Summary of lectures 6–7: Measures of association

Ernesto F. L. Amaral

September 15, 2022
Introduction to Sociological Data Analysis (SOCI 600)

Source: Healey, Joseph F. 2015. ”Statistics: A Tool for Social Research.” Stamford: Cengage Learning. 10th edition. Chapters 10 (pp. 247–275), 11 (pp. 276–306), 12 (pp. 308–341), 13 (pp. 342–378).

TEXAS A&M UNIVERSITY

Outline

- Measure of association for nominal-level variables
  - Chi Square
- Measure of association for ordinal-level variables
  - Spearman's Rho
- Measures of association for interval-ratio-level variables
  - Analysis of variance (ANOVA)
  - Scatterplots
  - Pearson's *r*

Measure of association for nominal-level variables

- Chi Square is a test of significance based on bivariate tables
  - Bivariate tables are also called cross tabulations, crosstabs, contingency tables
- We are looking for significant differences between
  - The actual cell frequencies observed in a table (f_o)
  - And those that would be expected by random chance or if cell frequencies were independent (f_e)

```
. ***Observed frequencies (fo)
. tab migrant sex
```

migrant	Sex: Male	Female	Total
Non-migrant	1,462,317	1,535,029	2,997,346
Internal migrant	88,155	81,712	169,867
International migrant	8,455	8,431	16,886
Total	1,558,927	1,625,172	3,184,099

```
.
. ***Expected frequencies (fe)
. tab migrant sex, exp nofreq
```

migrant	Sex: Male	Female	Total
Non-migrant	1467493.2	1529852.8	2997346.0
Internal migrant	83,166.5	86,700.5	169,867.0
International migrant	8,267.3	8,618.7	16,886.0
Total	1558927.0	1625172.0	3184099.0

$$f_e = \frac{Row\ marginal \times Column\ marginal}{n}$$

Chi square

$$f_e = \frac{Row\ marginal \times Column\ marginal}{n}$$

$$\chi^2(obtained) = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o = cell frequencies observed in the bivariate table

f_e = cell frequencies that would be expected if the variables were independent

Degrees of freedom (*df*) = (*r*–1)(*c*–1)

r = number of rows; *c* = number of columns

Limitations of chi square

- Difficult to interpret
  - When variables have many categories
  - Best when variables have four or fewer categories
- With small sample size
  - We cannot assume that chi square sampling distribution will be accurate
  - Small samples are those with a high percentage of cells with expected frequencies of 5 or less
- Like all tests of hypotheses
  - Chi square is sensitive to sample size
  - As *n* increases, obtained chi square increases
  - Large samples: Trivial relationships may be significant
- Statistical significance (statistical test) is not the same as substantive significance (importance, magnitude)

ACS: Migration by sex

- Is migration status different by sex?
  - The probability of not rejecting H_0 is small (p<0.00)
  - Migration status does depend on respondent's sex

```
. tab migrant sex, chi col
```

Key
frequency
column percentage

migrant	Sex Male	Female	Total
Non-migrant	1,462,317 93.80	1,535,029 94.45	2,997,346 94.13
Internal migrant	88,155 5.65	81,712 5.03	169,867 5.33
International migrant	8,455 0.54	8,431 0.52	16,886 0.53
Total	1,558,927 100.00	1,625,172 100.00	3,184,099 100.00

Pearson chi2(2) = 630.3698 Pr = 0.000

Source: 2018 American Community Survey.

Percentages, *N*, missing cases

```
. tab migrant sex [fweight=perwt], col // percentage & population size
```

Key
frequency
column percentage

migrant	Sex Male	Female	Total
Non-migrant	**149645178**	**155097362**	**304742540**
	93.99	**94.38**	**94.19**
Internal migrant	**8660884**	**8318528**	**16979412**
	5.44	**5.06**	**5.25**
International migrant	**900980**	**918570**	**1819550**
	0.57	**0.56**	**0.56**
Total	**159207042**	**164334460**	**323541502**
	100.00	**100.00**	**100.00**

```
. tab migrant sex, m // missing cases
```

migrant	Sex Male	Female	Total
Non-migrant	**1,462,317**	**1,535,029**	**2,997,346**
Internal migrant	**88,155**	**81,712**	**169,867**
International migrant	**8,455**	**8,431**	**16,886**
.	**15,691**	**14,749**	**30,440**
Total	**1,574,618**	**1,639,921**	**3,214,539**

Source: 2018 American Community Survey.

Edited table

Table 1. Distribution of U.S. population by migration status and sex, 2018

Migration status	Male	Female	Total
Non-migrant	93.99	94.38	94.19
Internal migrant	5.44	5.06	5.25
International migrant	0.57	0.56	0.56
Total	**100.00**	**100.00**	**100.00**
Population size (N)	159,207,042	164,334,460	323,541,502
Sample size (n)	1,558,927	1,625,172	3,184,099
Missing cases	15,691	14,749	30,440
Chi square (df=2)	630.37	p-value=0.000	

Source: 2018 American Community Survey.

ACS: Education by race/ethnicity

- Does education attainment vary by race/ethnicity?
  - The probability of not rejecting H_0 is small ($p<0.01$)
  - Education attainment is dependent on race/ethnicity

```
. tab educgr raceth [fweight=perwt], col nofreq
```

educgr	raceth White	African A	Hispanic	Asian	Native Am	Ohter rac	Total
Less than high school	23.19	30.14	49.76	27.23	20.66	47.04	35.24
High school	26.55	29.72	26.11	16.23	34.00	17.85	26.09
Some college	20.38	22.79	14.40	12.29	25.15	16.42	17.82
College	19.92	11.04	7.12	23.26	15.36	12.51	13.78
Graduate school	9.95	6.31	2.62	20.99	4.83	6.17	7.07
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

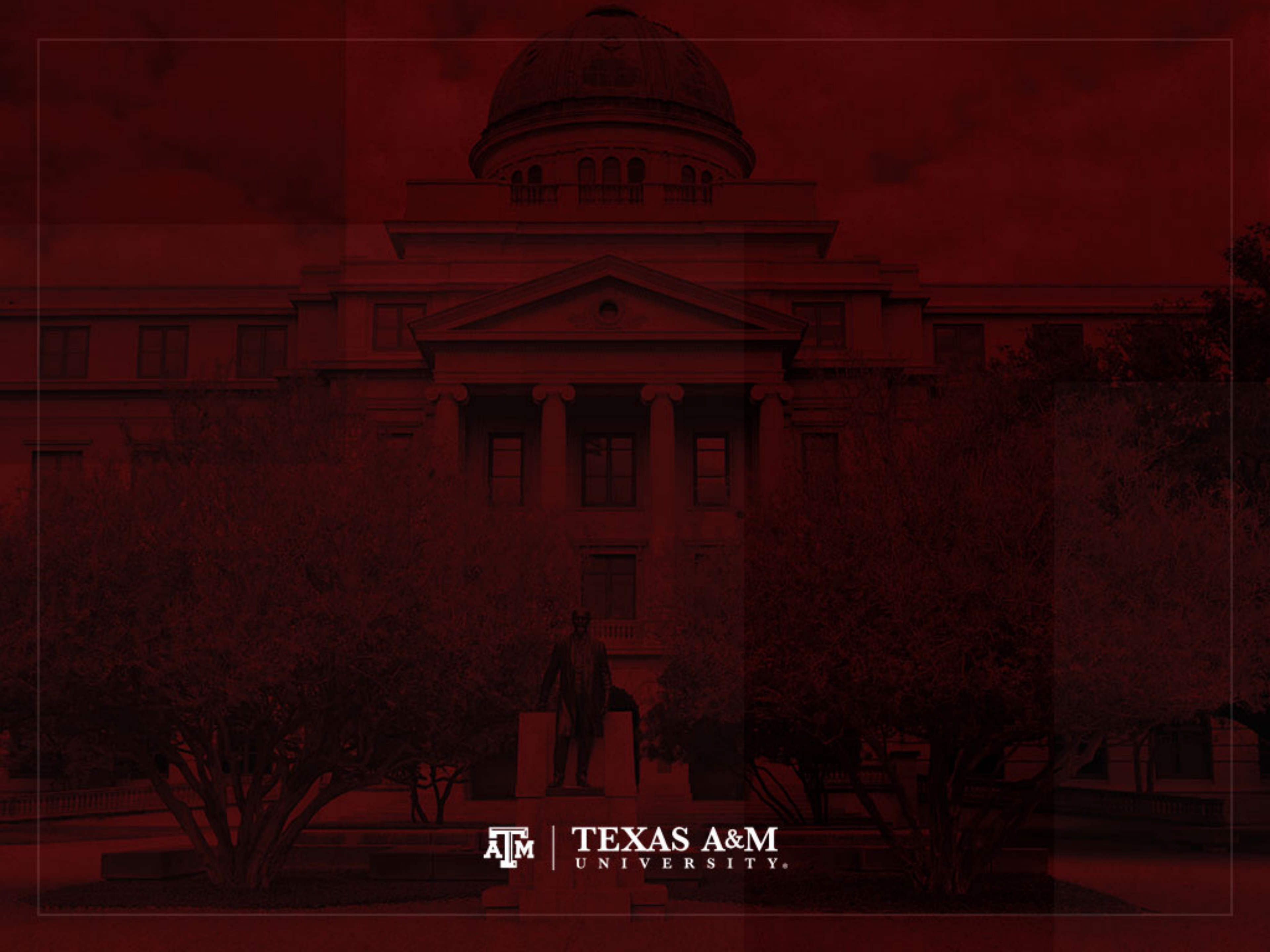
```
. svy: tab educgr raceth, col
(running tabulate on estimation sample)

Number of strata =       212                  Number of obs   =    272,776
Number of PSUs   =   114,016                  Population size = 28,995,881
                                              Design df       =    113,804
 Pearson:
   Uncorrected   chi2(20)          =  3.03e+04
   Design-based  F(19.11,  2.2e+06)=  676.9183     P = 0.0000
```

Edited table

Table 1. Percentage distribution of population by educational attainment and race/ethnicity, Texas, 2019

Educational attainment	Non-Hispanic White	Non-Hispanic Black	Hispanic	Non-Hispanic Asian	Non-Hispanic Native American	Other races	Total
Less than high school	23.19	30.14	49.76	27.23	20.66	47.04	35.24
High school	26.55	29.72	26.11	16.23	34.00	17.85	26.09
Some college	20.38	22.79	14.40	12.29	25.15	16.42	17.82
College	19.92	11.04	7.12	23.26	15.36	12.51	13.78
Graduate school	9.95	6.31	2.62	20.99	4.83	6.17	7.07
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00
Population size (*N*)	11,929,840	3,445,104	11,527,412	1,444,220	79,394	569,911	28,995,881
Chi square (*df* = 20)	3.03e+04						
Design-based *F* (19.11, 2.2e+06)	676.92						
p-value	0.0000						

Source: 2019 American Community Survey.


TEXAS A&M
UNIVERSITY

Measure of association for ordinal-level variables

- Spearman's Rho (r_s) is a measure of association for ordinal-level variables with a broad range of different scores and few ties between cases on either variable
- Computing Spearman's Rho, Spearman's ρ (r_s)
  1. It ranks cases from high to low on each variable
  2. It uses ranks, not the scores, to calculate Rho

$$r_s = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

where $\sum D^2$ is the sum of the squared differences in ranks

Interpreting Spearman's Rho

- Spearman's Rho is positive
  - As the rank of one variable increases, the rank of the other variable also increases
- Spearman's Rho is negative
  - As the rank of one variable increases, the rank of the other variable decreases

Example of Spearman's Rho (r_s)

Scores on Involvement in Jogging and Self-Esteem

Jogger	Involvement in Jogging (*X*)	Self-Esteem (*Y*)
Wendy	18	15
Debbie	17	18
Phyllis	15	12
Stacey	12	16
Evelyn	10	6
Tricia	9	10
Christy	8	8
Patsy	8	7
Marsha	5	5
Lynn	1	2

Source: Healey 2015, p.329.

Computing Spearman's Rho (r_s)

Computing Spearman's Rho

	Involvement (X)	Rank	Self-Image (Y)	Rank	D	D^2
Wendy	18	1	15	3	−2	4
Debbie	17	2	18	1	1	1
Phyllis	15	3	12	4	−1	1
Stacey	12	4	16	2	2	4
Evelyn	10	5	6	8	−3	9
Tricia	9	6	10	5	1	1
Christy	8	7.5	8	6	1.5	2.25
Patsy	8	7.5	7	7	0.5	0.25
Marsha	5	9	5	9	0	0
Lynn	1	10	2	10	0	0
					$\Sigma D = 0$	$\Sigma D^2 = 22.5$

Source: Healey 2015, p.330.

Result of Spearman's Rho (r_s)

- In the column headed D^2, each difference is squared to eliminate negative signs
- The sum of this column is $\sum D^2$, and this quantity is entered directly into the formula

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(22.5)}{10(100-1)} = 0.86$$

Interpreting Spearman's Rho (r_s)

- Rho is positive, therefore jogging and self-image share a positive association
  - As jogging rank increases, self-image rank also increases
- On its own, Rho does not have a good strength interpretation
  - But Rho^2 is a PRE measure...

PRE measures

- The logic of Proportional Reduction in Error (PRE) measures is based on two predictions
  - First prediction, E_1: How many errors in predicting the value of the dependent variable (Y) do we make if we **ignore** information about the independent variable (X)
  - Second prediction, E_2: How many errors in predicting the value of the dependent variable (Y) do we make if we take the independent variable (X) into account
- If the variables are associated, we should make fewer errors of the second kind (E_2) than we make of the first kind (E_1)

Spearman's Rho^2

- Rho^2 is a PRE measure

- For this example, $Rho^2 = (0.86)^2 = 0.74$

- We would make 74% fewer errors if we used the rank of jogging (X) to predict the rank on self-image (Y) compared to if we ignored the rank on jogging

ACS: Education by age

- Is educational attainment different by age group?

```
. tab educgr agegr, col
```

Key
frequency
column percentage

educgr	agegr 0	16	20	25	35	45	55	65	Total
Less than high school	571,701 99.97	89,702 52.61	10,262 5.51	25,198 6.49	30,960 8.25	35,040 8.52	39,879 8.44	74,522 11.67	877,264 27.29
High school	157 0.03	59,928 35.15	71,447 38.39	119,445 30.78	111,837 29.79	141,857 34.50	184,217 38.97	259,161 40.58	948,049 29.49
Some college	0 0.00	20,766 12.18	72,420 38.92	93,352 24.05	85,507 22.78	91,946 22.36	107,832 22.81	123,053 19.27	594,876 18.51
College	0 0.00	105 0.06	29,469 15.84	102,919 26.52	85,850 22.87	85,309 20.75	84,454 17.86	98,425 15.41	486,531 15.14
Graduate school	0 0.00	0 0.00	2,495 1.34	47,199 12.16	61,261 16.32	57,053 13.87	56,382 11.93	83,429 13.06	307,819 9.58
Total	571,858 100.00	170,501 100.00	186,093 100.00	388,113 100.00	375,415 100.00	411,205 100.00	472,764 100.00	638,590 100.00	3,214,539 100.00

Source: 2018 American Community Survey.

Spearman's Rho in Stata

```
. spearman educgr agegr

 Number of obs = 3214539
Spearman's rho =       0.4405

Test of Ho: educgr and agegr are independent
    Prob > |t| =       0.0000
```

$Rho^2 = (0.4405)^2 = 0.1940$

Source: 2018 American Community Survey.

ACS: Percentages with weight

- Use column percentages from this table

```
. tab educgr agegr [fweight=perwt], col
```

Key
frequency   *column percentage*

educgr	agegr 0	16	20	25	35	45	55	65	Total
Less than high school	64932988 99.97	9592001 55.79	1233939 5.67	3146621 6.95	3999381 9.59	4047164 9.73	4092972 9.68	6713748 12.81	97758814 29.88
High school	17628 0.03	5676286 33.02	8516860 39.11	14302836 31.59	12637092 30.31	14222739 34.20	16105938 38.09	20704168 39.51	92183547 28.18
Some college	0 0.00	1915448 11.14	8462363 38.86	11380862 25.14	9705561 23.28	9436932 22.69	9710019 22.96	10211276 19.48	60822461 18.59
College	0 0.00	8720 0.05	3288424 15.10	11420420 25.22	9104449 21.84	8441402 20.30	7508620 17.76	8093763 15.44	47865798 14.63
Graduate school	0 0.00	0 0.00	276404 1.27	5026278 11.10	6240807 14.97	5444101 13.09	4864635 11.51	6684594 12.76	28536819 8.72
Total	64950616 100.00	17192455 100.00	21777990 100.00	45277017 100.00	41687290 100.00	41592338 100.00	42282184 100.00	52407549 100.00	327167439 100.00

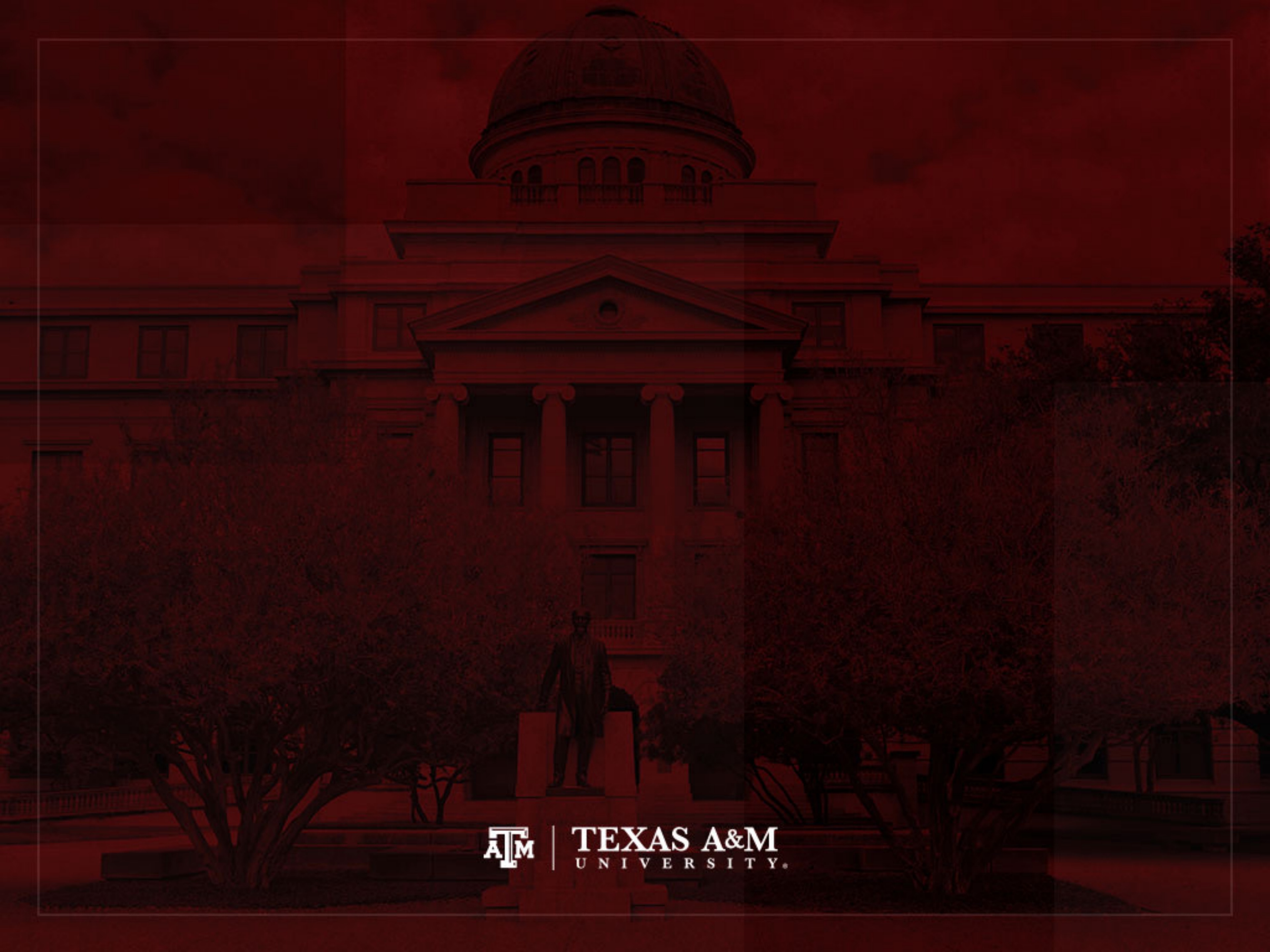
Source: 2018 American Community Survey.

Edited table

Table 1. Distribution of U.S. population by educational attainment and age group, 2018

Educational attainment	Age group							
	0–15	**16–19**	**20–24**	**25–34**	**35–44**	**45–54**	**55–64**	**65+**
Less than high school	99.97	55.79	5.67	6.95	9.59	9.73	9.68	12.81
High school	0.03	33.02	39.11	31.59	30.31	34.20	38.09	39.51
Some college	0.00	11.14	38.86	25.14	23.28	22.69	22.96	19.48
College	0.00	0.05	15.10	25.22	21.84	20.30	17.76	15.44
Graduate school	0.00	0.00	1.27	11.10	14.97	13.09	11.51	12.76
Total	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**
Population size (N)	64,950,616	17,192,455	21,777,990	45,277,017	41,687,290	41,592,338	42,282,184	52,407,549
Sample size (n)	571,858	170,501	186,093	388,113	375,415	411,205	472,764	638,590
Spearman's Rho	0.4405	p-value: 0.000						

Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Measures of association for interval-ratio-level variables

- Analysis of variance (ANOVA)
- Scatterplots
- Pearson's *r*

Analysis of variance (ANOVA)

- ANOVA can be used in situations where the researcher is interested in the differences in sample means across three or more categories
  - How do Protestants, Catholics, and Jews vary in terms of number of children?
  - How do Republicans, Democrats, and Independents vary in terms of income?
  - How do older, middle-aged, and younger people vary in terms of frequency of church attendance?

Extension of *t*-test

- We can think of ANOVA as an extension of *t*-test for more than two groups
  - Are the differences between the samples large enough to reject the null hypothesis and justify the conclusion that the populations represented by the samples are different?
- Null hypothesis, H_0
  - H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$
  - All population means are similar to each other
- Alternative hypothesis, H_1
  - At least one of the populations means is different

Between and within differences

- If the H_0 is true, the sample means should be about the same value
  - If the H_0 is true, there will be little difference between sample means
- If the H_0 is false
  - There should be substantial differences **between** sample means (between categories)
  - There should be relatively little difference **within** categories
    - The sample standard deviations should be small within groups

Likelihood of rejecting H_0

- The greater the difference **between** categories (as measured by the means)
  - Relative to the differences **within** categories (as measured by the standard deviations)
  - The more likely the H_0 can be rejected

- When we reject H_0
  - We are saying there are differences between the **populations** represented by the sample

Computation of ANOVA

1. Find total sum of squares (SST)

$$SST = \sum X_i^2 - n\bar{X}^2$$

2. Find sum of squares between (SSB)

$$SSB = \sum n_k(\bar{X}_k - \bar{X})^2$$

   - SSB = sum of squares between categories
   - n_k= number of cases in a category
   - $\bar{X}_k$= mean of a category

3. Find sum of squares within (SSW)

SSW = SST – SSB

4. Degrees of freedom

$$\text{dfb} = k - 1$$

- dfb = degrees of freedom between
- k = number of categories

$$\text{dfw} = n - k$$

- dfw = degrees of freedom within
- n = total number of cases
- k = number of categories

Final estimations

5. Find mean square estimates

$$Mean\ square\ between = \frac{SSB}{dfb}$$

$$Mean\ square\ within = \frac{SSW}{dfw}$$

6. Find the *F* ratio

$$F(obtained) = \frac{Mean\ square\ between}{Mean\ square\ within}$$

Limitations of ANOVA

- Requires interval-ratio level measurement of the dependent variable
- Requires roughly equal numbers of cases in the categories of the independent variable
- Statistically significant differences are not necessarily important (small magnitude)
- The alternative (research) hypothesis is not specific
  - It only asserts that at least one of the population means differs from the others

ACS: Income by race/ethnicity

- We know the average income by race/ethnicity

```
. tabstat income if income!=0 & income!=. [fweight=perwt], by(raceth) stat(mean sd n)

Summary for variables: income
Group variable: raceth
```

raceth	Mean	SD	N
White	63199.24	74601.04	6081513
African American	40079.03	40410.99	1766063
Hispanic	36595.08	38076.88	5250789
Asian	66528.88	73827.69	776722
Native American	44246.01	57666.53	44743
Other races	46151.98	58649.93	235029
Total	50285.44	60567.56	1.42e+07

- Does at least one category of race/ethnicity have average income different than the others?
  - This is not a perfect example for ANOVA, because race/ethnicity does not have equal numbers of cases across its categories

```
. svy, subpop(if income!=0 & income!=.): mean income, over(raceth)
(running mean on estimation sample)

. estat sd
```

(correct standard deviation)

Over	Mean	Std. dev.
c.income@ raceth		
White	63199.24	81952.97
African A..	40079.03	33729.03
Hispanic	36595.08	34417.96
Asian	66528.88	71633.26
Native Am..	44246.01	57876.89
Other races	46151.98	56501.55

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

. estat sd
```

	Mean	Std. dev.
income	50285.44	59920.72

Source: 2019 American Community Survey, Texas.

ANOVA in Stata

- The probability of not rejecting H_0 is small (p<0.01)
  - At least one category of the race/ethnicity variable has average income different than the others with a 99% confidence level
  - However, ANOVA does not inform which category has an average income significantly different than the others

```
. oneway income raceth if income!=0 & income!=. [aweight=perwt]

                        Analysis of variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      2.2032e+13      5   4.4065e+12   1259.17     0.0000
 Within groups      4.5608e+14 130325   3.4995e+09
                                                  (statistical significance)
------------------------------------------------------------------------
    Total           4.7811e+14 130330   3.6685e+09

Bartlett's equal-variances test: chi2(5) =  1.2e+04    Prob>chi2 = 0.000
```

Source: 2019 American Community Survey, Texas.

ACS: n, N

```
. ***Sample size of each category of race/ethnicity and missing cases
. tab raceth if income!=0 & income!=., m
```

raceth	Freq.	Percent	Cum.
White	69,043	52.98	52.98
African American	11,574	8.88	61.86
Hispanic	40,359	30.97	92.82
Asian	6,879	5.28	98.10
Native American	424	0.33	98.43
Other races	2,052	1.57	100.00
Total	130,331	100.00	

```
. ***Population size of each category of race/ethnicity
. tab raceth if income!=0 & income!=. [fweight=perwt]
```

(correct percentage distribution)

raceth	Freq.	Percent	Cum.
White	6,081,513	42.96	42.96
African American	1,766,063	12.48	55.44
Hispanic	5,250,789	37.10	92.54
Asian	776,722	5.49	98.02
Native American	44,743	0.32	98.34
Other races	235,029	1.66	100.00
Total	14,154,859	100.00	

Source: 2019 American Community Survey, Texas.

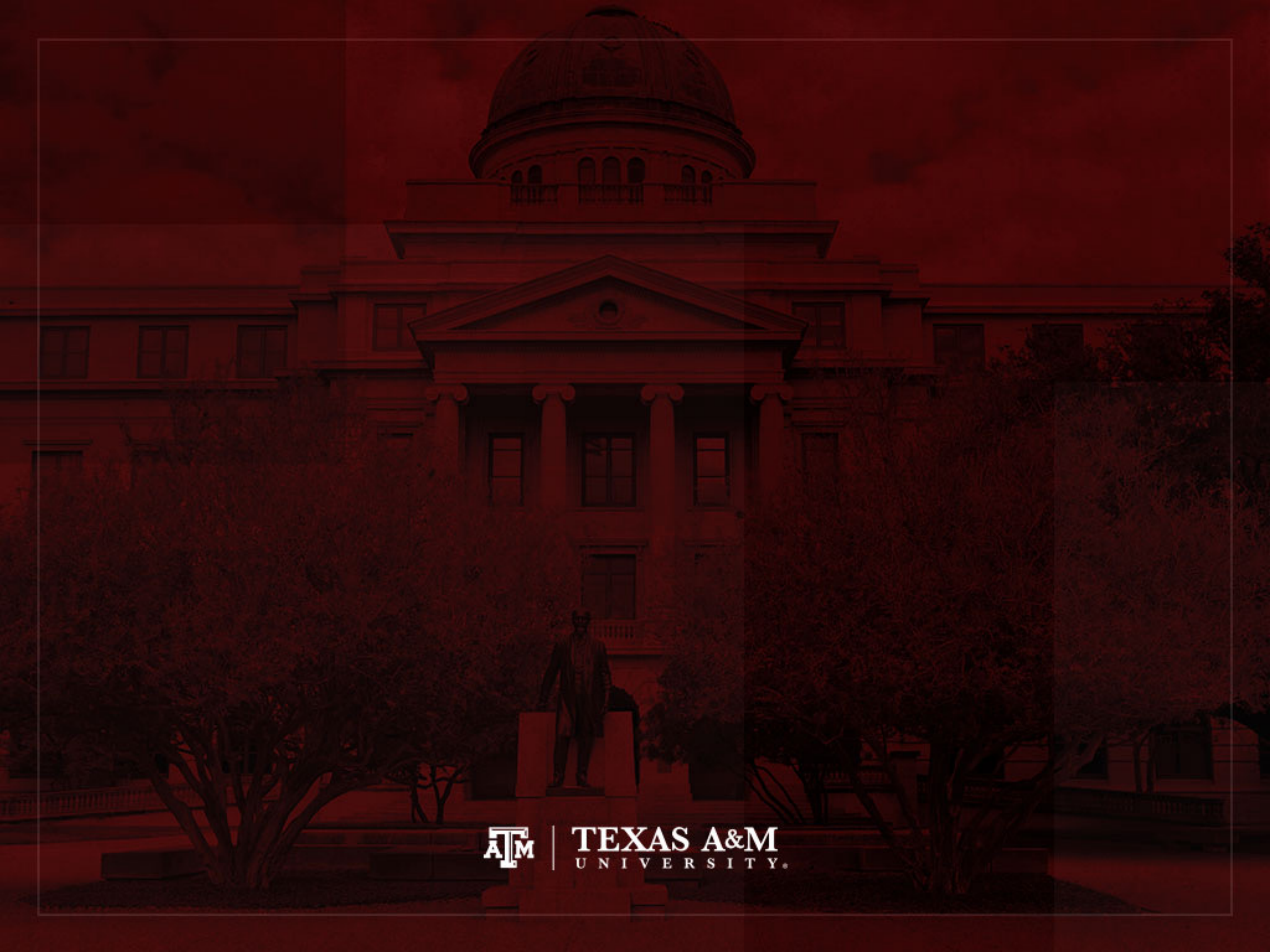
Edited table

Table 1. One-way analysis of variance for wage and salary income by race/ethnicity, Texas, 2019

Race/ethnicity	Income		Population percentage
	Mean	Standard deviation	
White	63,199.24	81,952.97	42.96
African American	40,079.03	33,729.03	12.48
Hispanic	36,595.08	34,417.96	37.10
Asian	66,528.88	71,633.26	5.49
Native American	44,246.01	57,876.89	0.32
Other races	46,151.98	56,501.55	1.66
Total	50,285.44	59,920.72	100.00
Population size	—	—	14,154,859
Sample size	—	—	130,331

ANOVA	Sum of squares	Degrees of freedom	Mean of squares	F-test	Prob > F
Between groups	2.20e+13	5	4.41e+12	1,259.17	0.0000
Within groups	4.56e+14	130,325	3.50e+09		
Total	4.78e+14	130,330	3.67e+09		

Source: 2019 American Community Survey.


TEXAS A&M
UNIVERSITY

Scatterplots

- Scatterplots can be used to answer these questions

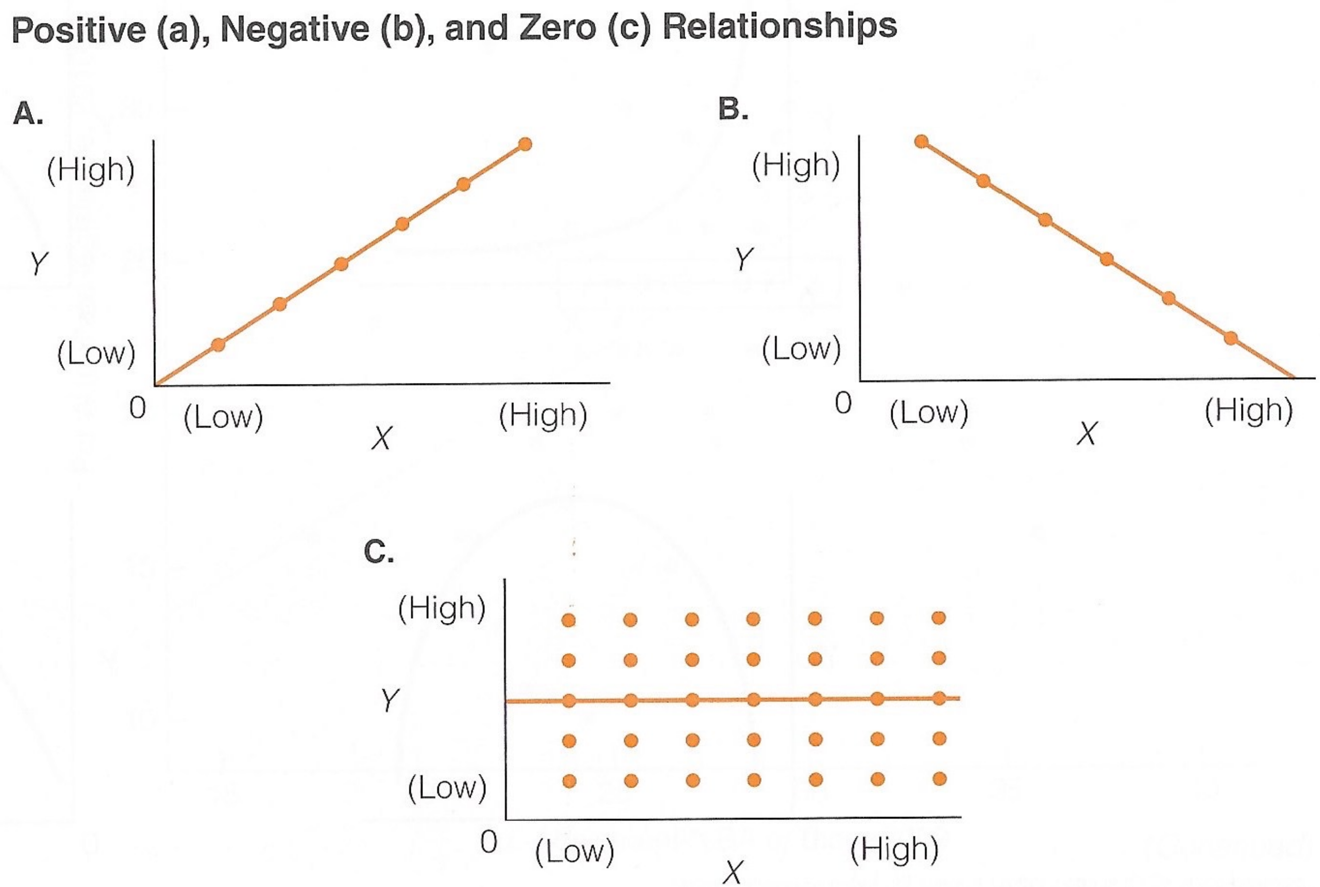
1. Is there an association?

2. How strong is the association?

3. What is the pattern of the association?

Pattern of the association

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

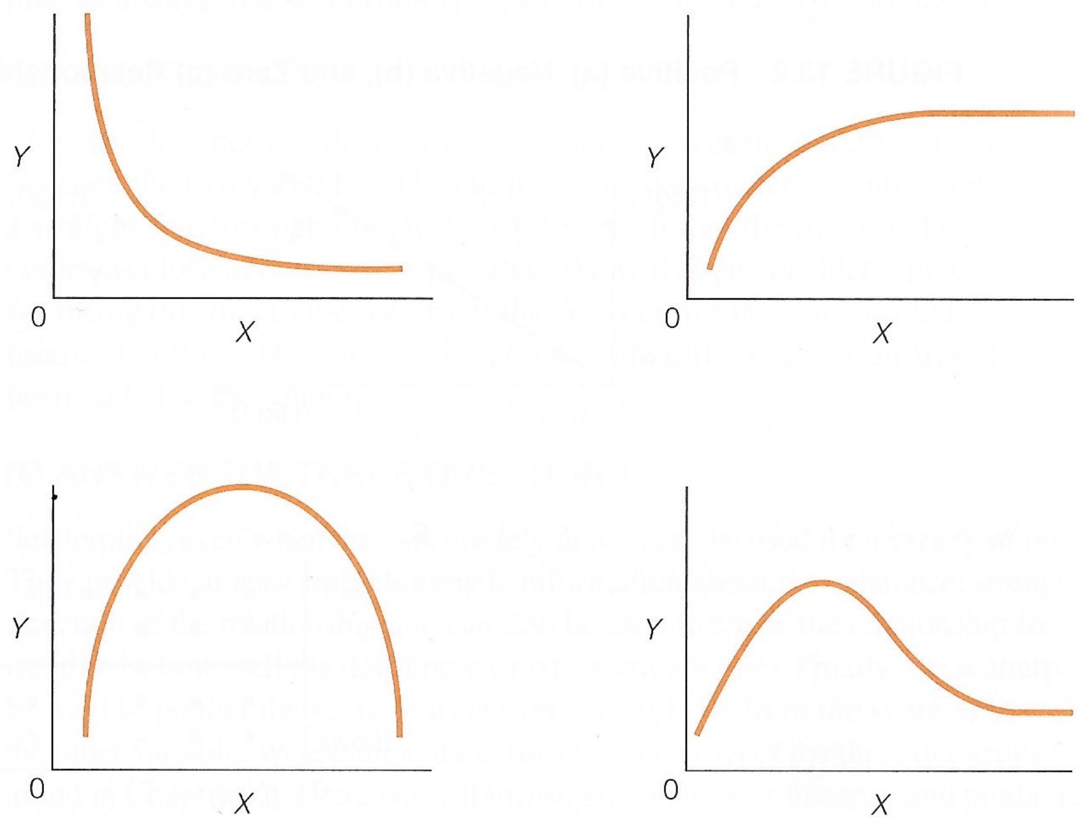


Source: Healey 2015, p.345.

Nonlinear associations

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

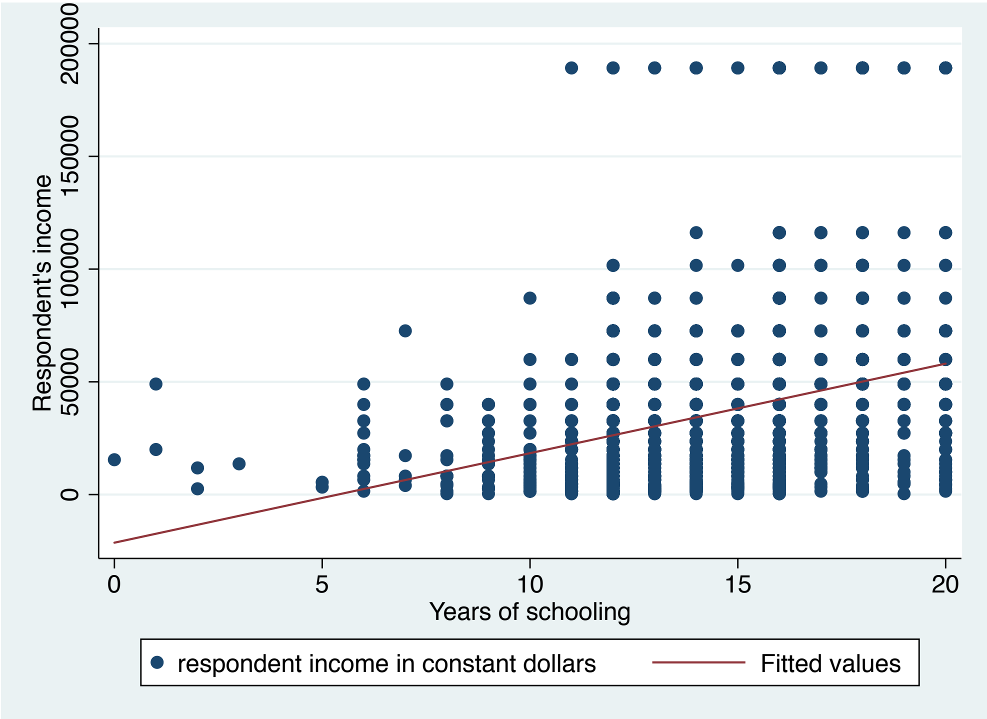
Some Nonlinear Relationships



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
```

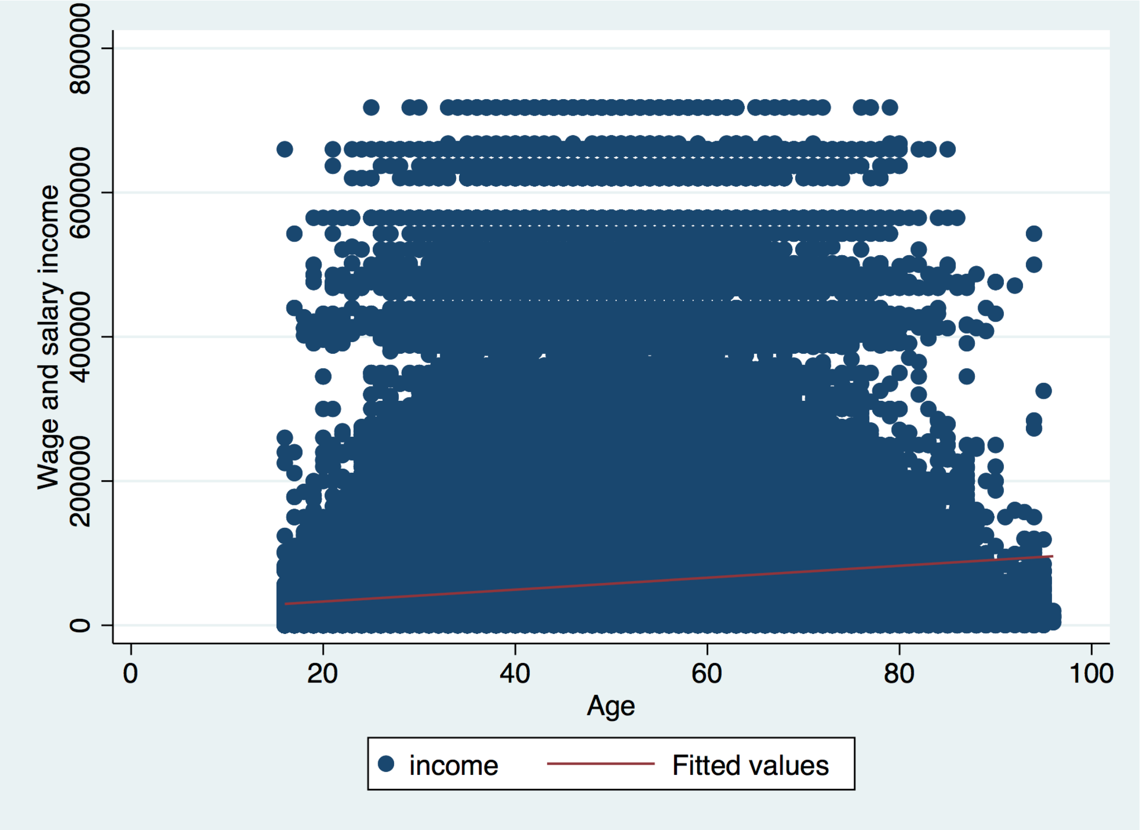
```
. svy: reg conrinc educ
(running regress on estimation sample)

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)
```

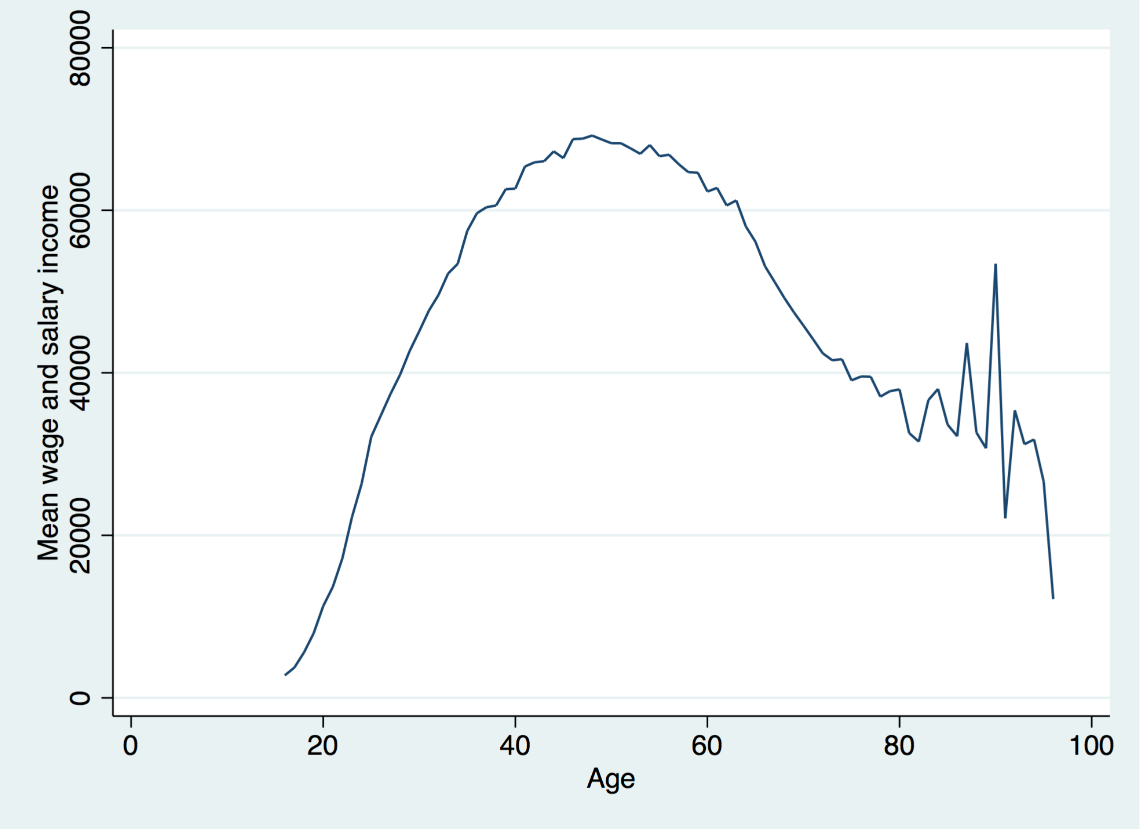
```
. svy, subpop(if income!=. & income!=0): reg income age
(running regress on estimation sample)

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

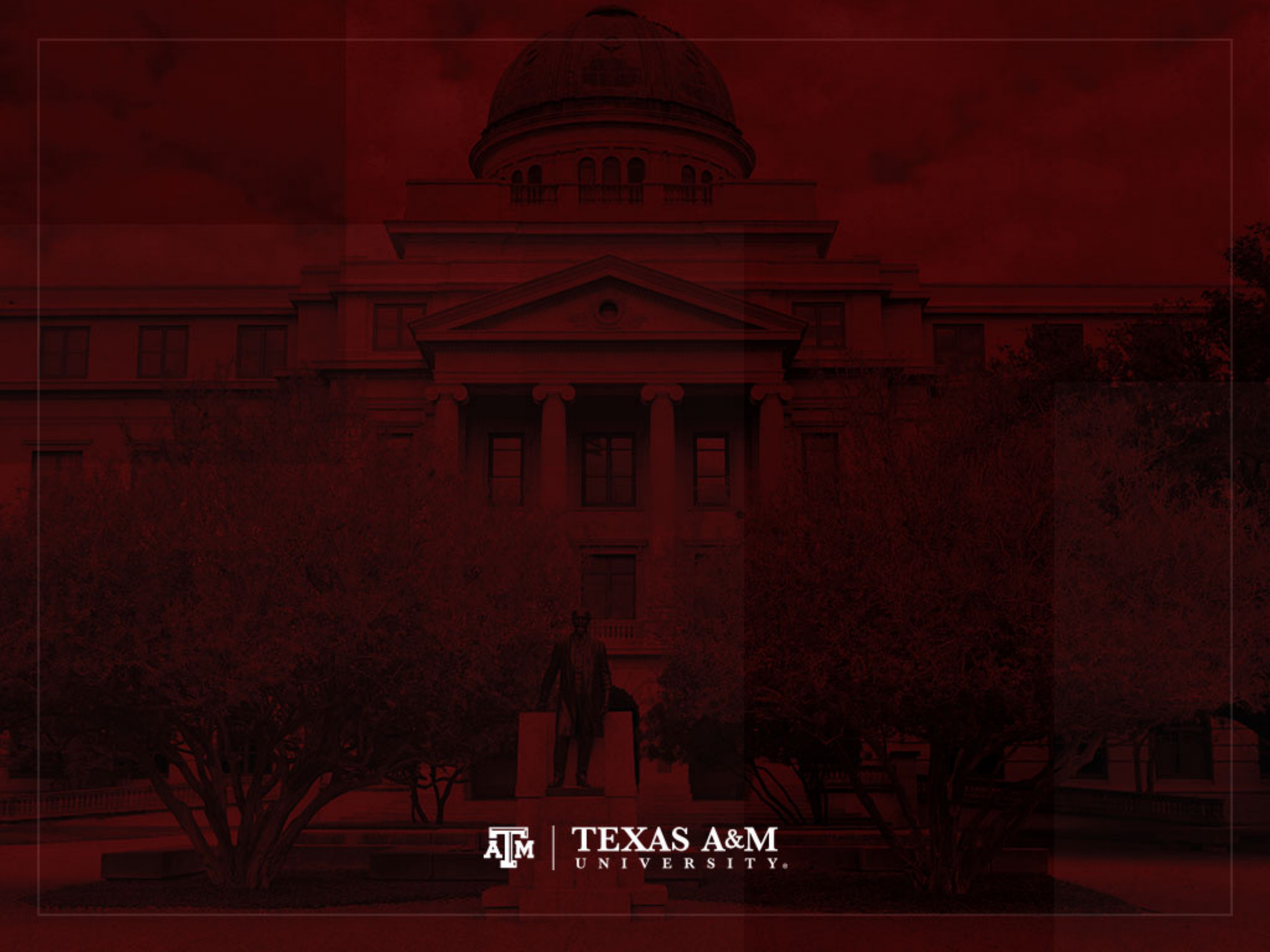
Source: 2018 American Community Survey.

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.


ATM
TEXAS A&M
UNIVERSITY

Pearson's *r*

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

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

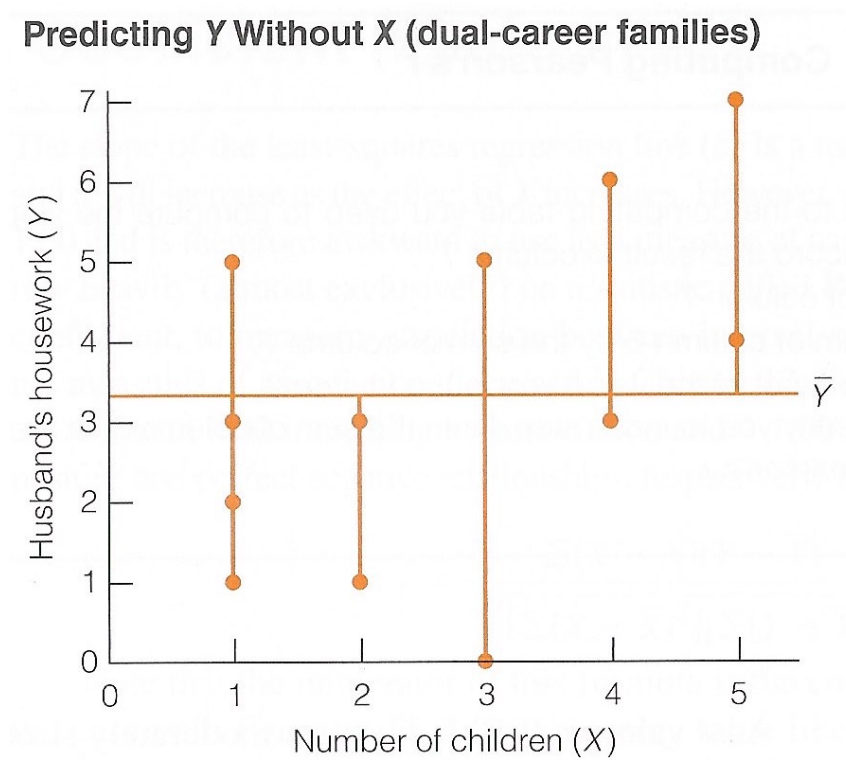
- Pearson's *r* indicate the direction of association
  - –1.00 indicates perfect negative association
  - 0.00 indicates no association
  - +1.00 indicates perfect positive association
- It doesn't have a direct interpretation of strength

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

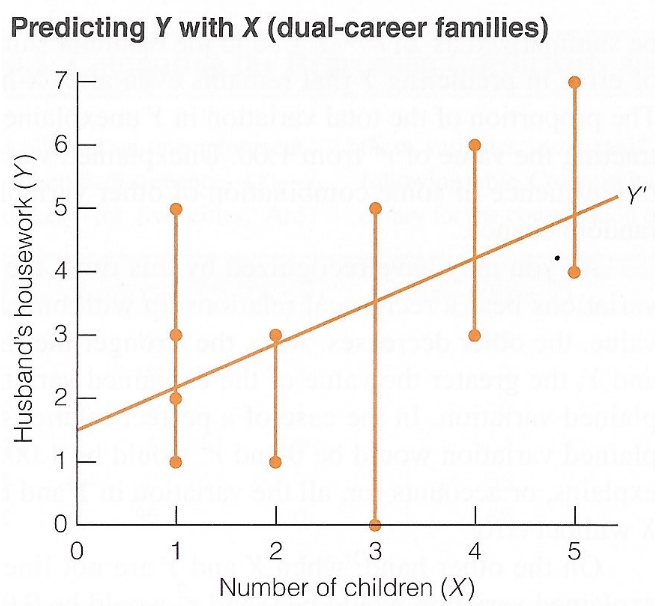


Source: Healey 2015, p.356.

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

$Y' = a + bX$



Estimating r^2

- **<u>Total variation</u>**: $\sum(Y - \bar{Y})^2$
  - Gives the error we incur by predicting ***Y without knowledge of X***
- **<u>Explained variation</u>**: $\sum(Y' - \bar{Y})^2 = \sum\left(\hat{Y} - \bar{Y}\right)^2$
  - 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

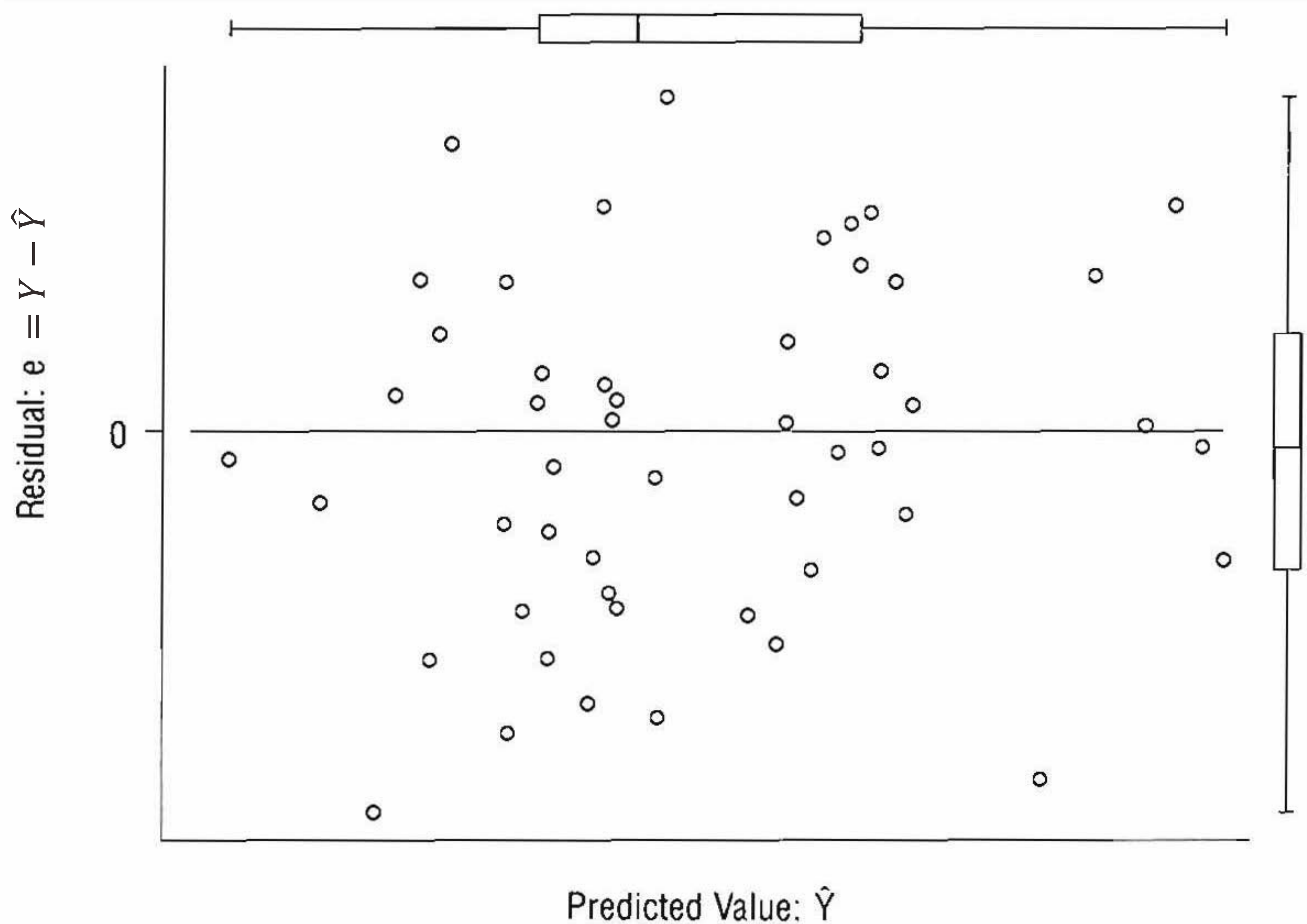


Figure 2.10 "All clear" e-versus-$\hat{Y}$ plot (artificial data).

Fonte: Hamilton, 1992: 52.

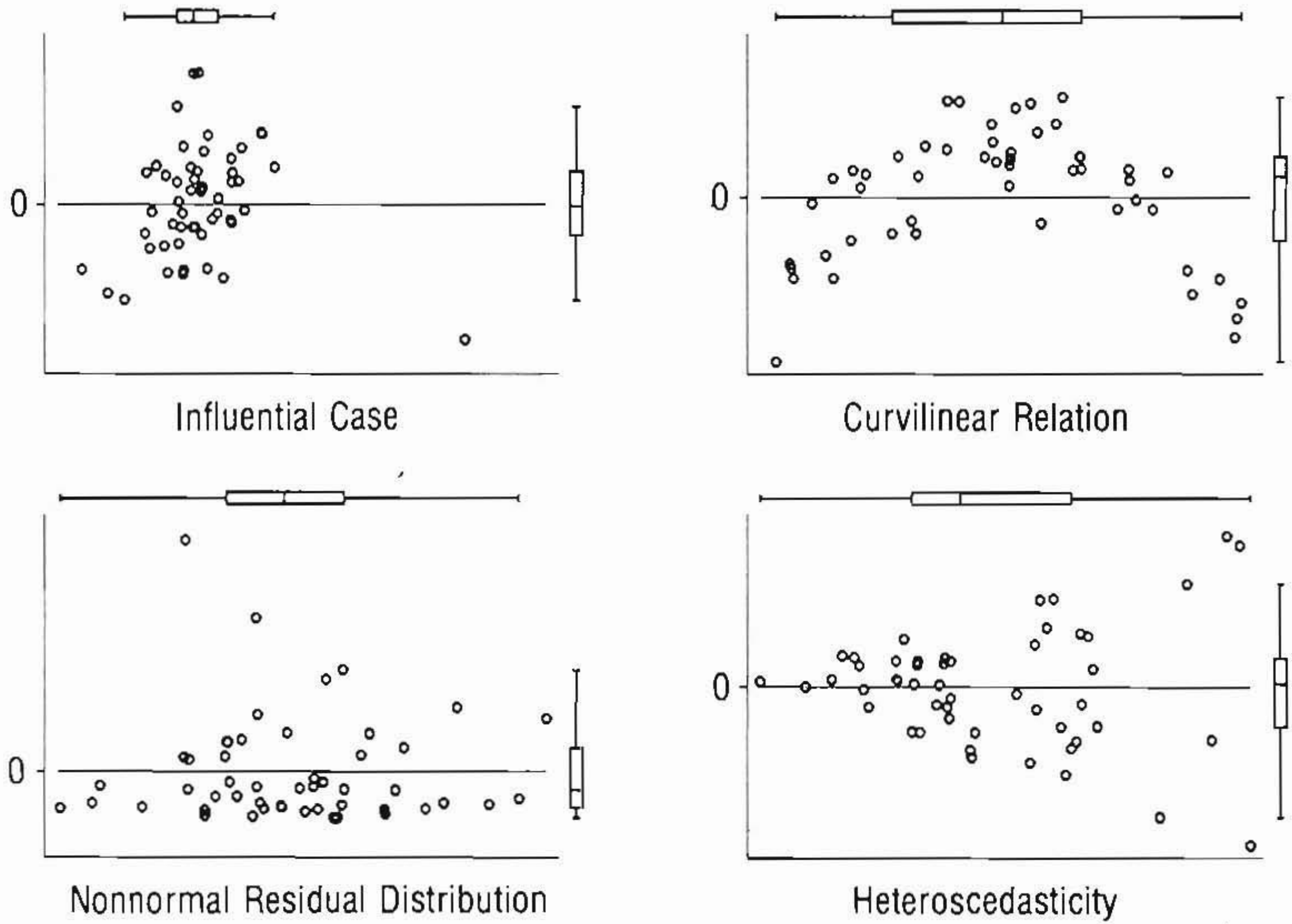


Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

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	2	3	4
	Birth Rate	Education	Poverty	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
```

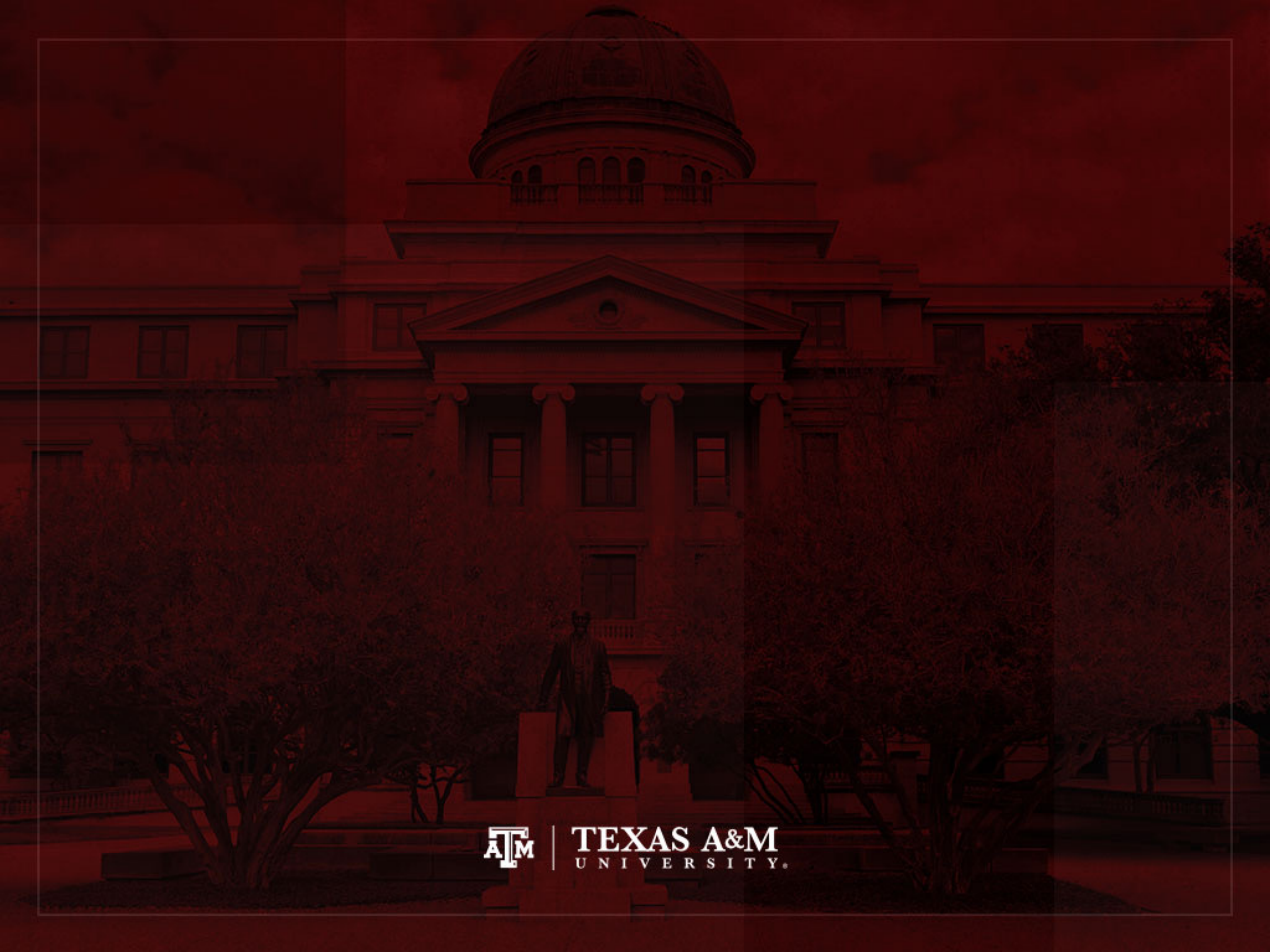
Source: 2018 American Community Survey.

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.


ATM
TEXAS A&M
UNIVERSITY