Lecture (chapter 15): Partial correlation, multiple regression, and correlation

Ernesto F. L. Amaral

April 23–25, 2018
Advanced Methods of Social Research (SOCI 420)

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 15 (pp. 405–441).

TEXAS A&M UNIVERSITY

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

Family	Husband's Housework (*Y*)	Number of Children (*X*)	Husband's Years of Education (*Z*)
A	1	1	12
B	2	1	14
C	3	1	16
D	5	1	16
E	3	2	18
F	1	2	16
G	5	3	12
H	0	3	12
I	6	4	10
J	3	4	12
K	7	5	10
L	4	5	16

Source: Healey 2015, p.409.

Correlation matrix

- The bivariate (zero-order) correlation between husbands' housework and number of children is +0.50
  - This indicates a positive relationship

Zero-Order Correlations

↓	Husband's Housework (*Y*)	Number of Children (*X*)	Husband's Years of Education (*Z*)
Husband's Housework (*Y*)	1.00	0.50	−0.30
Number of Children (*X*)		1.00	−0.47
Husband's Years of Education (*Z*)			1.00

Source: Healey 2015, p.410.

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*)

$$r_{yx.z} = \frac{r_{yx} - \left(r_{yz}\right)\left(r_{xz}\right)}{\sqrt{1 - r_{yz}^2}\sqrt{1 - r_{xz}^2}}$$

$$r_{yx.z} = \frac{(0.50) - (-0.30)(-0.47)}{\sqrt{1 - (-0.30)^2}\sqrt{1 - (-0.47)^2}}$$

$$r_{yx.z} = 0.43$$

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

$$Y = a + bX = \beta_0 + \beta_1 X$$

  - $a = \beta_0 =$ *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

$$Y = a + b_1X_1 + b_2X_2 = \beta_0 + \beta_1X_1 + \beta_2X_2$$

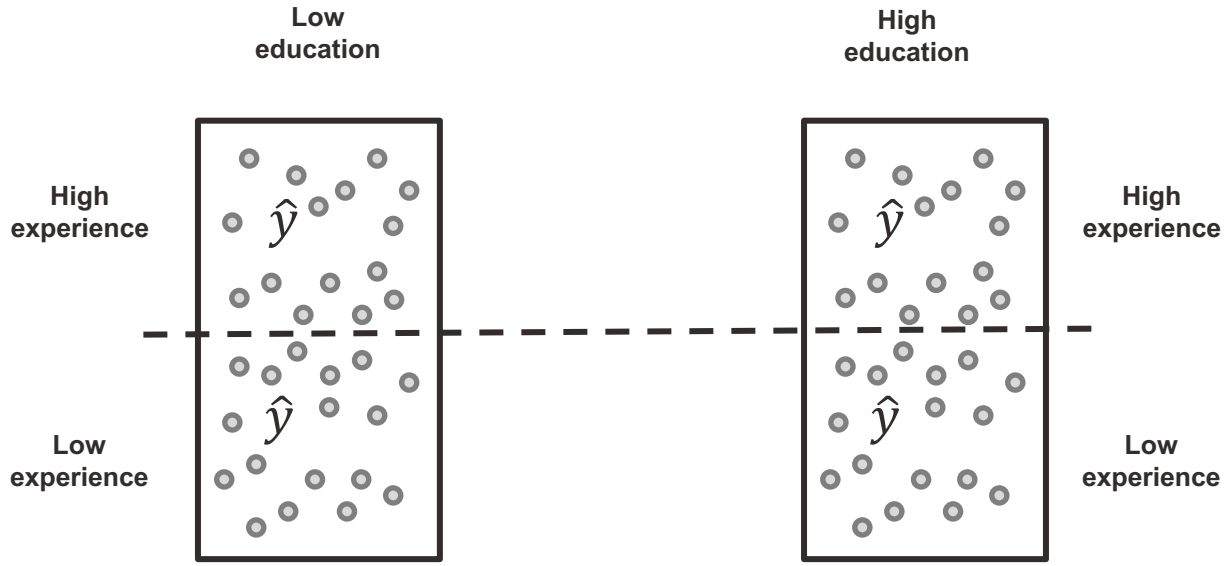
- $a = \beta_0$ = the Y intercept, where the regression line crosses the Y axis
- $b_1 = \beta_1$ = partial slope for X_1 on Y
  - β_1 indicates the change in Y for one unit change in X_1, controlling for X_2
- $b_2 = \beta_2$ = partial slope for X_2 on Y
  - β_2 indicates the change in Y for one unit 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

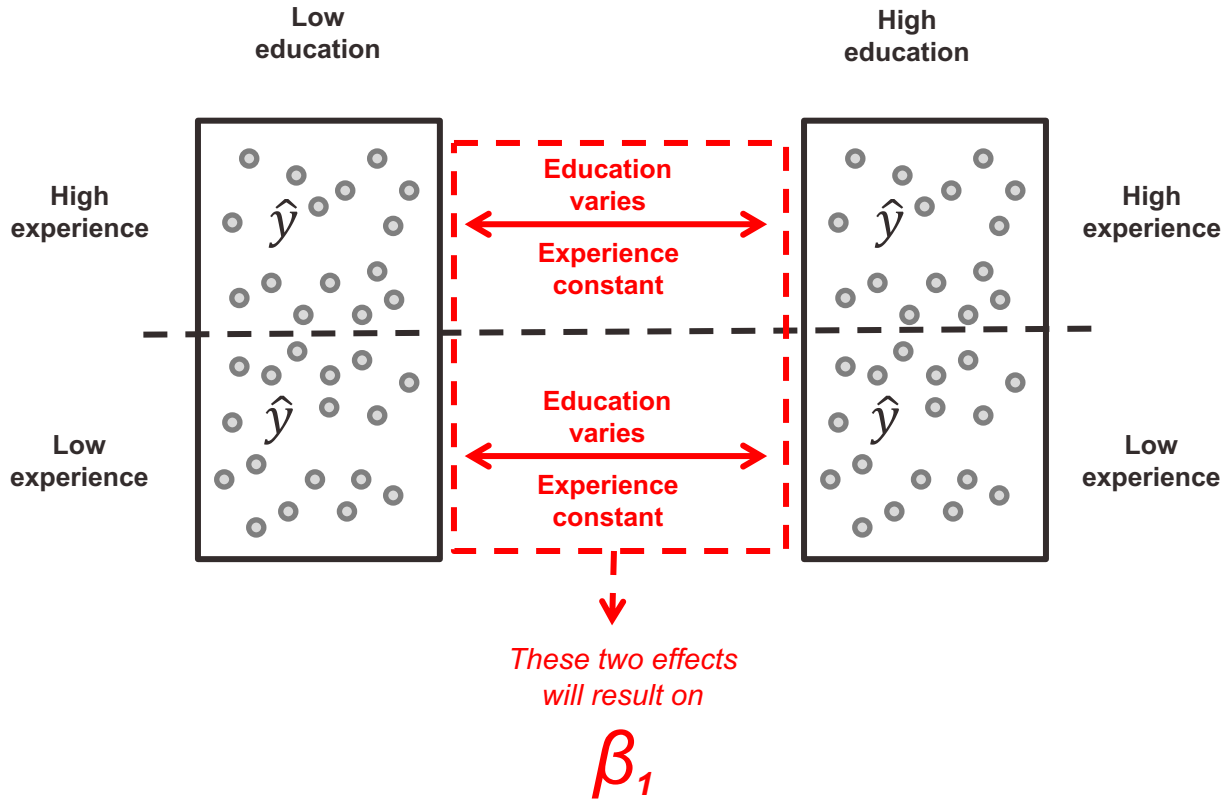
Ceteris paribus

$$Income = \beta_0 + \beta_1 education + \beta_2 experience + u$$



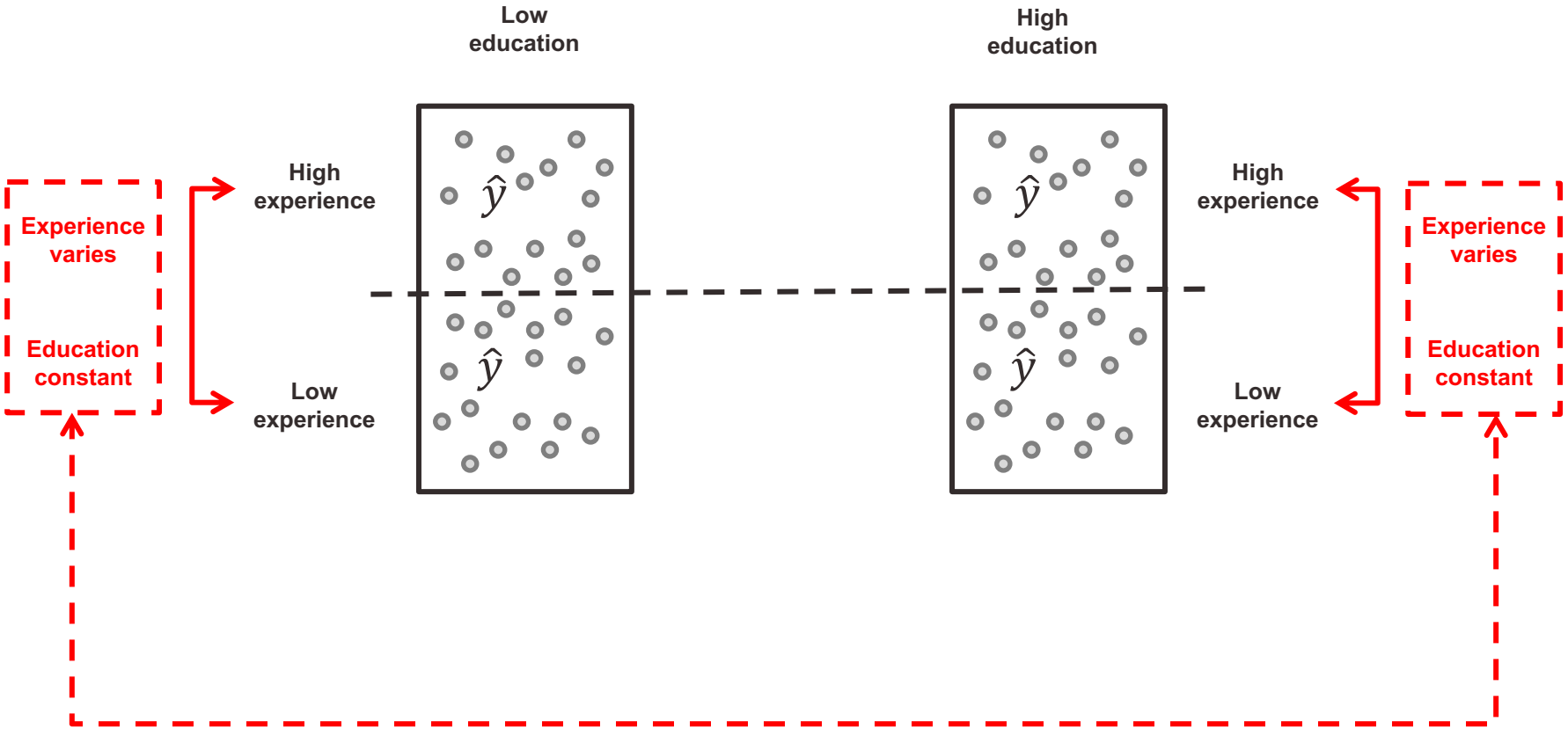
Ceteris paribus

$$Income = \beta_0 + \beta_1 education + \beta_2 experience + u$$



Ceteris paribus

$$Income = \beta_0 + \beta_1 education + \beta_2 experience + u$$

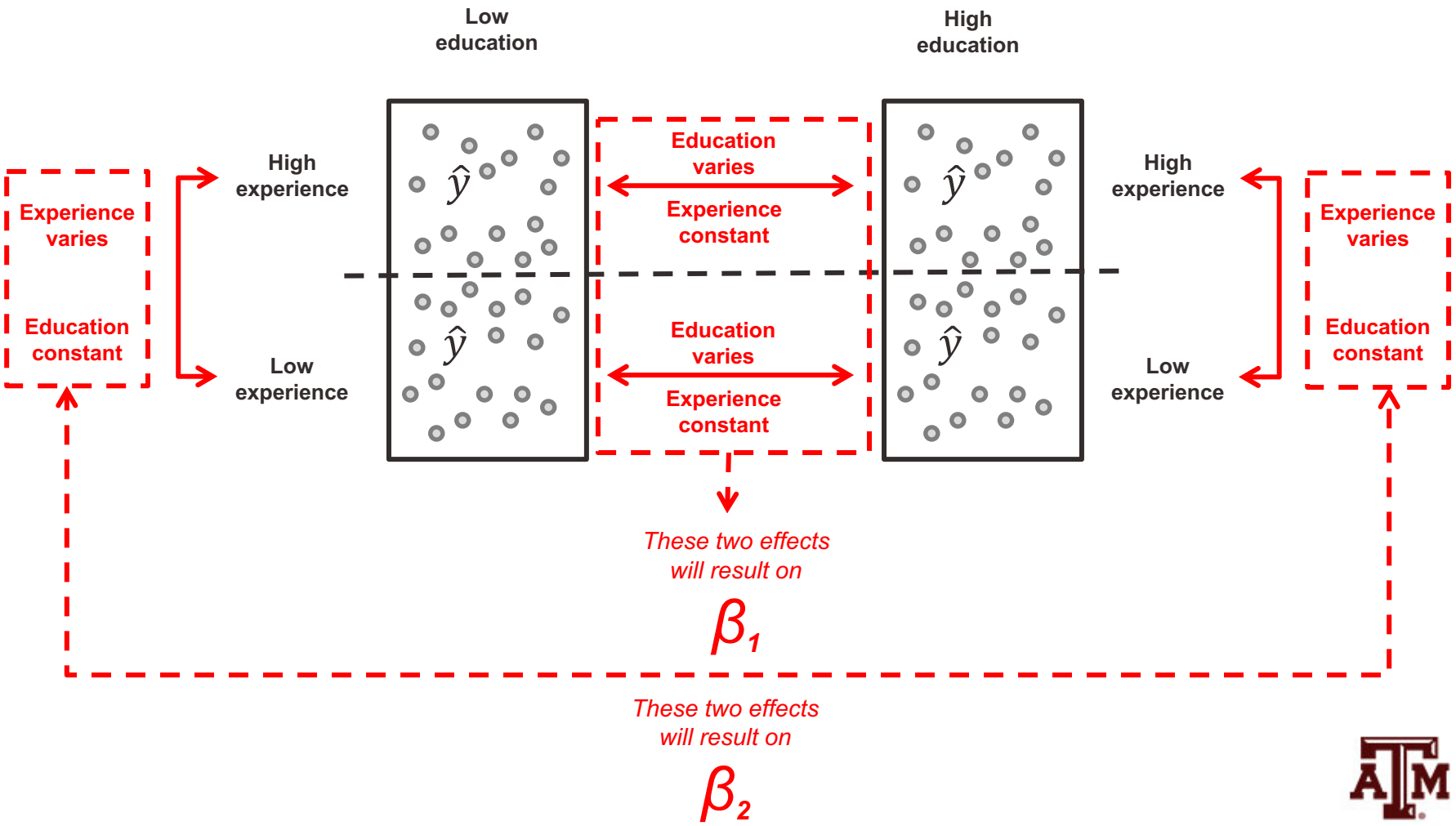


These two effects will result on

β_2

Ceteris paribus

$$Income = \beta_0 + \beta_1 education + \beta_2 experience + u$$



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 (β_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 = β_1 = partial slope of X_1 on Y

b_2 = β_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 (β_1) and b_2 (β_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

Husband's Housework	Number of Children	Husband's Education
$\overline{Y} = 3.3$	$\overline{X}_1 = 2.7$	$\overline{X}_2 = 13.7$
$s_y = 2.1$	$s_1 = 1.5$	$s_2 = 2.6$
Zero-Order Correlations		
$r_{y1} = 0.50$		
$r_{y2} = -0.30$		
$r_{12} = -0.47$		

Source: Healey 2015, p.414.

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

$$Y = a + b_1X_1 + b_2X_2$$

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2$$

$$Y = 2.5 + (0.65)X_1 + (-0.07)X_2$$

$$Y = 2.5 + 0.65X_1 - 0.07X_2$$

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

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

  - 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)

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

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

$$R^2 = 0.255$$

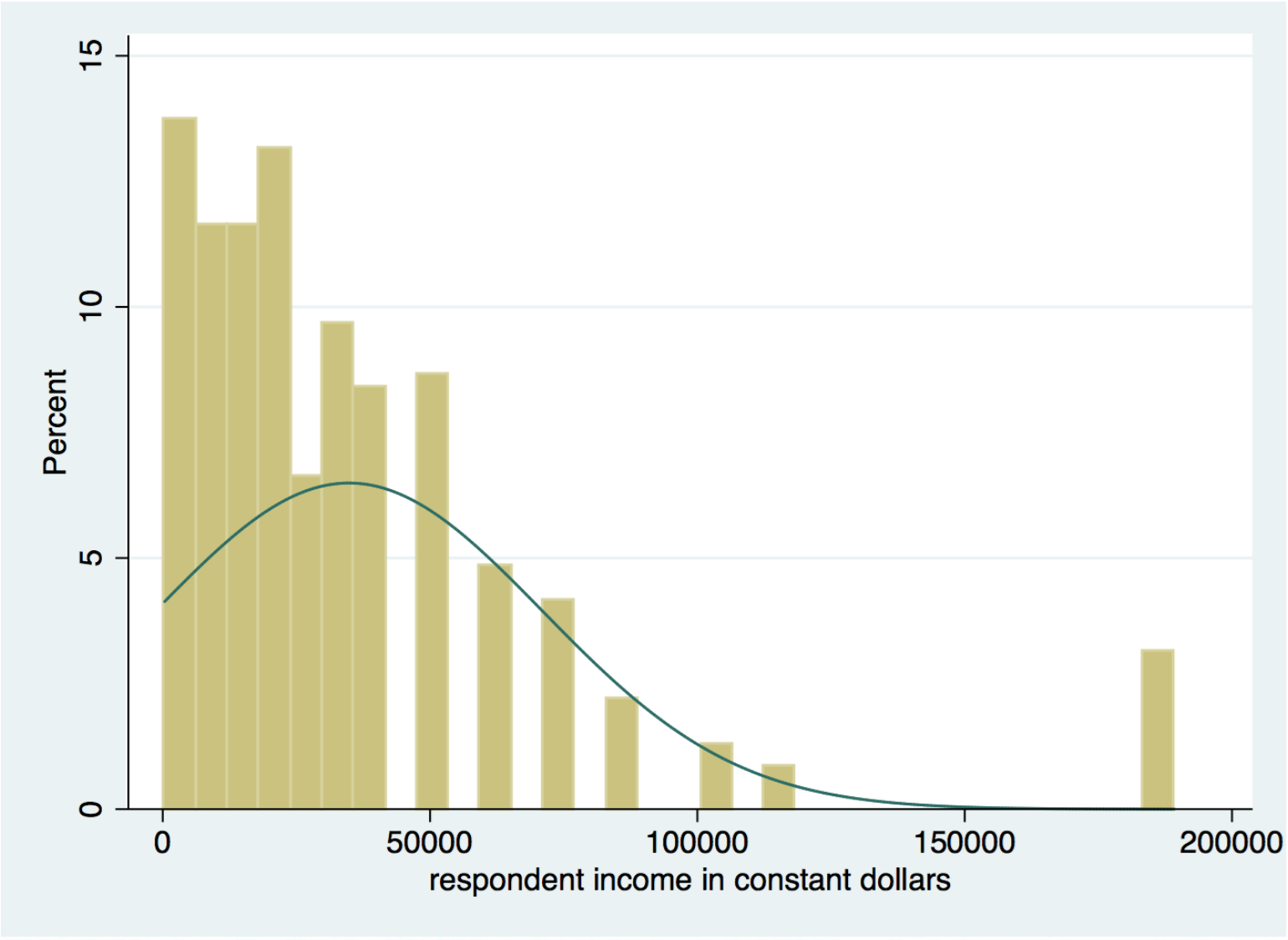
- Number of children and husband's education explain 25.5% of the variation in husband's housework

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

Number of strata   =        65                  Number of obs     =      1,626
Number of PSUs     =       130                  Population size   = 1,688.1407
                                                Design df         =         65
                                                F(   2,     64)   =      74.94
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1434
```

conrinc	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
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
```

Source	SS	df	MS		
Model	2.5113e+11	2	1.2557e+11	Prob > F	= 0.0000
Residual	1.8608e+12	1,623	1.1465e+09	R-squared	= 0.1189
Total	2.1120e+12	1,625	1.2997e+09	Root MSE	= 33861

Number of obs	=	1,626			
F(2, 1623)	=	109.52			
Adj R-squared	=	0.1178			

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)

$$Y^3 \longrightarrow q = 3$$

$$Y^2 \longrightarrow q = 2$$

$$Y^1 \longrightarrow q = 1$$

$$Y^{0.5} \longrightarrow q = 0.5$$

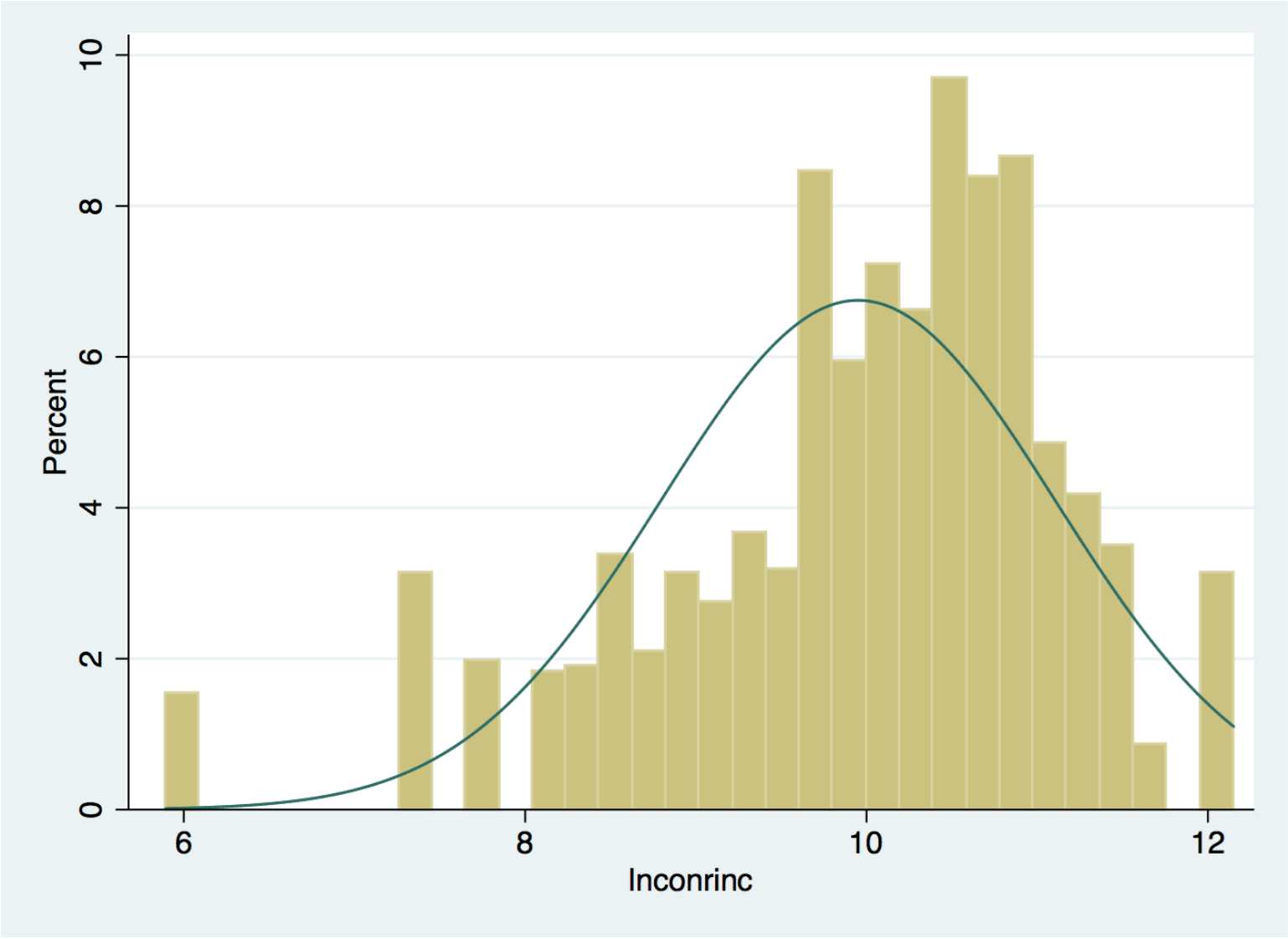
$$\log(Y) \longrightarrow q = 0$$

$$-(Y^{-0.5}) \longrightarrow q = -0.5$$

$$-(Y^{-1}) \longrightarrow q = -1$$

- 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 lnconrinc age educ if year==2016
(running regress on estimation sample)

Survey: Linear regression

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

lnconrinc	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0157926	.0021115	7.48	0.000	.0115756	.0200096
educ	.1229175	.0109476	11.23	0.000	.1010536	.1447814
_cons	7.506136	.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(β_1) times
    - When age increases by one unit, income increases on average by **1.0161 times**, controlling for education
  - 100*[exp(β_1)–1] percent
    - When age increases by one unit, income increases on average by **1.61%**, controlling for education
  - 100*β_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
```

Source	SS	df	MS		
Model	240.448673	2	120.224336	Prob > F	0.0000
Residual	1926.88361	1,623	1.18723574	R-squared	0.1109
Total	2167.33229	1,625	1.33374294	Root MSE	1.0896

Number of obs	=	1,626			
F(2, 1623)	=	101.26			
Prob > F	=	0.0000			
R-squared	=	0.1109			
Adj R-squared	=	0.1098			
Root MSE	=	1.0896			

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	.

Source: 2016 General Social Survey.

Dummy variables

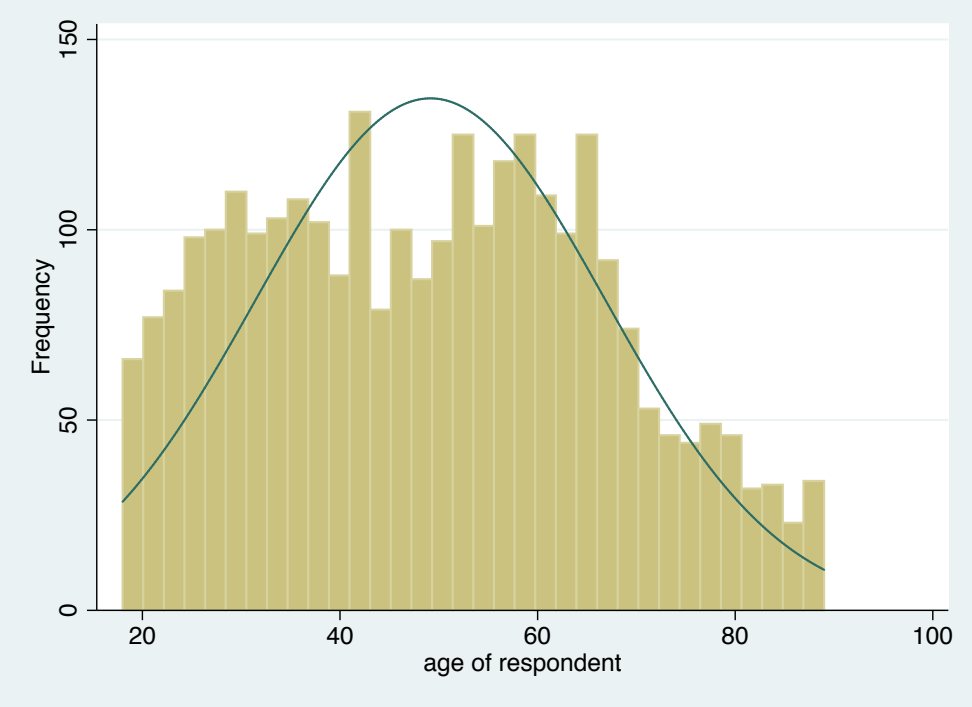
- Many variables that are important in social life are nominal-level variables
  - They cannot be included in a regression equation or correlational analysis (e.g., sex, race/ethnicity)
- We can create dummy variables
  - Two categories, one coded as 0 and the other as 1

Sex	Male	Female
1	1	0
2	0	1

Race/ ethnicity	White	Black	Hispanic	Other
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	0	0	0	1

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

```
. egen agegr = cut(age), at(18,25,35,50,65,90)
(22 missing values generated)

. table agegr, contents(min age max age count age)
```

agegr	min(age)	max(age)	N(age)
18	**18**	**24**	**671**
25	**25**	**34**	**1,452**
35	**35**	**49**	**2,116**
50	**50**	**64**	**2,022**
65	**65**	**89**	**1,440**

- Generate dummy variables for age...

Source: 2016 General Social Survey.

Dummies for age

- Generate dummy variables for age group

Age group	Age 18–24	Age 25–35	Age 35–49	Age 50–64	Age 65–89
18–24	1	0	0	0	0
25–34	0	1	0	0	0
35–49	0	0	1	0	0
50–64	0	0	0	1	0
65–89	0	0	0	0	1

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

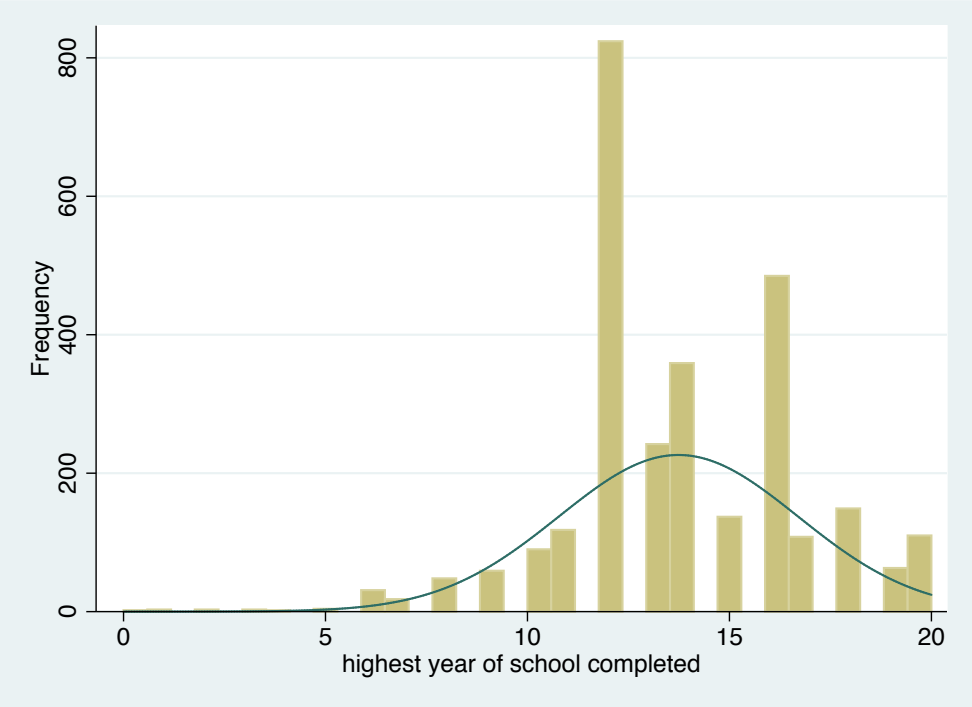
```
. tab agegr
```

agegr	Freq.	Percent	Cum.
18	671	8.71	8.71
25	1,452	18.85	27.57
35	2,116	27.48	55.04
50	2,022	26.26	81.30
65	1,440	18.70	100.00
Total	7,701	100.00	

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
```

rs highest degree	Freq.	Percent	Cum.
lt high school	997	12.92	12.92
high school	3,897	50.52	63.44
junior college	585	7.58	71.03
bachelor	1,418	18.38	89.41
graduate	817	10.59	100.00
Total	7,714	100.00	

- Generate dummy variables for education...

Source: 2016 General Social Survey.

Dummies for education

- Generate dummy variables for education group

Education group	<HS	HS	JC	BA	GR
Less than high school	1	0	0	0	0
High school	0	1	0	0	0
Junior college	0	0	1	0	0
Bachelor	0	0	0	1	0
Graduate	0	0	0	0	1

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

```
. tab degree
```

rs highest degree	Freq.	Percent	Cum.
lt high school	997	12.92	12.92
high school	3,897	50.52	63.44
junior college	585	7.58	71.03
bachelor	1,418	18.38	89.41
graduate	817	10.59	100.00
Total	7,714	100.00	

Source: 2016 General Social Survey.

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

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

lnconrinc	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
agegr1	-1.166963	.1220959	-9.56	0.000	-1.410805	-.9231207
agegr2	-.3345438	.0736023	-4.55	0.000	-.4815379	-.1875498
agegr4	-.0050007	.0638917	-0.08	0.938	-.1326013	.1225999
agegr5	-.4155278	.096474	-4.31	0.000	-.6081997	-.2228559
educgr1	-.4276264	.1163403	-3.68	0.000	-.6599739	-.1952789
educgr3	.2367316	.0940649	2.52	0.014	.0488711	.4245921
educgr4	.4559903	.0843136	5.41	0.000	.2876045	.6243761
educgr5	.8516728	.0920326	9.25	0.000	.667871	1.035475
_cons	9.949482	.0471336	211.09	0.000	9.855349	10.04361

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(β_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(β_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*β_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		
Model	395.582999	8	49.4478749	Number of obs	= 1,626
Residual	1771.74929	1,617	1.09570148	F(8, 1617)	= 45.13
				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	.0810907	9.86	0.000	.2359353
_cons	9.927145	.0541188	183.43	0.000	.

Source: 2016 General Social Survey.

Edited table

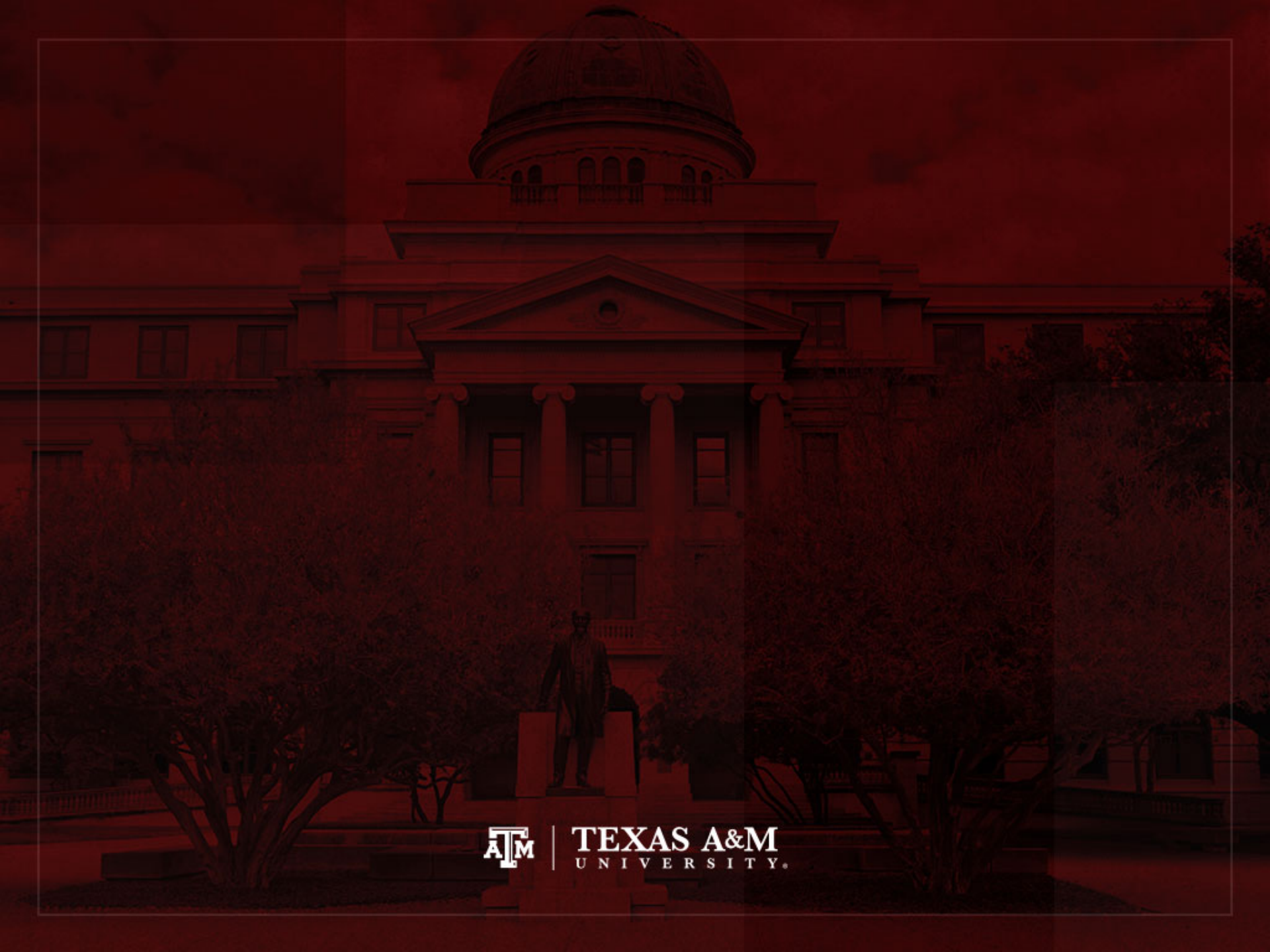
Table 1. Coefficients and standard errors estimated with ordinary least squares models for the logarithm of respondent's income as the dependent variable, U.S. adult population, 2004, 2010, and 2016

Independent variables	2004		2010		2016	
	Coefficients	Standardized coefficients	Coefficients	Standardized coefficients	Coefficients	Standardized coefficients
Constant	10.030***		9.919***		9.949***	
	(0.063)		(0.090)		(0.047)	
Age groups						
18–24	-1.114***	-0.269	-1.438***	-0.327	-1.167***	-0.257
	(0.104)		(0.188)		(0.122)	
25–34	-0.306***	-0.118	-0.406***	-0.140	-0.335***	-0.109
	(0.074)		(0.102)		(0.074)	
35–49	ref.	ref.	ref.	ref.	ref.	ref.
50–64	0.132*	0.041	0.043	0.015	-0.005	0.006
	(0.068)		(0.092)		(0.064)	
65+	-0.596***	-0.120	-0.720***	-0.168	-0.416***	-0.111
	(0.165)		(0.175)		(0.097)	
Education groups						
Less than high school	-0.410***	-0.101	-0.477***	-0.139	-0.428***	-0.097
	(0.117)		(0.125)		(0.116)	
High school	ref.	ref.	ref.	ref.	ref.	ref.
Junior college	0.276***	0.071	0.142	0.018	0.237**	0.052
	(0.097)		(0.122)		(0.094)	
Bachelor	0.620***	0.219	0.579***	0.197	0.456***	0.183
	(0.062)		(0.099)		(0.084)	
Graduate	0.785***	0.233	0.983***	0.251	0.852***	0.236
	(0.097)		(0.088)		(0.092)	
R^2	0.242	0.222	0.288	0.272	0.198	0.183
Number of observations	1,685	1,685	1,201	1,201	1,626	1,626

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.
Source: 2004, 2010, 2016 General Social Surveys.

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


TEXAS A&M
UNIVERSITY