Lecture 13: Bivariate associations for interval-ratio-level variables

Ernesto F. L. Amaral

October 19–24, 2023
Advanced Methods of Social Research (SOCI 420)

www.ernestoamaral.com

Source: Healey, Joseph F. 2015. "Statistics: A Tool for Social Research." Stamford: Cengage Learning. 10th edition. Chapter 13 (pp. 342–378).

TEXAS A&M UNIVERSITY

Outline

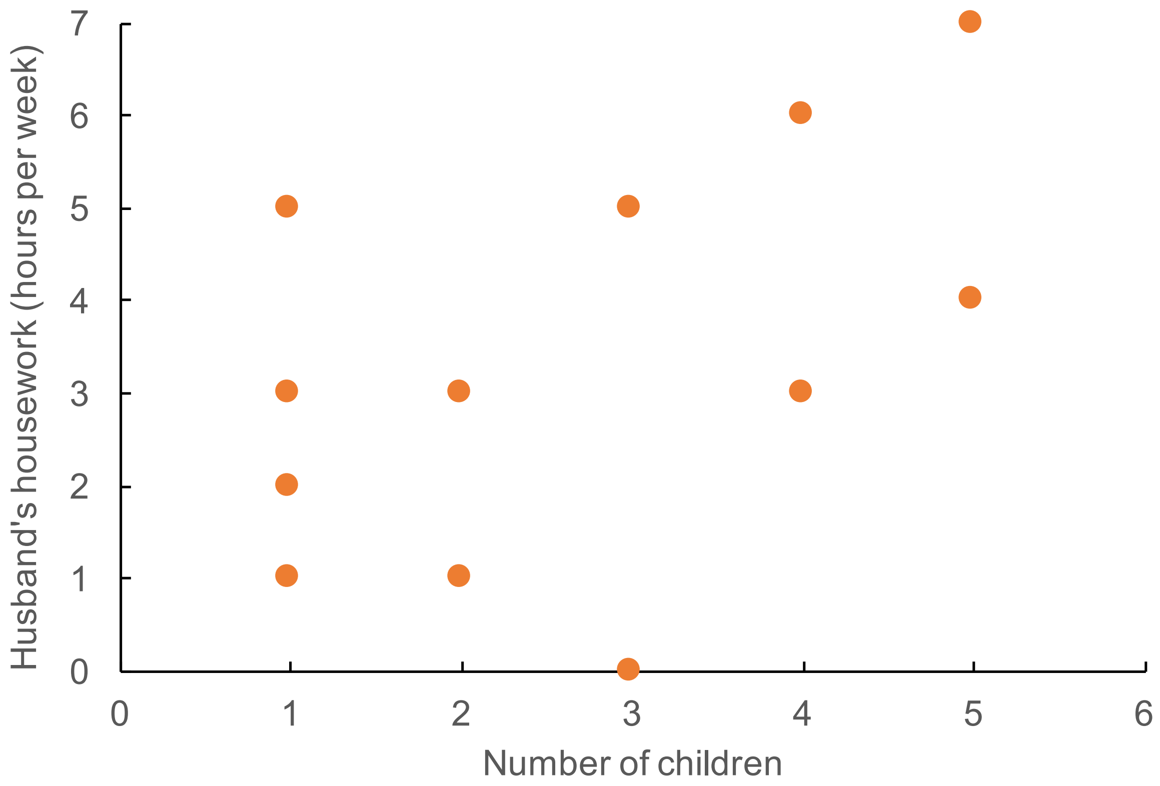
- Scatterplots
- Pearson's *r* and r^2
  – Explain the concepts of total, explained, and unexplained variance
  – Test Pearson's *r* for significance: five-step model

Scatterplots

- Scatterplots have two dimensions
  - The independent variable (X) is displayed along the horizontal axis
  - The dependent variable (Y) is displayed along the vertical axis
- Each dot on a scatterplot is a case
  - The dot is placed at the intersection of the case's scores on X and Y
- Inspection of a scatterplot should always be the first step in assessing the association between two interval-ratio level variables

Example of a scatterplot

- Number of children (X) and hours per week husband spends on housework (Y) at dual-career households

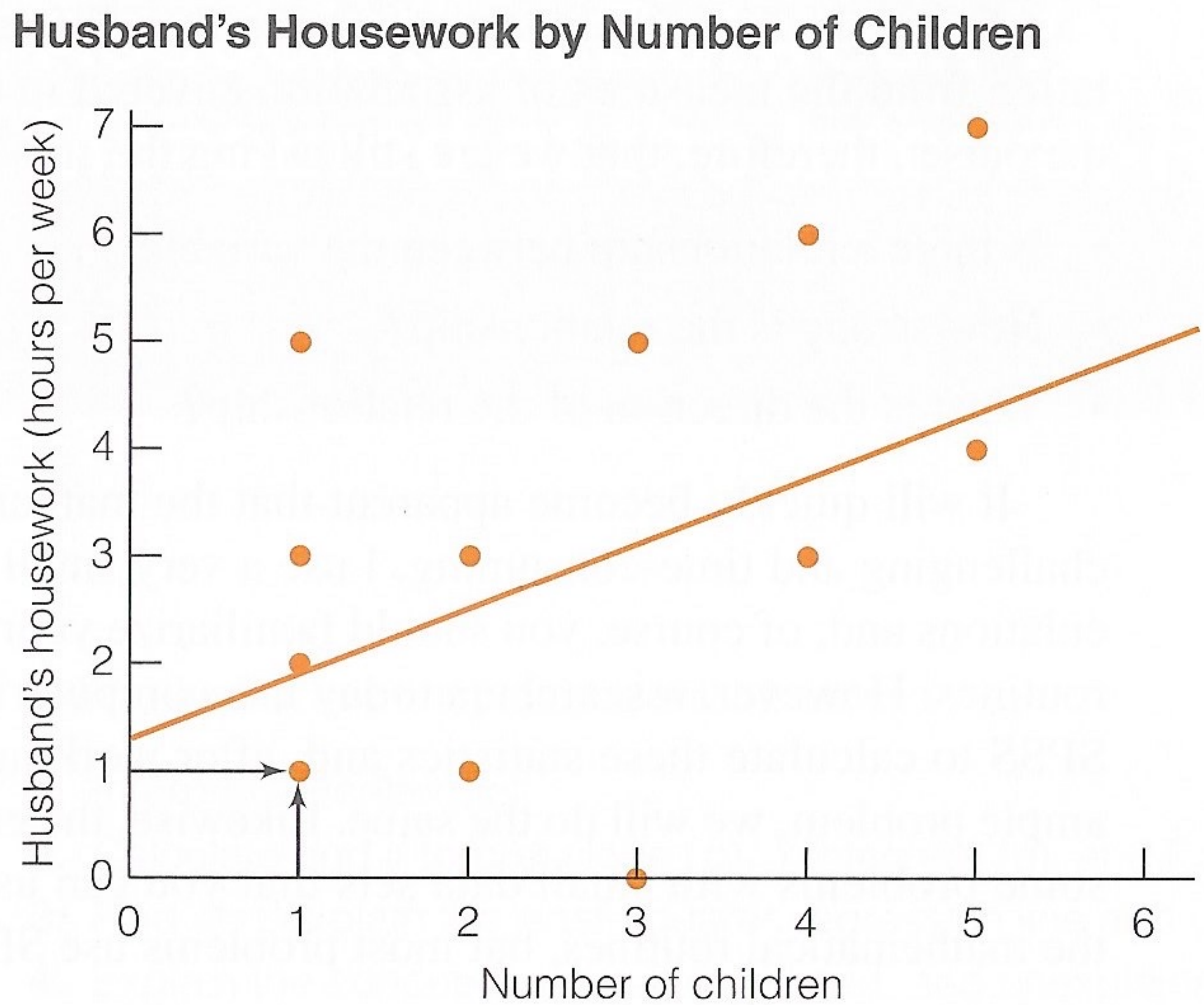


Source: Healey 2015, p.343.

Regression line

- A regression line is added to the graph
- It summarizes the linear correlation between X and Y
  - This straight line connects all of the dots
  - Or this line comes as close as possible to connecting all of the dots

Scatterplot with regression line



Source: Healey 2015, p.344.

Use of scatterplots

- Scatterplots can be used to answer these questions

1. Is there an association?

2. How strong is the association?

3. What is the pattern of the association?

1. Is there an association?

- An association exists if the conditional means of Y change across values of X

- If the regression line has an angle to the X axis
  - We can conclude that an association exists between the two variables
  - The line is not parallel to the X axis

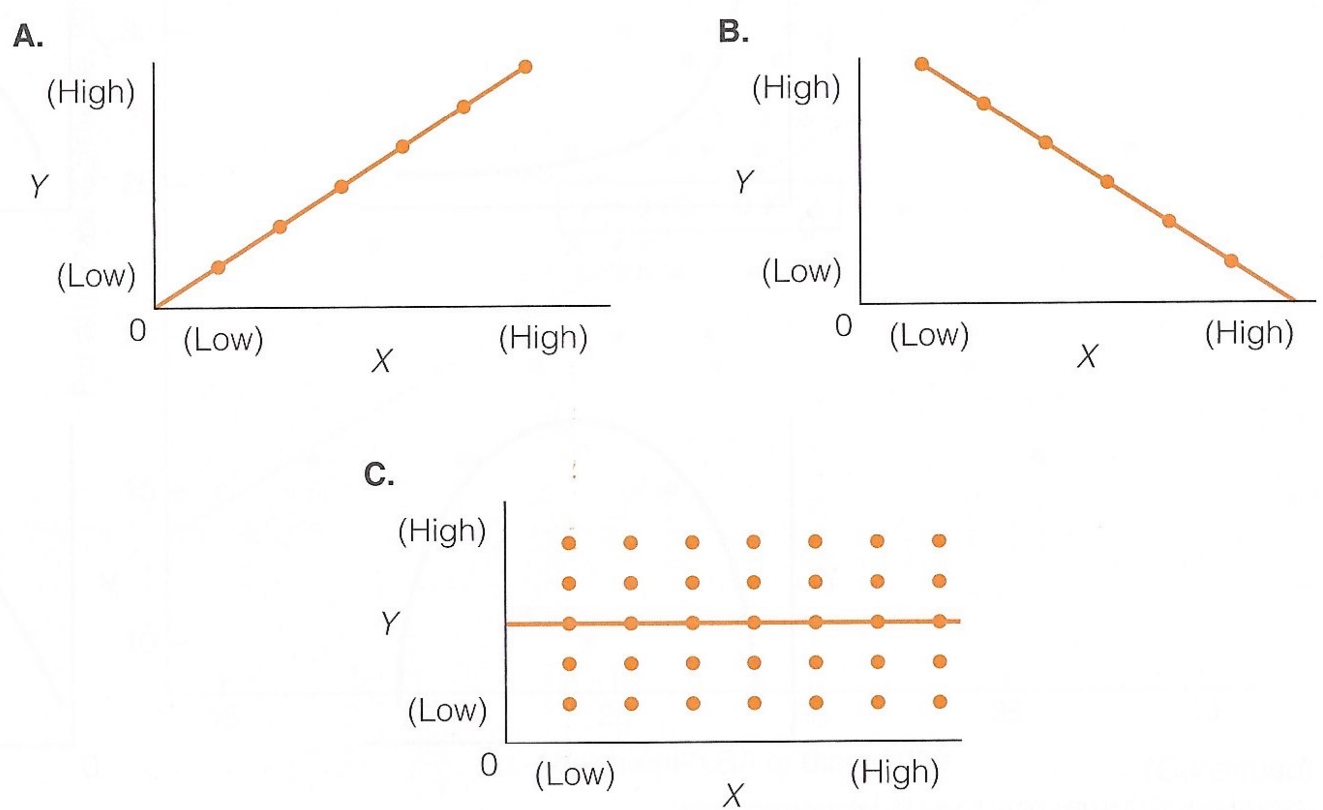
2. How strong is the association?

- Strength of the correlation is determined by the spread of the dots around the regression line
- In a perfect association
  - All dots fall on the regression line
- In a stronger association
  - The dots fall close to the regression line
- In a weaker association
  - The dots are spread out relatively far from the regression line

3. Pattern of the association

- The pattern or direction of association is determined by the angle of the regression line

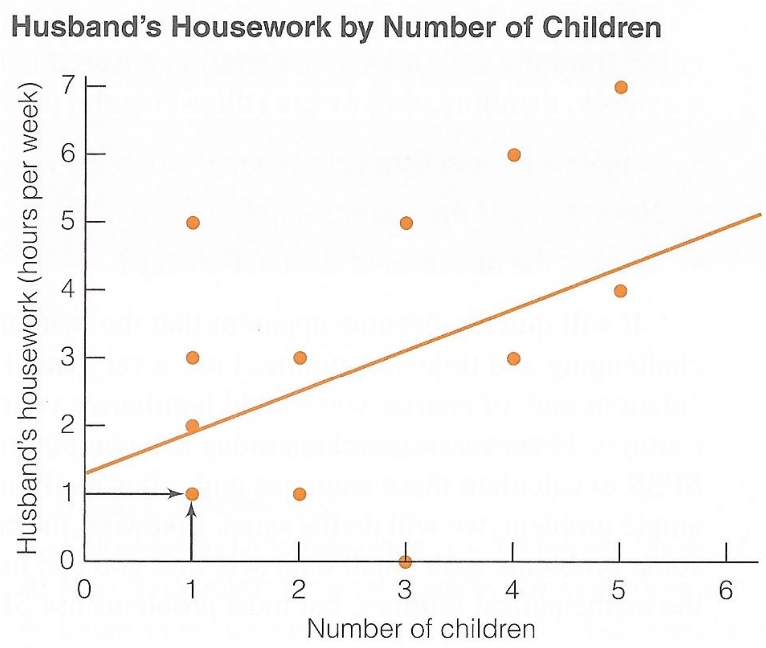
Positive (a), Negative (b), and Zero (c) Relationships



Source: Healey 2015, p.345.

Check for linearity

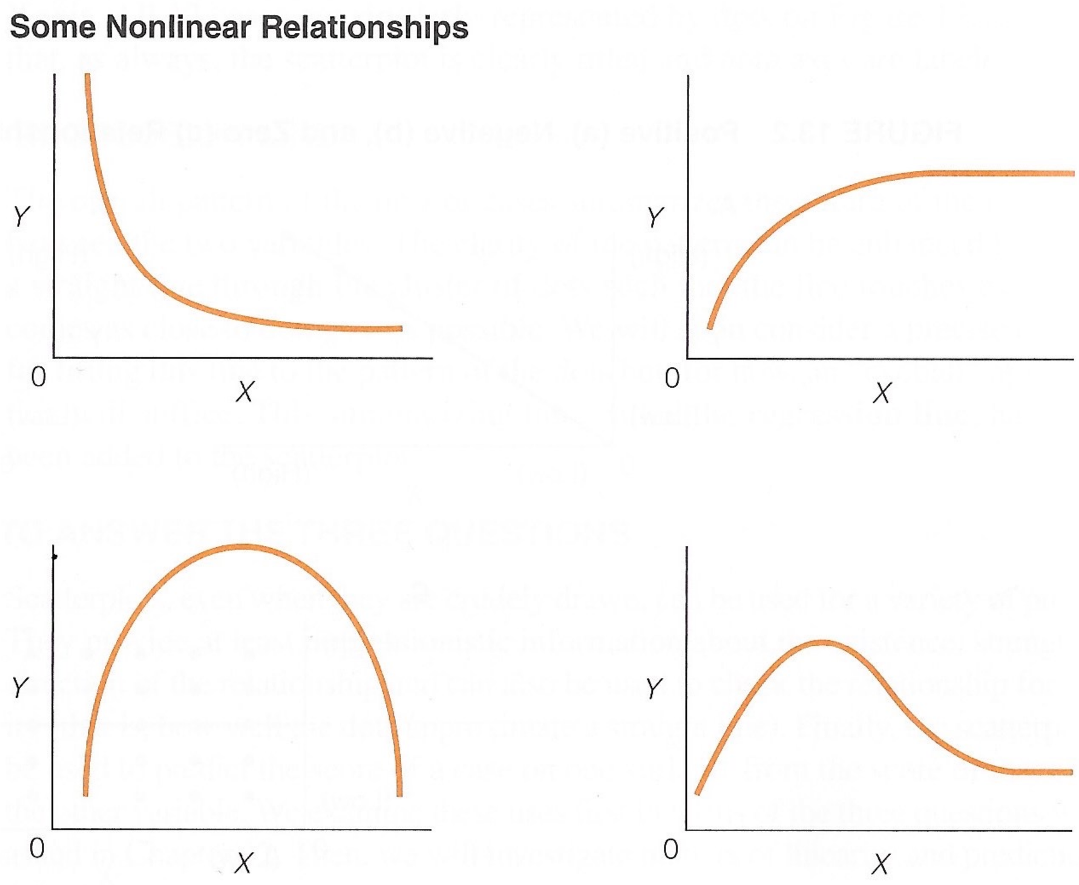
- Scatterplots can be used to check for linearity
  - An assumption of scatterplots and linear regression analysis is that X and Y have a linear correlation
  - In a linear association, the dots of a scatterplot form a straight line pattern



Source: Healey 2015, p.344.

Nonlinear associations

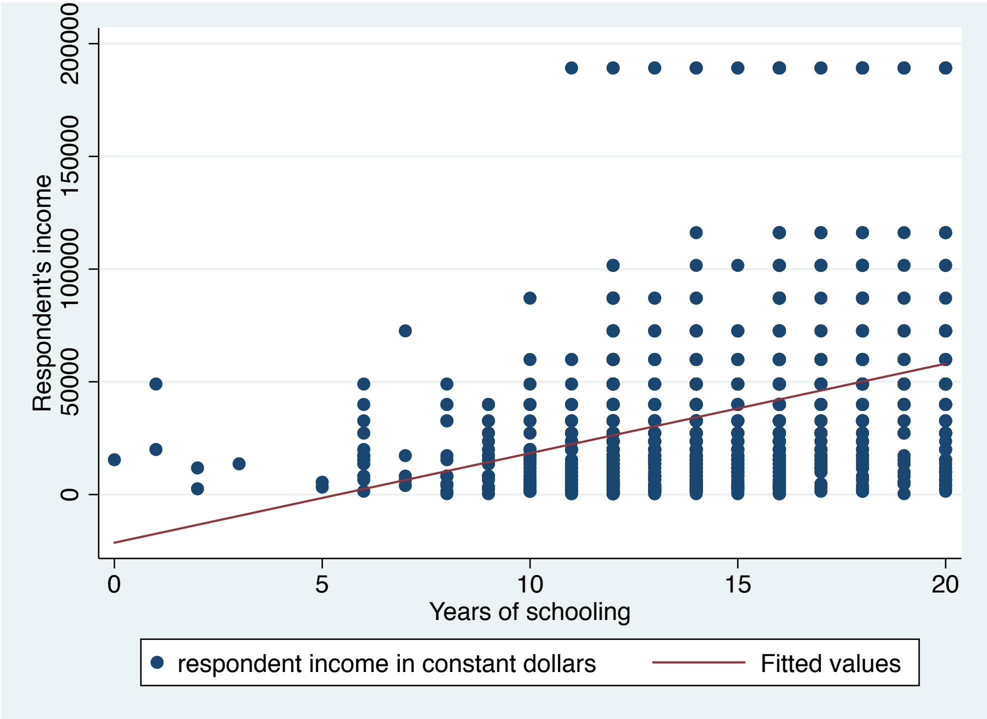
- In a nonlinear association, the dots do not form a straight line pattern



Source: Healey 2015, p.346.

GSS: Income by education

Figure 1. Respondent's income by years of schooling, U.S. adult population, 2016



Income = –26,219.18 + 4,326.10(Years of schooling)

Note: The scatterplot was generated without the complex survey design of the General Social Survey. The regression was generated taking into account the complex survey design of the General Social Survey.
Source: 2016 General Social Survey.

GSS: Income = F(Education)

```
***Dependent variable: Respondent's income (conrinc)
***Independent variable: Years of schooling (educ)

***Scatterplot with regression line
twoway scatter conrinc educ || lfit conrinc educ, ytitle(Respondent's income) xtitle(Years of schooling)

***Regression coefficients
***Least-squares regression model
***They can be reported in the footnote of the scatterplot
svy: reg conrinc educ
```

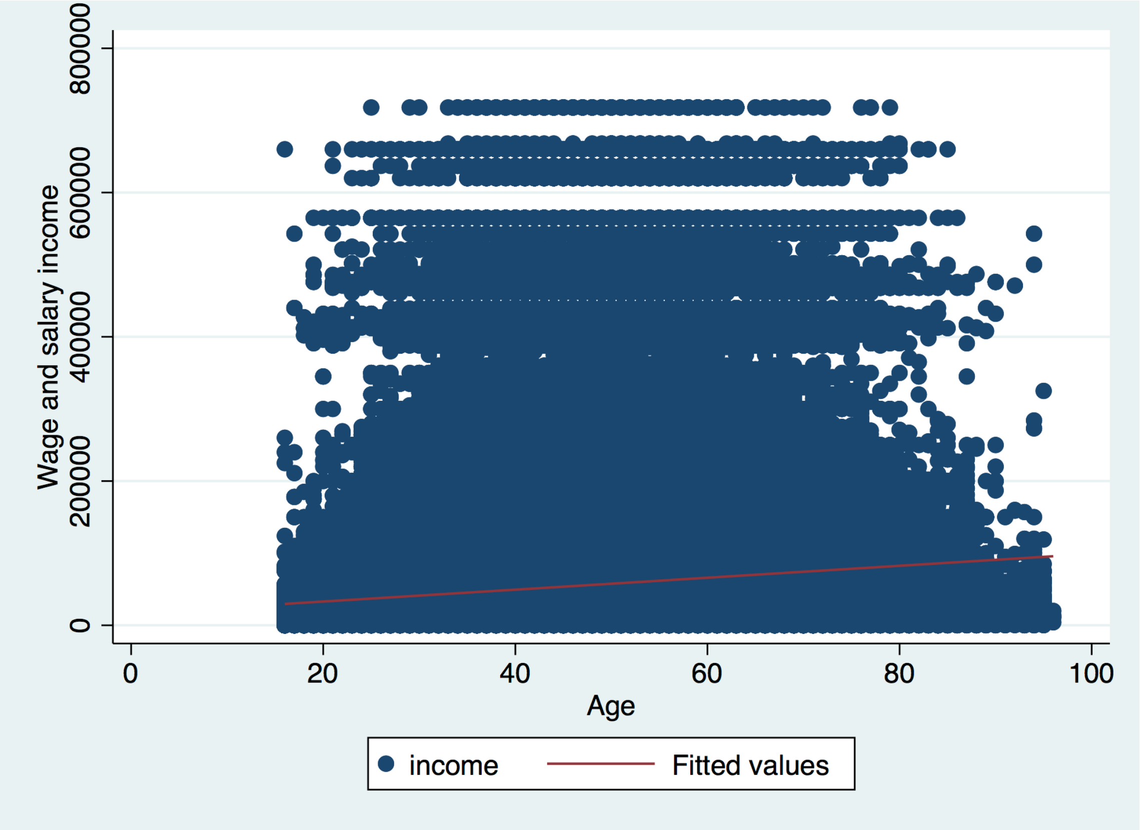
```
. svy: reg conrinc educ
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        65          Number of obs     =      1,631
Number of PSUs     =       130          Population size   = 1,694.7478
                                        Design df         =         65
                                        F(   1,      65)  =      88.15
                                        Prob > F          =     0.0000
                                        R-squared         =     0.1147
```

conrinc	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	4326.103	460.7631	9.39	0.000	3405.896	5246.311
_cons	-26219.18	5819.513	-4.51	0.000	-37841.55	-14596.81

Source: 2016 General Social Survey.

ACS: Income by age

Figure 1. Wage and salary income by age, U.S. 2018



Income = 13,447.38 + 888.23(Age)

Note: The scatterplot was generated without the ACS complex survey design. The regression was generated taking into account the ACS complex survey design. Only people with some wage and salary income are included.
Source: 2018 American Community Survey (ACS).

ACS: Income = F(Age)

```
***Dependent variable: Wage and salary income (income)
***Independent variable: Age (age)

***Scatterplot with regression line
twoway (scatter income age) (lfit income age) if income!=0, ytitle(Wage and salary income) xtitle(Age)
```

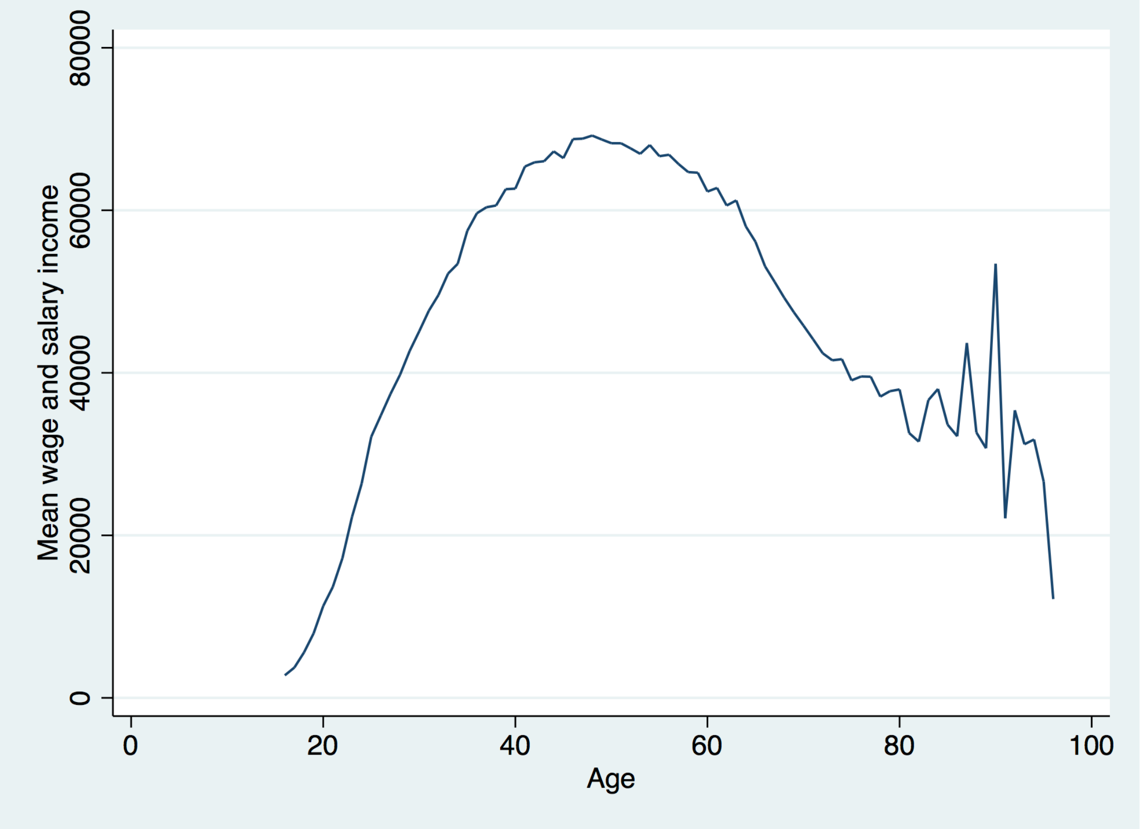
```
. svy, subpop(if income!=. & income!=0): reg income age
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =     2,351              Number of obs     =    3,214,539
Number of PSUs     = 1,410,976              Population size   =  327,167,439
                                            Subpop. no. obs   =    1,574,313
                                            Subpop. size      =  163,349,075
                                            Design df         =    1,408,625
                                            F(   1,1408625)   =     57648.04
                                            Prob > F          =       0.0000
                                            R-squared         =       0.0449
```

income	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	888.2282	3.699409	240.10	0.000	880.9775	895.479
_cons	13447.38	138.3572	97.19	0.000	13176.21	13718.56

Source: 2018 American Community Survey.

ACS: Mean income by age

Figure 1. Mean wage and salary income by age, U.S. 2018



Income = –73,956.52 + 5,492.81(Age) – 53.36(Age squared)

Note: The line graph was generated taking into account the ACS sample weight. The regression was generated taking into account the ACS complex survey design. Only people with some wage and salary income are included.
Source: 2018 American Community Survey (ACS).

ACS: Income = F(Age, Age^2)

```
***Dependent variable: Wage and salary income (income)
***Independent variables: Age (age), age squared (agesq)

***Generate variable with mean income by age
bysort age: egen mincage=mean(income) if income!=0

***Line graph of income by age
twoway line mincage age [fweight=perwt], ytitle("Mean wage and salary income") ylabel(0(20000)80000)

***Generate age squared
gen agesq=age * age
```

```
. svy, subpop(if income!=. & income!=0): reg income age agesq
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =     2,351          Number of obs     =   3,214,539
Number of PSUs     = 1,410,976          Population size   = 327,167,439
                                        Subpop. no. obs   =   1,574,313
                                        Subpop. size      = 163,349,075
                                        Design df         =   1,408,625
                                        F(   2,1408624)   =    85652.78
                                        Prob > F          =      0.0000
                                        R-squared         =      0.0839
```

income	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	5492.806	20.13499	272.80	0.000	5453.342	5532.27
agesq	-53.36376	.2435244	-219.13	0.000	-53.84106	-52.88646
_cons	-73956.52	352.3116	-209.92	0.000	-74647.03	-73266

Source: 2018 American Community Survey.

ACS: Income by age group

```
. ***Use aweight to get sample size by age group
. table agegr [aweight=perwt] if income!=0, c(mean income sd income n income)
```

agegr	mean(income)	sd(income)	N(income)
0			0
16	6255.097	10792.61	82,884
20	18744.6	19610.05	146,813
25	42093.8	39527.84	315,787
35	60282.16	65996.67	296,932
45	66337.25	74647.34	315,072
55	63089.86	73052.64	296,653
65	47947.36	72828.89	120,172

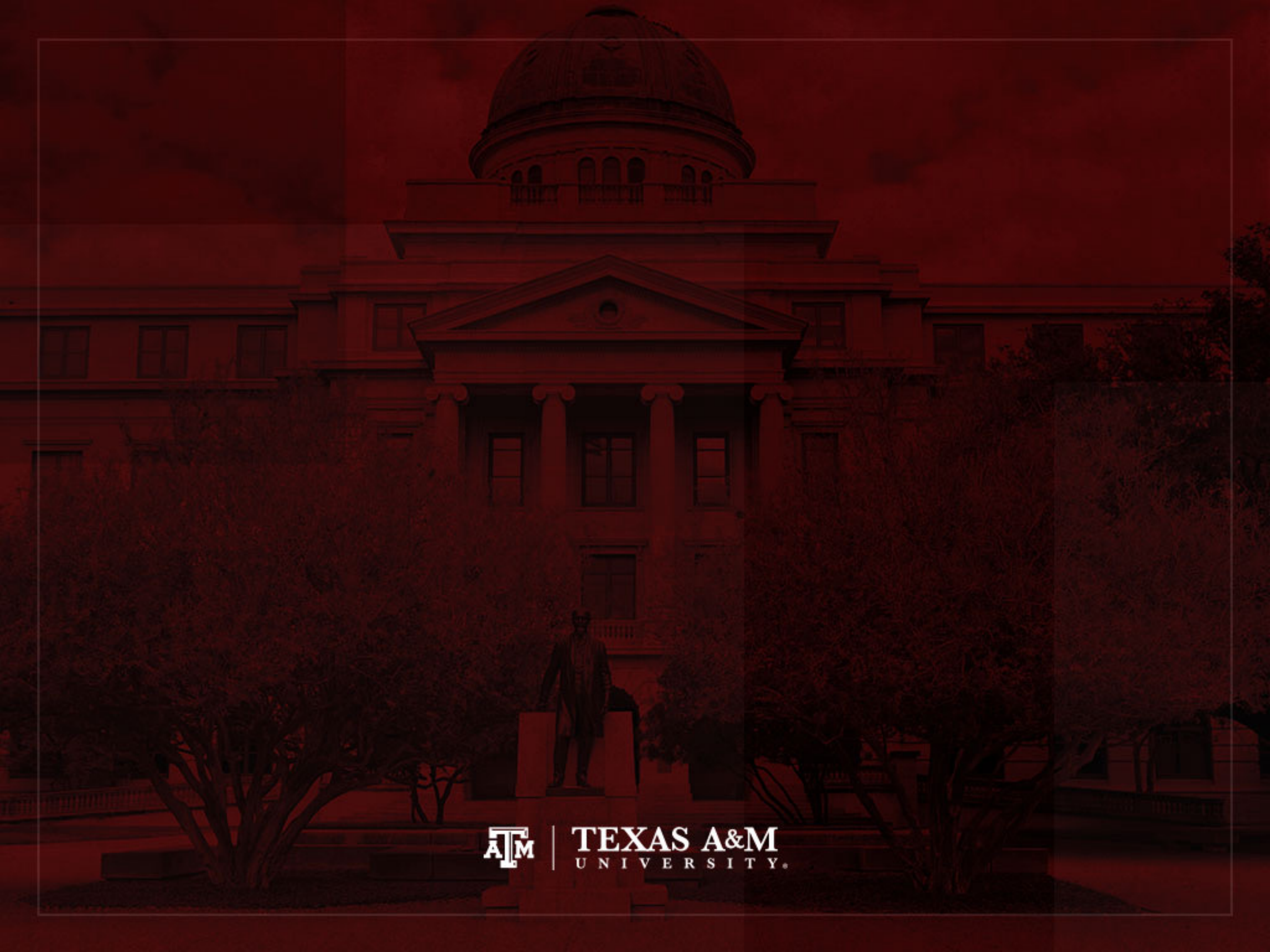
Source: 2018 American Community Survey.

ACS: Income = F(Age groups)

```
. ***Reference category: 45-54
. char agegr[omit] 45

.
. ***Income <- Age groups
. xi: svy, subpop(if income!=. & income!=0): reg income i.agegr
i.agegr           _Iagegr_0-65        (naturally coded; _Iagegr_45 omitted)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =     2,351              Number of obs      =   3,214,539
Number of PSUs     = 1,410,976              Population size    = 327,167,439
                                            Subpop. no. obs    =   1,574,313
                                            Subpop. size       = 163,349,075
                                            Design df          =   1,408,625
                                            F(   6,1408620)    =    62649.13
                                            Prob > F           =      0.0000
                                            R-squared          =      0.0808
```

income	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
_Iagegr_0	0	(omitted)				
_Iagegr_16	-60082.15	166.6691	-360.49	0.000	-60408.82	-59755.48
_Iagegr_20	-47592.64	172.1686	-276.43	0.000	-47930.09	-47255.2
_Iagegr_25	-24243.44	181.4771	-133.59	0.000	-24599.13	-23887.76
_Iagegr_35	-6055.089	215.5623	-28.09	0.000	-6477.584	-5632.594
_Iagegr_55	-3247.394	225.8159	-14.38	0.000	-3689.985	-2804.802
_Iagegr_65	-18389.89	299.2292	-61.46	0.000	-18976.37	-17803.41
_cons	66337.25	158.7966	417.75	0.000	66026.01	66648.48

Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Pearson's *r*

- Pearson's *r* is a measure of association for interval-ratio level variables

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

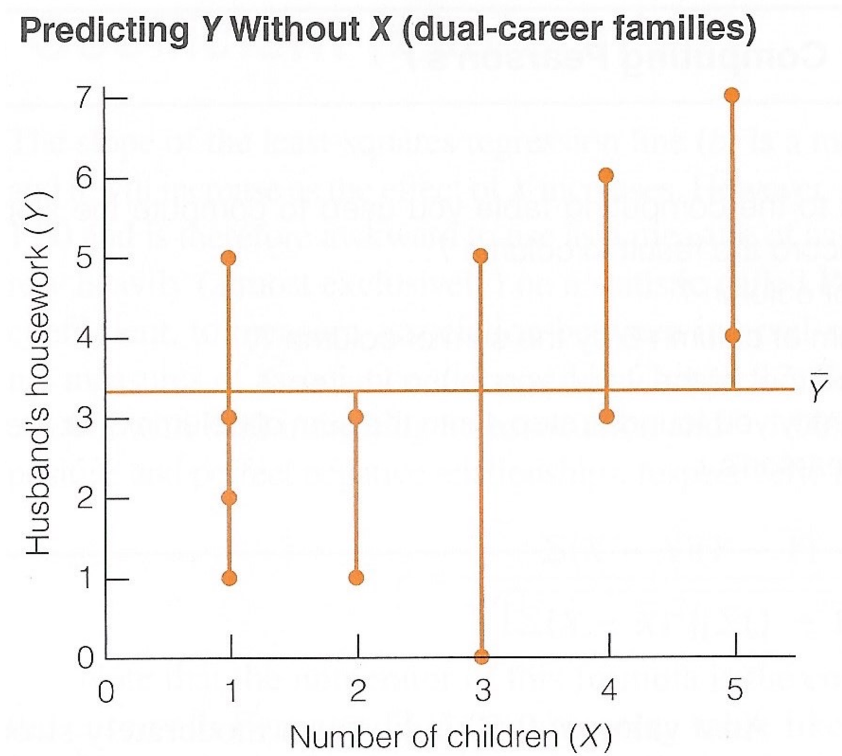
- Pearson's *r* indicate the direction of association
  - –1.00 indicates perfect negative association
  - 0.00 indicates no association
  - +1.00 indicates perfect positive association
- It doesn't have a direct interpretation of strength

Coefficient of determination (r^2)

- For a more direct interpretation of the strength of the linear association between two variables
  - Calculate the coefficient of determination (r^2)
- The coefficient of determination informs the percentage of the variation in Y explained by X
- It uses a logic similar to the proportional reduction in error (PRE) measure
  - Y is predicted while ignoring the information on X
    - Mean of the Y scores: $\bar{Y}$
  - Y is predicted taking into account information on X

Predicting Y without X

- The scores of any variable vary less around the mean than around any other point
  - The vertical lines from the actual scores to the predicted scores represent the amount of error of predicting Y while ignoring X

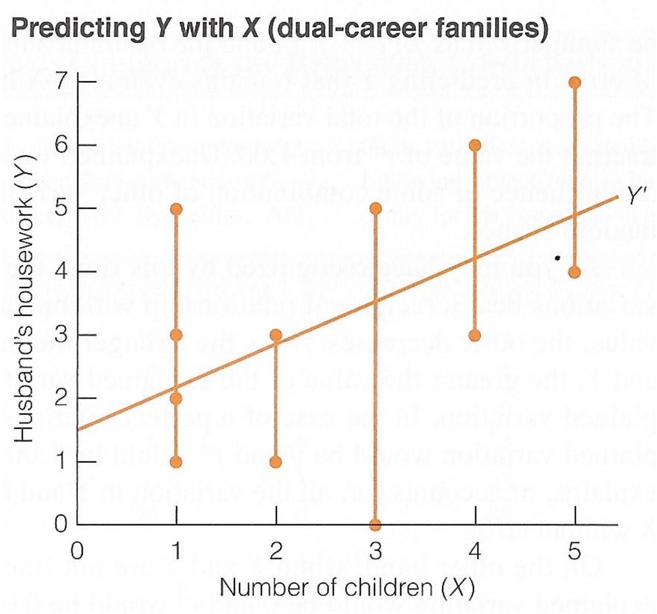


Source: Healey 2015, p.356.

Predicting Y with X

- If the Y and X have a linear association
  - Predicting scores on Y from the least-squares regression equation will incorporate knowledge of X
  - The vertical lines from each data point to the regression line represent the amount of error in predicting Y that remains even after X has been taking into account

Y’ = *a* + *b*X



Source: Healey 2015, p.357.

Estimating r^2

- **Total variation**: $\sum(Y - \bar{Y})^2$
  - Gives the error we incur by predicting ***Y without knowledge of X***
- **Explained variation**: $\sum(Y' - \bar{Y})^2 = \sum\left(\hat{Y} - \bar{Y}\right)^2$
  - Improvement in our ability to predict ***Y when taking X into account***
- r^2 indicates how much X helps us predict Y

$$r^2 = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum(Y - \bar{Y})^2} = \frac{Explained\ variation}{Total\ variation}$$

Source: Healey 2015, p.357.

Unexplained variation

- **<u>Unexplained variation</u>**: $\sum(Y - Y')^2 = \sum\left(Y - \hat{Y}\right)^2$
  - Difference between our best prediction of Y with X (Y') and the actual scores (Y)
  - It is the aggregation of vertical lines from the actual scores to the regression line
  - This is the amount of error in predicting Y that remains after X has been taken into account
  - It is caused by omitted variables, measurement error, and/or random chance
  - This is the residual of the regression

Example: Pearson's *r*

- Number of children (X) and hours per week husband spends on housework (Y)

Computation of Pearson's *r*

1	2	3	4	5	6	7
X	$X - \bar{X}$	Y	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
1	−1.67	1	−2.33	3.89	2.79	5.43
1	−1.67	2	−1.33	2.22	2.79	1.77
1	−1.67	3	−0.33	0.55	2.79	0.11
1	−1.67	5	1.67	−2.79	2.79	2.79
2	−0.67	3	−0.33	0.22	0.45	0.11
2	−0.67	1	−2.33	1.56	0.45	5.43
3	0.33	5	1.67	0.55	0.11	2.79
3	0.33	0	−3.33	−1.10	0.11	11.09
4	1.33	6	2.67	3.55	1.77	7.13
4	1.33	3	−0.33	−0.44	1.77	0.11
5	2.33	7	3.67	8.55	5.43	13.47
5	2.33	4	0.67	1.56	5.43	0.45
32	−0.04	40	0.04	18.32	26.68	50.68

Source: Healey 2015, p.354.

Example: calculate *r*

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

$$r = \frac{18.32}{\sqrt{(26.68)(50.68)}}$$

$$r = 0.50$$

Example: interpretation

- $r = 0.50$
  - The association between X and Y is positive
  - As the number of children increases, husbands' hours of housework per week also increases
- $r^2 = (0.50)^2 = 0.25$
  - The number of children explains 25% of the total variation in husbands' hours of housework per week
  - We make 25% fewer errors by basing the prediction of husbands' housework hours on number of children
    - We make 25% fewer errors by using the regression line
    - As opposed to ignoring the X variable and predicting the mean of Y for every case

Test Pearson's *r* for significance

- Use the five-step model

1. Make assumptions and meet test requirements
2. Define the null hypothesis (H_0)
3. Select the sampling distribution and establish the critical region
4. Compute the test statistic
5. Make a decision and interpret the test results

Step 1: Assumptions,requirements

- Random sampling
- Interval-ratio level measurement
- Bivariate normal distributions
- Linear association
- Homoscedasticity
  - The variance of Y scores is uniform for all values of X
  - If the Y scores are evenly spread above and below the regression line for the entire length of the line, the association is homoscedastic
- Normal sampling distribution

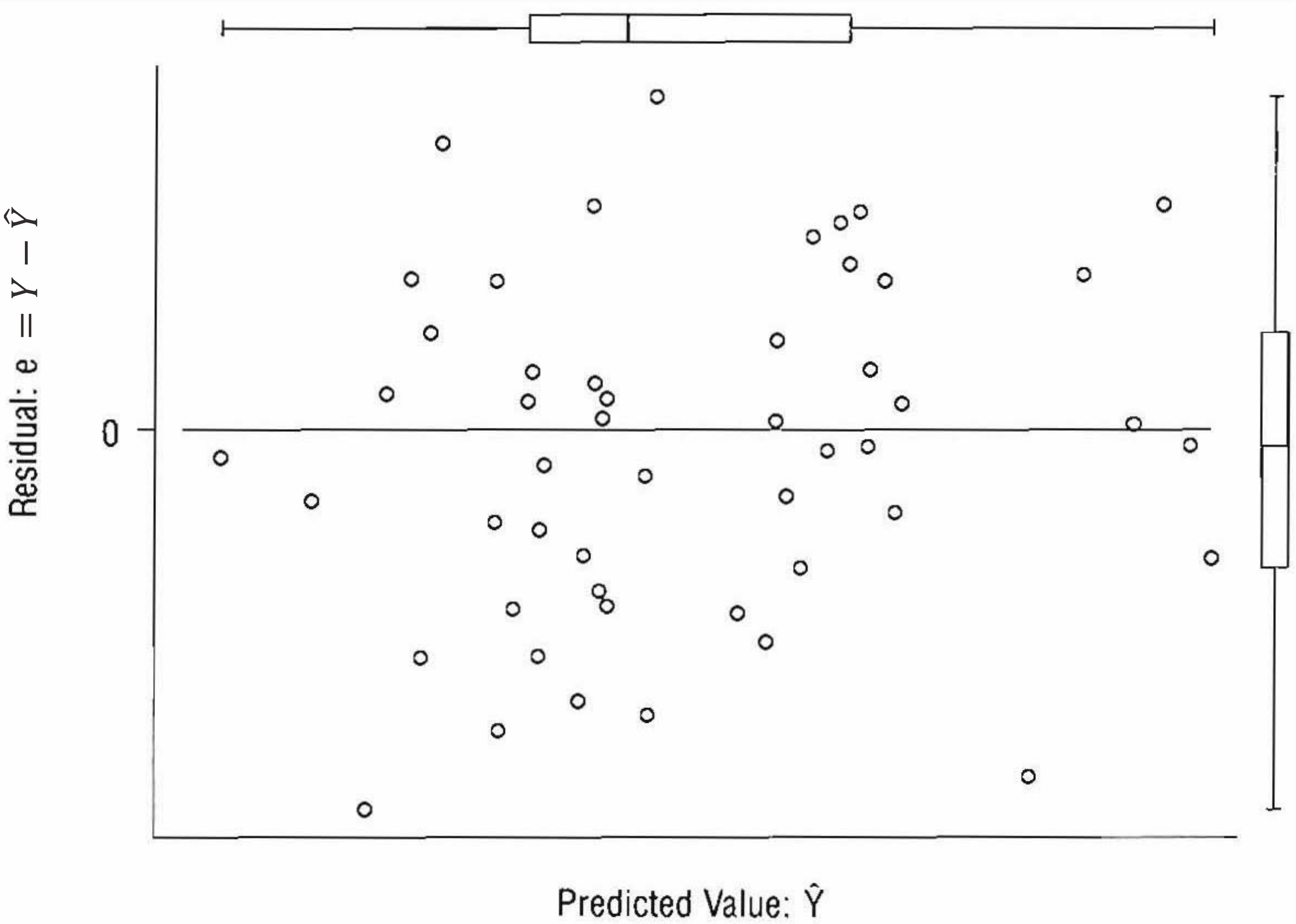


Figure 2.10 "All clear" e-versus-$\hat{Y}$ plot (artificial data).

Fonte: Hamilton, 1992: 52.

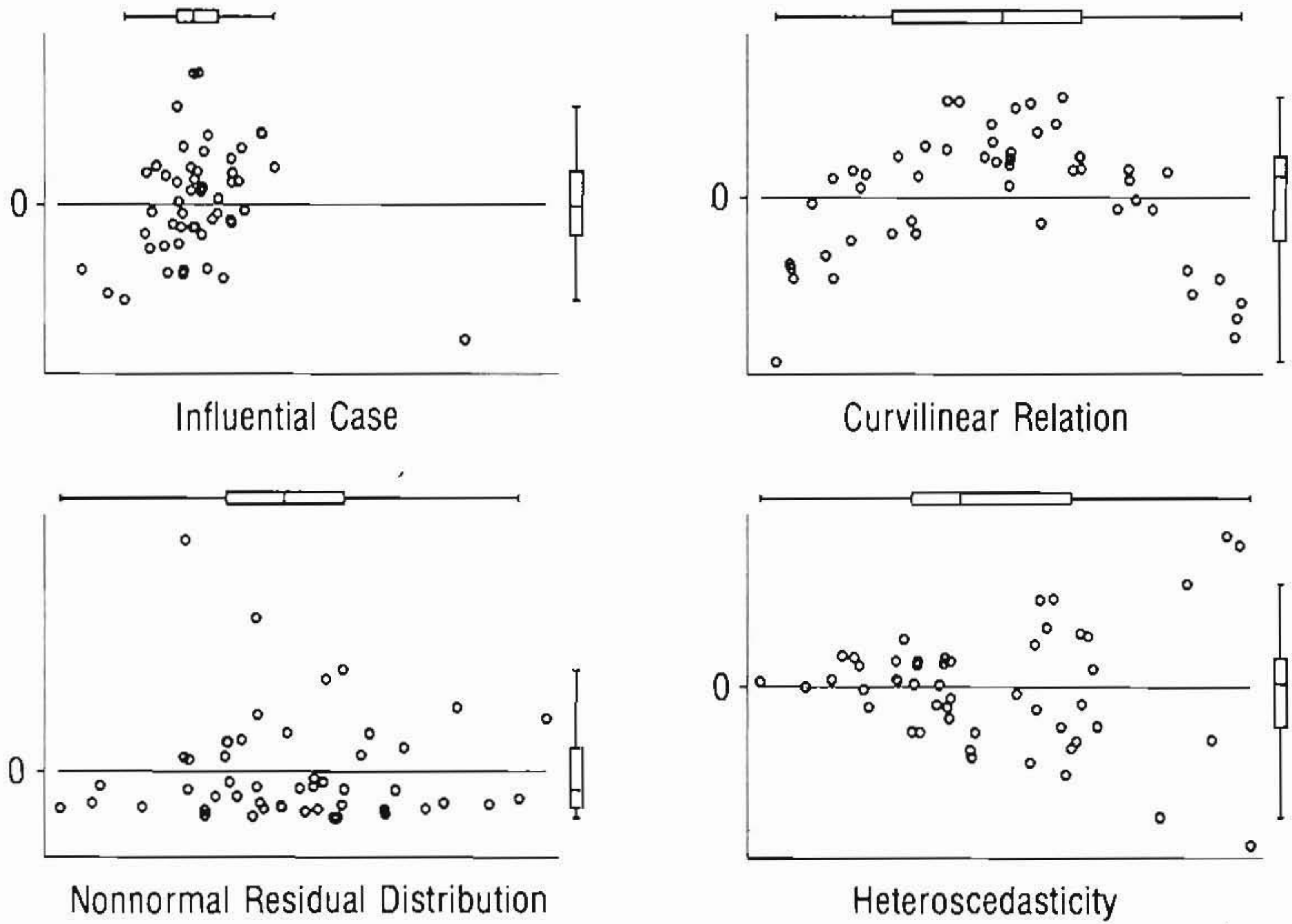


Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

Fonte: Hamilton, 1992: 53.

Step 2: Null hypothesis

- Null hypothesis, H_0: $\rho = 0$
  - H_0 states that there is no correlation between the number of children (X) and hours per week husband spends on housework (Y)

- Alternative hypothesis, H_1: $\rho \neq 0$
  - H_1 states that there is a correlation between the number of children (X) and hours per week husband spends on housework (Y)

Step 3: Distribution, critical region

- Sampling distribution: Student's *t*
- Alpha = 0.05 (two-tailed)
- Degrees of freedom = $n - 2 = 12 - 2 = 10$
- t(critical) = ±2.228

Step 4: Test statistic

$$t(obtained) = r\sqrt{\frac{n-2}{1-r^2}}$$

$$t(obtained) = (0.50)\sqrt{\frac{12-2}{1-(0.50)^2}}$$

$$t(obtained) = 1.83$$

Step 5: Decision, interpret

- *t(obtained)* = 1.83
  - This is not beyond the *t(critical)* = ±2.228
  - The *t*(obtained) does not fall in the critical region, so we ***do not reject*** the H_0
- The two variables are not correlated in the population
  - The correlation between number of children (X) and hours per week husband spends on housework (Y) is not statistically significant

Correlation matrix

- Table that shows the associations between all possible pairs of variables
  - Which are the strongest and weakest associations among birth rate, education, poverty, and teen births?

A Correlation Matrix Showing the Relationships Among Four Variables

	1 Birth Rate	2 Education	3 Poverty	4 Teen Births
1. Birth Rate	1.00	−0.24	0.16	0.26
2. Education	−0.24	1.00	−0.71	−0.78
3. Poverty	0.16	−0.71	1.00	0.88
4. Teen Births	0.26	−0.78	0.88	1.00

KEY: "Birth Rate" is number of births per 1000 population.
"Education" is percentage of the population with a college degree or more.
"Poverty" is percentage of families below the poverty line.
"Teen Births" is the percentage of all births to teenagers.

Source: Healey 2015, p.360.

GSS: Income, Age, Education

```
. ***Respondent's income income, age, education
. pwcorr conrinc age educ [aweight=wtssall], sig

             |  conrinc      age     educ
-------------+---------------------------
     conrinc |   1.0000
             |
             |
         age |   0.1852   1.0000
             |   0.0000
             |
        educ |   0.3387  -0.0131   1.0000
             |   0.0000   0.4857
             |

.
. ***Coefficient of determination (r-squared)
. ***Respondent's income and age
. di .1852^2
.03429904

.
. ***Coefficient of determination (r-squared)
. ***Respondent's income and education
. di .3387^2
.11471769
```

Source: 2016 General Social Survey.

Edited table

Table 1. Pearson's *r* and coefficient of determination (r^2) for the association of respondent's income with age and years of schooling, U.S. adult population, 2016

Independent variable	Pearson's *r*	Coefficient of determination (r^2)
Age	0.1852***	0.0343
Years of schooling	0.3387***	0.1147

Note: Pearson's *r* and coefficient of determination (r^2) were generated taking into account the survey weight of the General Social Survey. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Source: 2016 General Social Survey.

ACS: Income, Age, Education

```
. ***Wage and salary income, age, education
. pwcorr income age educ if income!=0 [aweight=perwt], sig

             |   income      age     educ
-------------+---------------------------
      income |   1.0000 
             |
             |
         age |   0.2118   1.0000 
             |   0.0000
             |
        educ |   0.3360   0.6768   1.0000 
             |   0.0000   0.0000
             |

.
. ***Coefficient of determination (r-squared)
. ***Income and age
. di .2118^2
.04485924

.
. ***Coefficient of determination (r-squared)
. ***Income and education
. di .3360^2
.112896
```

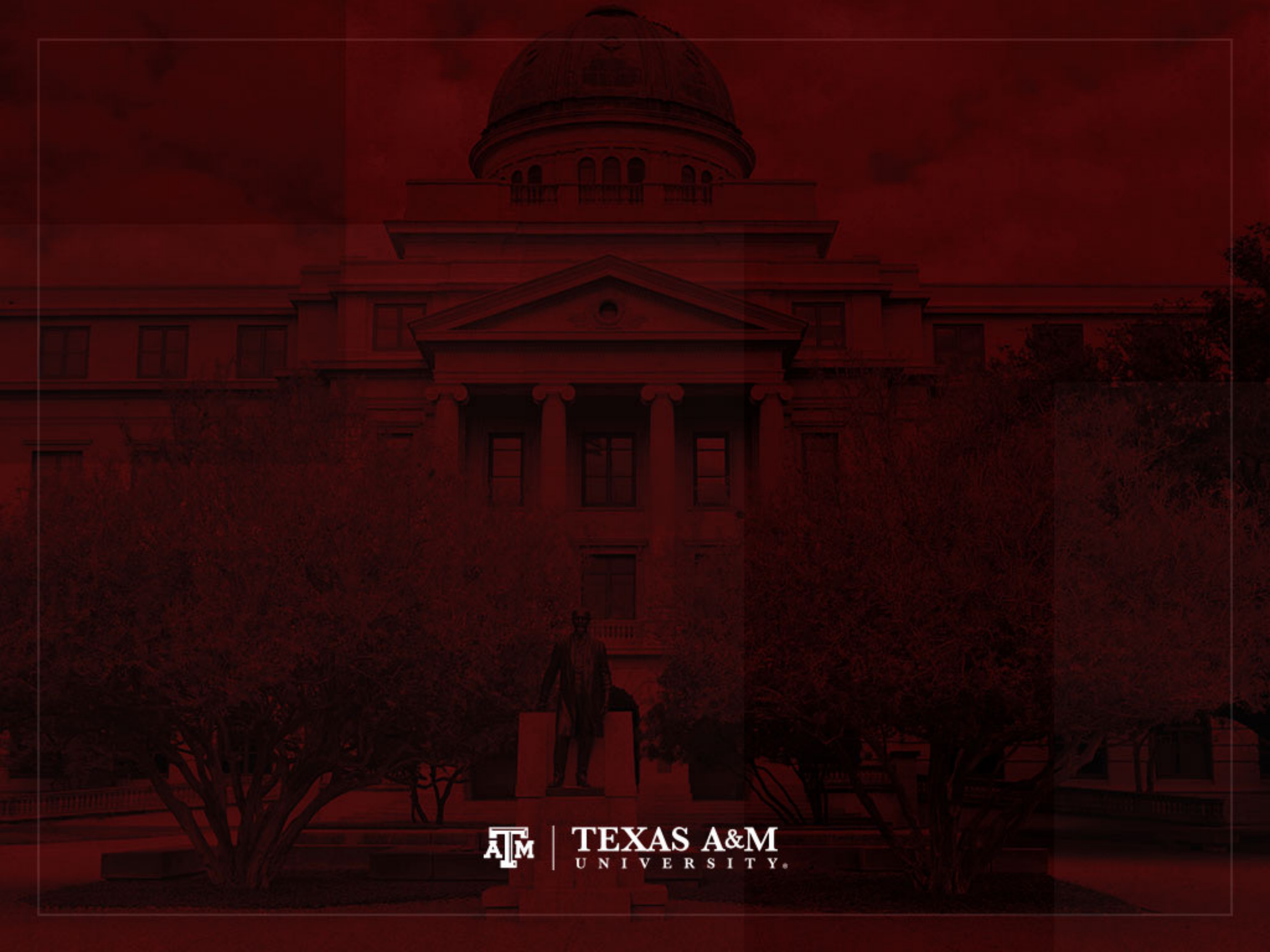
Source: 2018 American Community Survey.

Edited table

Table 1. Pearson's *r* and coefficient of determination (r^2) for the association of wage and salary income with age and educational attainment, United States, 2018

Independent variable	Pearson's *r*	Coefficient of determination (r^2)
Age	0.2118***	0.0449
Educational attainment	0.3360***	0.1129

Note: Pearson's *r* and coefficient of determination (r^2) were generated taking into account the survey weight of the American Community Survey. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY