes up, average income goes down..

Negative or Inverse relationship
Without getting into the technical details

The best fitting “regression line” is obtained via:

“the method of least squares”

Must satisfy the condition:

the difference between the observed values on Y and predicted values (regression line)

-> must be at a minimum
The formula:

\[ \hat{Y} = a + bX \]

*represents the regression line that best represents the data.*

\[ a = y \text{ intercept} \]
\[ b = \text{slope} \]
\[ \hat{Y} = \text{predicted value} \]
\[ Y \text{ represents your dependent variable or} \]
\[ X \text{ represents your independent variable} \]
Y intercept \((a)\) is the point on the Y axis when \(X = 0\)

In this case, when \(X = 0\), we would predict \(Y = \$40,300\)
In this example, our y intercept is about 1.5
What about the “slope” \((b)\)?
What about the “slope” (b)

**Slope:**

For each unit increment in X (i.e. for each additional hour worked)

**Y = 1.5 + 0.5X**

B = 0.5
Slope (b) for each unit increment in X
what increase do we expect in Y

Going up by 1% unemployment average income declines by about $2000 \ (b = -2000)

\[ Y = 40,300 - 2000 \times X \]
The formula:

\[ \hat{Y} = a + bX \]

regression line when working with a sample

\[ \hat{Y} \]
represents the values of Y along the straight line
predicted values of Y (not observed values)
Relationship between Distance and Airfare, with 12 cases
Relationship between Distance and Airfare, with 12 cases

The graph shows a positive correlation between distance and airfare, with data points indicating that as distance increases, airfare also tends to increase.
How to choose a line:

- In our example:
- with each case we obtain an error (residual): $e_i = (\hat{Y} - Y)$
• Solve for a and b,
• Must satisfy the condition:

• the difference between the observed values on Y and predicted values (regression line)
• -> these residuals must be at a minimum
How to compute regression line?
(don’t sweat it)

Formula for the slope:

\[ b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} \]

Formula for the y intercept:

\[ a = \bar{Y} - b\bar{X} \]

Should be able to interpret computer output

SPSS can produce above, as well as significance tests!!

Our estimates again have sampling error!!
• Example... cardio exercise (weekly basis) and pulse rate

<table>
<thead>
<tr>
<th>ID # persons</th>
<th>Hours of exercise</th>
<th>Pulse rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>92.0</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>65.0</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>52.0</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>80.0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>82.0</td>
</tr>
<tr>
<td>6</td>
<td>7</td>
<td>71.0</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>68.0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>92.0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>94.0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>88.0</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>50.0</td>
</tr>
<tr>
<td>12</td>
<td>6</td>
<td>75.0</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>66</td>
</tr>
</tbody>
</table>

1000s of cases, large sample
<table>
<thead>
<tr>
<th>pulse</th>
<th>brkfast</th>
<th>pet</th>
<th>roomates</th>
<th>movies</th>
<th>job</th>
<th>haircut</th>
</tr>
</thead>
<tbody>
<tr>
<td>70</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>11</td>
<td>5</td>
<td>45.00</td>
</tr>
<tr>
<td>78</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>8</td>
<td>2</td>
<td>15.00</td>
</tr>
<tr>
<td>64</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>1</td>
<td>14</td>
<td>120.00</td>
</tr>
<tr>
<td>84</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>49.00</td>
</tr>
<tr>
<td>84</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>6</td>
<td>65.00</td>
</tr>
<tr>
<td>80</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>9</td>
<td>75</td>
<td>9.00</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>29.00</td>
</tr>
<tr>
<td>72</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>150.00</td>
</tr>
<tr>
<td>80</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>34.00</td>
</tr>
<tr>
<td>72</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>78.00</td>
</tr>
<tr>
<td>79</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>20</td>
<td>4.00</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>pulse</th>
<th>brkfast</th>
<th>pet</th>
<th>roomates</th>
<th>movies</th>
<th>job</th>
<th>haircut</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5.57</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>5.32</td>
<td>1982</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>5.87</td>
<td>1976</td>
<td>2</td>
</tr>
<tr>
<td>21</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>5.17</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>5.50</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>23</td>
<td>1</td>
<td>5</td>
<td>6</td>
<td>6.00</td>
<td>1980</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>5.75</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>5.58</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>6</td>
<td>1</td>
<td>5.50</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>27</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>5.58</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>28</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>5.57</td>
<td>1983</td>
<td>2</td>
</tr>
<tr>
<td>29</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>5.75</td>
<td>1981</td>
<td>2</td>
</tr>
<tr>
<td>30</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>5.58</td>
<td>1983</td>
<td>2</td>
</tr>
</tbody>
</table>

---

**SPSS Processor is ready**

**Data View**

**Variable View**
**Coefficients**

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>76.500</td>
<td>2.300</td>
<td></td>
<td>33.268</td>
</tr>
<tr>
<td>EXERCIS</td>
<td>-1.149</td>
<td>.020</td>
<td>-.063</td>
<td>57.450</td>
</tr>
</tbody>
</table>

- **76.500** is our “Y intercept” (Constant)
- **-1.149** is our slope… (unstandardized)

**Slope (b) tells us:**
- for every 1 unit change in the
- independent variable (each additional hour of exercise)
- we expect a 1.149 unit change (reduction) in the dependent

**Y intercept (a) tells us:**
- when the independent variable is zero,
- our regression predicts that the dependent would be 76.5

**Significance Test**
- if less than 0.05 (great!!!!)
What is that Significance test about??..

Can we infer from the sample to the larger population?

Might your result merely be the by-product of "random sampling error"?

Is your result "statistically significant"??

A P-value greater than .05 suggests that your results may have merely been the by-product of sampling error and might not be true of the population.
• Return to previous example,
• Exercise and pulse rate (smaller sample)

• .. Random sample of about 60 individuals
• rather than several thousand...
Y intercept

Slope

Note: we don’t care about the significance test on the Y intercept (constant)

But here, P=0.650@!!

If not significant, we can not trust the slope, hence report relationship as “non-sign”
• I will skip over this slide quickly in the “lecture”, but for your information:

• Standard error (called STD ERROR on output):
  • the standard deviation of the sampling distribution (theoretical distribution used in significance tests).
  • As a general rule, the larger the sample, the smaller the STD ERROR

• In this example N=60 cases
  • SEs are very large... Slope = -1.149
  • SE = 2.446
  • Can’t even say that the slope is different from 0
  • Might in reality be a negative association????

• The sample size is just too small to say anything decisive..
  • If a statistic (in this case the slope) is more than 1.96 SE’s away from 0 (given a specific sample size), we can state that the slope is “significantly different from 0., i.e. not the by-product of sampling error (the t-value tells us how many SE’s the slope estimate is away from zero)
• Slope (unstandardized)
  “direction” and “character” of the relationship

• PROBLEM:

• But sometimes the “slope” is difficult to interpret..

• “Bigger isn’t always Better”

• A large slope doesn’t necessary imply a “stronger relationship”

• Why??

• Its value is intrinsically linked to the units of measurement of the variables involved
  • Example:
    • If we measured Age in months, rather than years, the slope would be smaller;
    • Or how do we compare the effect of age with education?
Researchers deal with this problem,
by “standardizing” the “unstandardized slope”

- Standardised -> ranges from -1 thru +1
- units of measurement are irrelevant

General rule in interpretation:
- Standardized slope < 0.1 weak association
- 0.1 < 0.4 moderate
- >= 0.4 strong
• With two variables:

- $s_y \to$ standard deviation of the variable Y
- $s_x \to$ standard deviation of the variable X

• Simply put:
  
  Pearson’s r is obtained as:

  $$r = \left( \frac{s_x}{s_y} \right) b$$

  With $b \to$ unstandardized slope in simple regression
• Computational formula:

\[
r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{N \sum X^2 - (\sum X)^2} \sqrt{N \sum Y^2 - (\sum Y)^2}}
\]

• Above all: you “must” be
• capable of interpreting it!!!
Standardized slope = -.063

For every 1 standard deviation change in your independent variable… you can expect a .063 standard deviation decline in your dependent variable.

Again, as a general rule:

<table>
<thead>
<tr>
<th>Standardized slope</th>
<th>weak association</th>
<th>moderate</th>
<th>strong</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; .1</td>
<td>.1 &lt; .4</td>
<td>&gt;= 4</td>
<td></td>
</tr>
</tbody>
</table>

a. Dependent Variable: PULSE
Multiple Regression

What is multiple regression?
• Involves working with several independent variables and one dependent variable
• .. a more complex model is needed
• This is where it gets more interesting...
Multiple Regression

• Regression allows us to examine the relationship between several “independent variables” and one dependent variable, simultaneously!!!
Multiple Regression

What is multiple regression?

• Involves working with several independent variables and one dependent variable

• .. a more complex model is needed:

\[ \hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_k X_k \]

Where we are working with a “sample” with:

- \( X_1, X_2, \ldots, X_k \) are independent variables
- \( b_1, b_2, \ldots, b_k \) are their respective slopes
Multiple Regression

• We can model the relationship between our dependent variable $Y$, and several independent variables

Purpose:
• Estimate the effect of each independent variable on the dependent variable, while controlling for all other variables
• i.e. the effect of $X_1$ on $Y$, controlling for all other independent variables...
• the effect of $X_2$ on $Y$, controlling for all others

• Provides more complete explanation of our DV, because several IVs are taken into account simultaneously
• The relationship between an IV and the DV when controlling for other variables is called a *partial* relationship
Multiple Regression

In working with SPSS software:

• Estimate and interpret all regression coefficients
• Determine the significance of each relationship: is the documented association significantly different from zero (p-value < .05)?

• Note: rule of thumb
• If the *unstandardized* slope is more than twice its SE (1.96), it’s not likely that the true population slope is equal to 0
• A slope of 0 means no relationship, so…
• An *unstandardized* slope that is 2X its SE means the association is statistically significant
Multiple Regression

Example of SPSS output:
• Using NLSCY
• Y → Depression score (1-35) for parents

• 2 independent variables:
• X₁ → social support (0-18)
• X₂ → family dysfunction (0-35 high score is a problem)

• What do we anticipate?
• Direct and Indirect effects

• Conduct three separate regressions:
  1) X₁ → Y
  2) X₂ → Y
  3) both variables simultaneously
Multiple Regression

Hypothesize causal chain (direct & indirect effects).

Social support → Depression

Social support

Family dysfunction

Depression

• We anticipate that the effect of social support would be modified by controlling for family functioning
Consider the following results from two simple regressions (SPSS)

1. “social support” ($X_1$) as the independent variable:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFNHS01 FAMILY FUNCTIONING SCORE</td>
<td>10.573</td>
<td>.185</td>
<td>-.392</td>
<td>.013</td>
</tr>
</tbody>
</table>

2. “family dysfunctioning” ($X_2$) as the independent variable:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Constant)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASPHS01 SOCIAL SUPPORT SCORE</td>
<td>1.951</td>
<td>.064</td>
<td>.358</td>
<td>.007</td>
</tr>
</tbody>
</table>

a. Dependent Variable: ADPPS01 DEPRESSION SCORE
Consider the following results from two simple regressions (SPSS)

1. “social support” \( (X_1) \) as the independent variable:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>10.573</td>
<td>.185</td>
</tr>
<tr>
<td>ASPHS01 SOCIAL SUPPORT SCORE</td>
<td>-.392</td>
<td>.013</td>
</tr>
</tbody>
</table>

a. Dependent Variable: ADPPS01 DEPRESSION SCORE

For every one unit increase on the social support scale, we predict -.392 lower score on the depression scale.

The direct effect of social support on depression is of “moderate” strength with a standardized slope of -0.205 (P<.001).
Consider the following results from two simple regressions (SPSS)

1. “social support” ($X_1$) as the independent variable:

2. “family dysfunctioning” ($X_2$) as the independent variable:

For every one unit increase on the family dysfunction scale, we predict +.358 higher score on the depression scale.

### Table 1

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>10.573</td>
<td>.185</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASPHS01 SOCIAL SUPPORT SCORE</td>
<td>-.392</td>
<td>.013</td>
<td>-.205</td>
<td>-31.280</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B</td>
<td>Std. Error</td>
<td>Beta</td>
<td></td>
</tr>
<tr>
<td>1 (Constant)</td>
<td>1.951</td>
<td>.064</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AFNHS01 FAMILY FUNCTIONING SCORE</td>
<td>.358</td>
<td>.007</td>
<td>342</td>
<td>54.112</td>
</tr>
</tbody>
</table>

a. Dependent Variable: ADPPS01 DEPRESSION SCORE

Significant

Moderate effect
3. Multivariate regression, with both social support and family functioning entered into the equation:

<table>
<thead>
<tr>
<th>Model</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B</td>
<td>Std. Error</td>
</tr>
<tr>
<td>(Constant)</td>
<td>3.121</td>
<td>.243</td>
</tr>
<tr>
<td>ASPHS01 SOCIAL SUPPORT SCORE</td>
<td>-.070</td>
<td>.014</td>
</tr>
<tr>
<td>AFNHS01 FAMILY FUNCTIONING SCORE</td>
<td>.338</td>
<td>.008</td>
</tr>
</tbody>
</table>

Don’t usually report Y intercept: Merely tells us that when social support and family dysfunction = 0, we predict a depression score of 3.12.

A 1 point increase in social support leads us to predict a .07 point decrease in depression.

A 1 point increase in family dysfunction leads to predict a .338 point increase in depression.

Family functioning has a much stronger impact on depression than social support.
Multiple Regression

When do we use standardized slopes ($b^*$)?
• When comparing the relative effect of different variables
• Which is most important in explaining $Y$?

What about unstandardized slopes ($b$)?
• A more meaningful interpretation
• Direct interpretation within the initial metric of variables
• Useful in comparing variables with the same metric or the effect of same variable across different sub-samples
Multiple Regression

With two independent variables, $X_1$ and $X_2$;

\[ b^*_1 \rightarrow \text{partial effect of } X_1 \text{ on } Y \text{ when controlling for } X_2 \]

\[ b^*_2 \rightarrow \text{partial effect of } X_2 \text{ on } Y \text{ when controlling for } X_1 \]

Again:

We interpret in “standard deviation units”:

- A 1 SD unit change in the independent variable corresponds to a $b^*$ SD unit change in the dependent variable, when controlling for the other variables in the model.
The regression equations:

1. \( IV \rightarrow Social \ support \)
   \[
   \hat{Y} = 10.573 - .392X_1 \quad \text{Beta } X_1 = -.205
   \]

2. \( IV \rightarrow Family \ functioning \)
   \[
   \hat{Y} = 1.951 + .358X_2 \quad \text{Beta } X_2 = .342
   \]

3. Multivariate model:
   \[
   \hat{Y} = 3.121 - .070X_1 + .338X_2 \quad \text{Beta } X_1 = -.037
   \]
   \[
   \text{Beta } X_2 = .323
   \]
Multiple Regression

• Notice that the effect of “social support” becomes much weaker when controlling for “family functioning,” but is still sig.

• This might be considered consistent with expectations of a causal chain:

  social support \[\rightarrow\] depression

  family functioning
Multiple Regression

What is the utility of this regression?

• Provides insight beyond simple bivariate relationships
• By holding constant other variables, we can more accurately assess the impact of a given IV on the DV
• Can gain a better understanding of the “causes” or “correlates” of a given phenomenon
Regression

Example

→ with 2 independent variables
→ explaining the crime rate (from page 384 of Agresti text)
Multiple Regression

Recall that in simple regression, we estimate

- Y-intercept
- Unstandardized slopes
- Standardized slope

We can also calculate the standardized slopes for all independent variables in MR

- denoted as $b_{1}^{*}$, $b_{2}^{*}$, $b_{3}^{*}$..., $b_{k}^{*}$
- In SPSS output: “beta” values

Recall:

- Standardized slopes are not influenced by the metric of the variables involved
Regression

Example:
→ with 2 independent variables
→ explaining the crime rate (from page 384 in Agresti)

Suppose we have information for all 67 counties in Florida. For each county, we document:
Regression

Example:
→ with 2 independent variables
→ explaining the crime rate (from page 384 of Agresti)

Suppose we have information for all 67 counties in Florida. For each county, we document:

→ crime rate (annual crimes per 1,000)  (DEPENDENT)
→ education (% with at least HS)  (INDEPENDENT)
→ urbanization (% in urban areas)  (INDEPENDENT)
Regression

Example:
→ with 2 independent variables
→ explaining the crime rate (from page 384 of text)

Suppose we have information for all 67 counties in Florida. For each county, we document:

→ crime rate (annual crimes per 1,000)  (DEPENDENT)
→ education (% with at least HS)      (INDEPENDENT)
→ urbanization (% in urban areas)     (INDEPENDENT)

How many cases? 67
Multiple Regression

• Begin with simple bivariate relationship
• Agresti and Finlay report the following results:

\[ E(Y) = 51.3 + 1.5X_1 \]

• Does this make sense?
• Implies that, as the % of people with HS education increases, the crime rate also increase

• Perhaps we are missing an important control. “urbanization”
\[ E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 \]

where:
- \( E(Y) \) expected value on dependent variable (crime rate)
- \( X_1 \) level of education
- \( X_2 \) level of urbanization
\[ E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 \] (our population equation)

where:

- \( E(Y) \) expected value on dependent variable (crime rate)
- \( X_1 \) level of education
- \( X_2 \) level of urbanization

\[ E(Y) = 58.9 - 0.6X_1 + 0.7X_2 \]

How do we interpret the regression coefficients??
\[ E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 \]  
(our population equation)

where:

\[ E(Y) \] expected value on dependent variable (crime rate)
\[ X_1 \] level of education
\[ X_2 \] level of urbanization

\[ E(Y) = 58.9 - .6 X_1 + .7 X_2 \]

How do we interpret the regression coefficients??

- \( \rightarrow \) 58.9 is our Y-intercept (value of Y when \( X_1 \) and \( X_2 = 0 \))
- \( \rightarrow \) - .6 is the slope of education, controlling for urbanization
- \( \rightarrow \) +.7 is the slope of urbanization, controlling for education
Multiple Regression

In multiple regression:
• the slopes are “partial regression coefficients”

Interpretation:
• “the effect of an independent variable while controlling for the other independent variable(s)”

• Meaning: the effect of an independent variable, if cases did not differ on other independent variable(s) → if we held other variables constant

• as with crosstabs, we have “statistical control groups”
Multiple Regression

From our example:

\[ E(Y) = 58.9 - 0.6X_1 + 0.7X_2 \]

With regard to \( X_1 \):

- Holding urbanization constant, education has an inverse relationship with crime

Direct interpretation:

- When controlling for “urbanization”, a 1% increase in the proportion of people with high school education results in a mean decrease of 0.6 crimes per 1,000 population
\[ E(Y) = 58.9 - 0.6X_1 + 0.7X_2 \]

With regard to \( X_2 \).
\[ E(Y) = 58.9 - 0.6X_1 + 0.7X_2 \]

With regard to \( X_2 \):
- Urbanization has a positive relationship with crime, when holding education constant
- Parts of Florida with higher levels of urbanization also have higher crime rates
\[ E(Y) = 58.9 - 0.6X_1 + 0.7X_2 \]

With regard to \(X_2\):
- Urbanization has a positive relationship with crime, when holding education constant.
- Parts of Florida with higher levels of urbanization also have higher crime rates.

Direct interpretation:
- Controlling for “education,” a 1% increase in the proportion of people living in urban areas results in an average increase of .7 crimes per 1,000 people.