Measures of association

Ernesto F. L. Amaral

February 20–March 6, 2020
Introduction to Social Statistics Using Stata

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

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 example: Chi square

- 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	Sex: 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	Sex: 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**

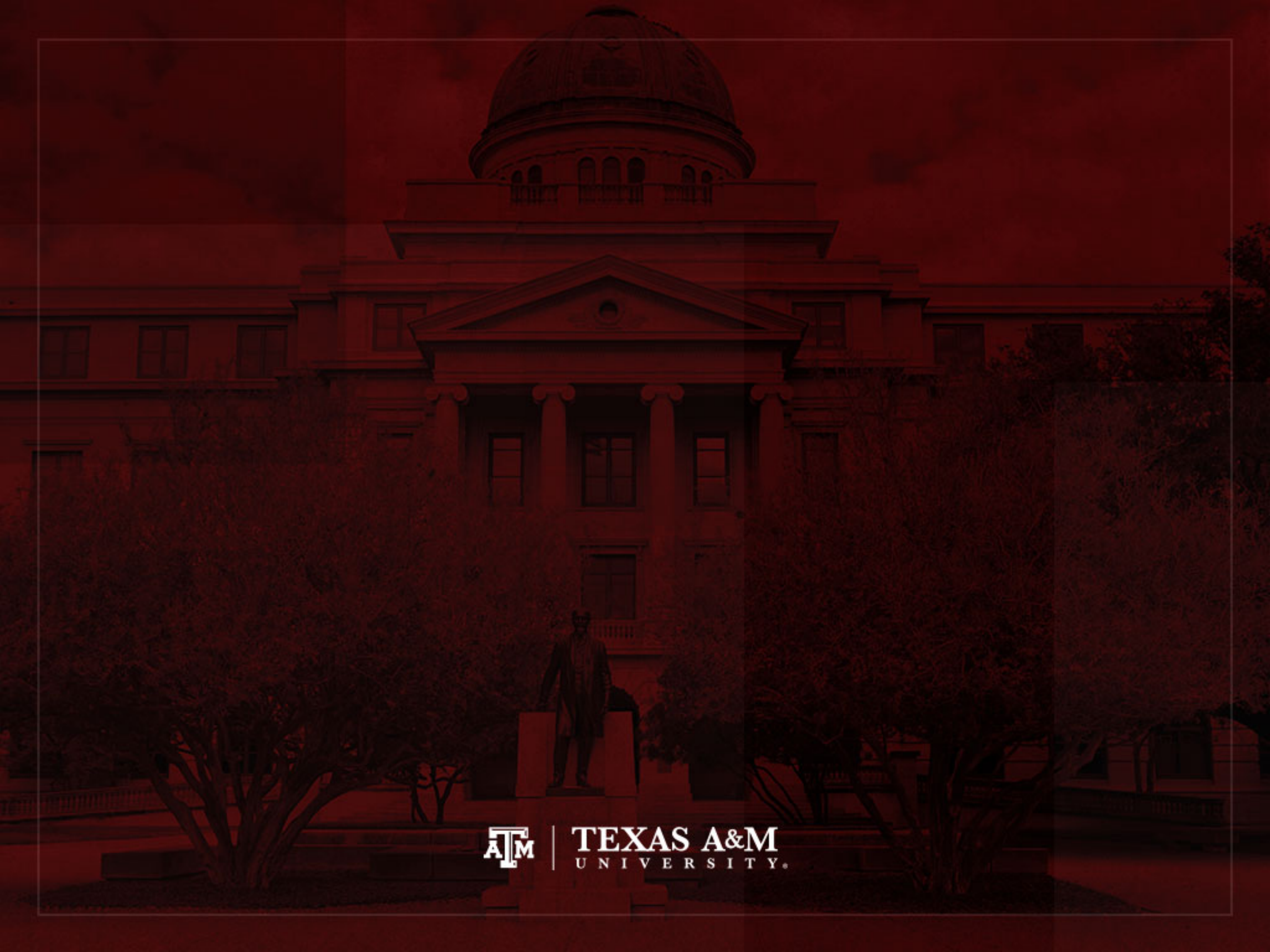
Source: 2018 American Community Survey.

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.


TEXAS A&M
UNIVERSITY

Measure of association for ordinal-level variables

- 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

ACS example: Spearman's Rho

- 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
```

Source: 2018 American Community Survey.

ACS example: percentages

- 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

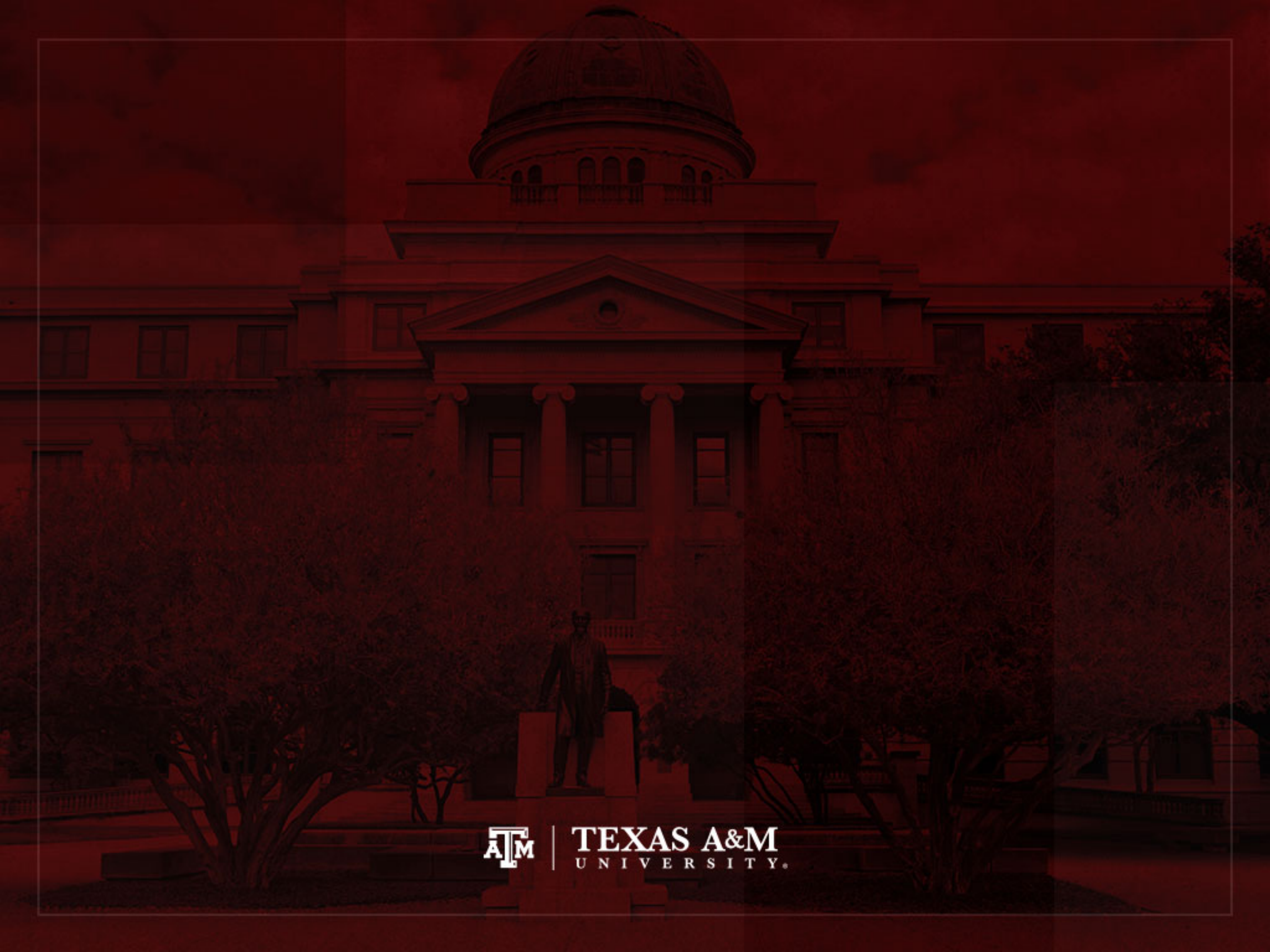
Source: 2018 American Community Survey.

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.


ATM
TEXAS A&M
UNIVERSITY

Measures of association for interval-ratio-level variables

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

Scatterplots

- Scatterplots can be used to answer these questions

1. Is there an association?

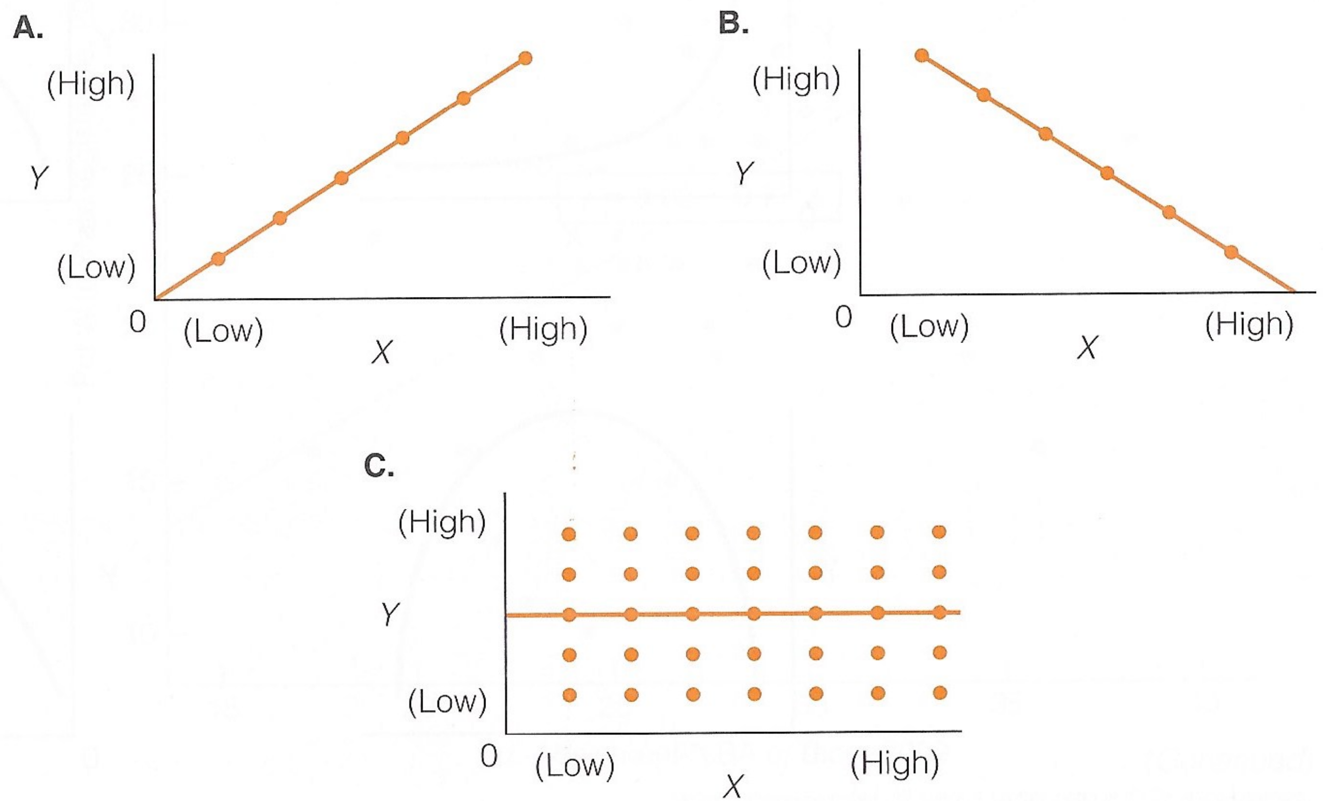
2. How strong is the association?

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

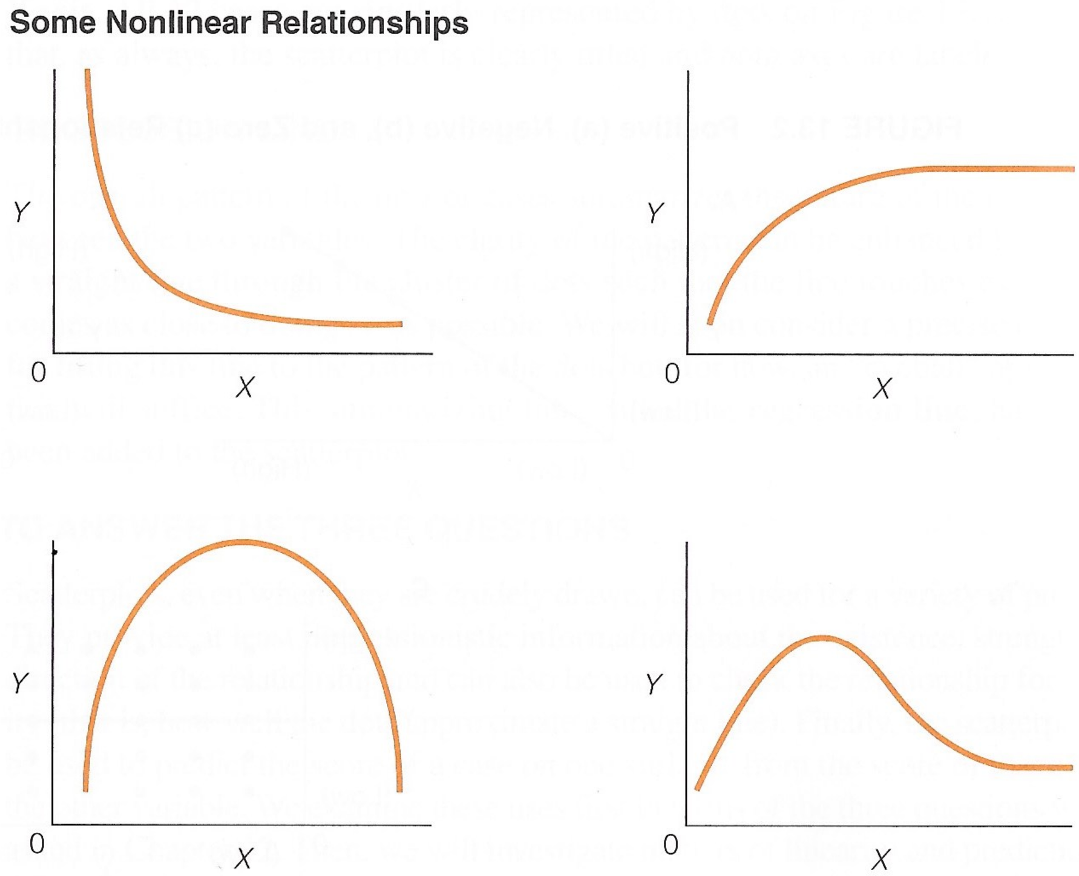
Positive (a), Negative (b), and Zero (c) Relationships



Source: Healey 2015, p.345.

Nonlinear associations

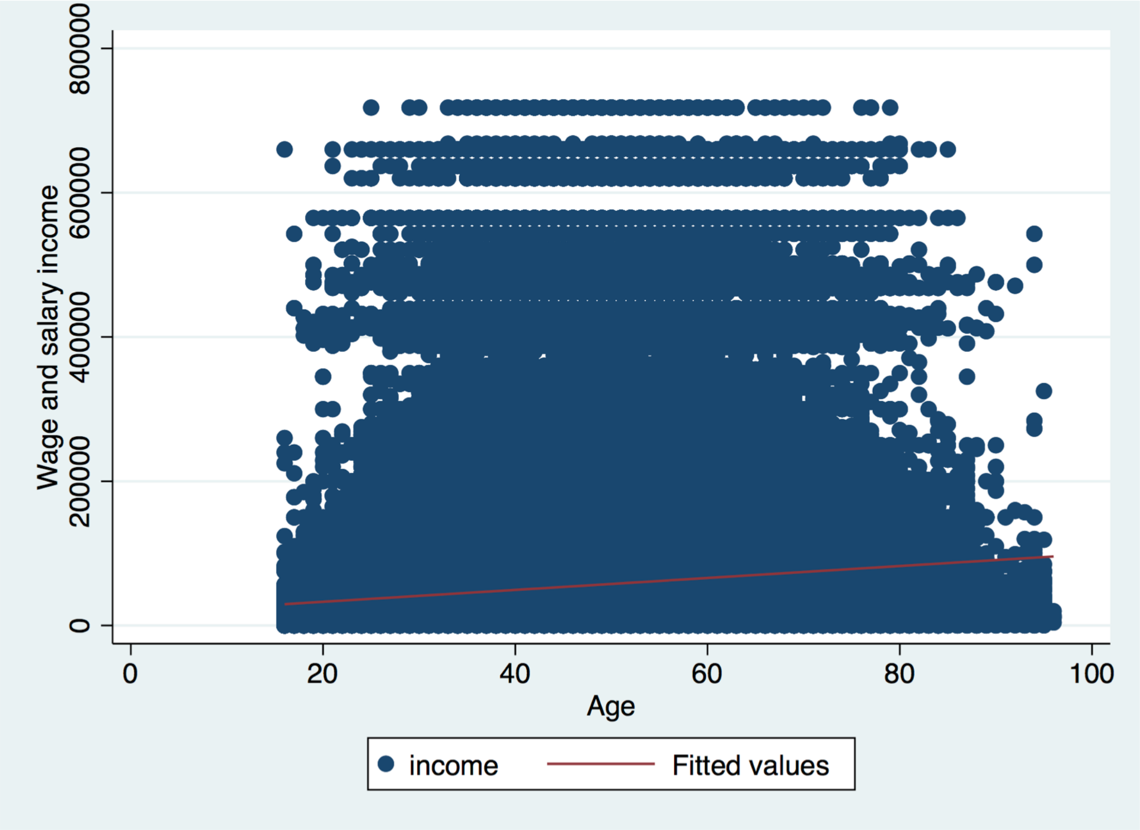
- In a nonlinear association, the dots do not form a straight line pattern



Source: Healey 2015, p.346.

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).

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)
```

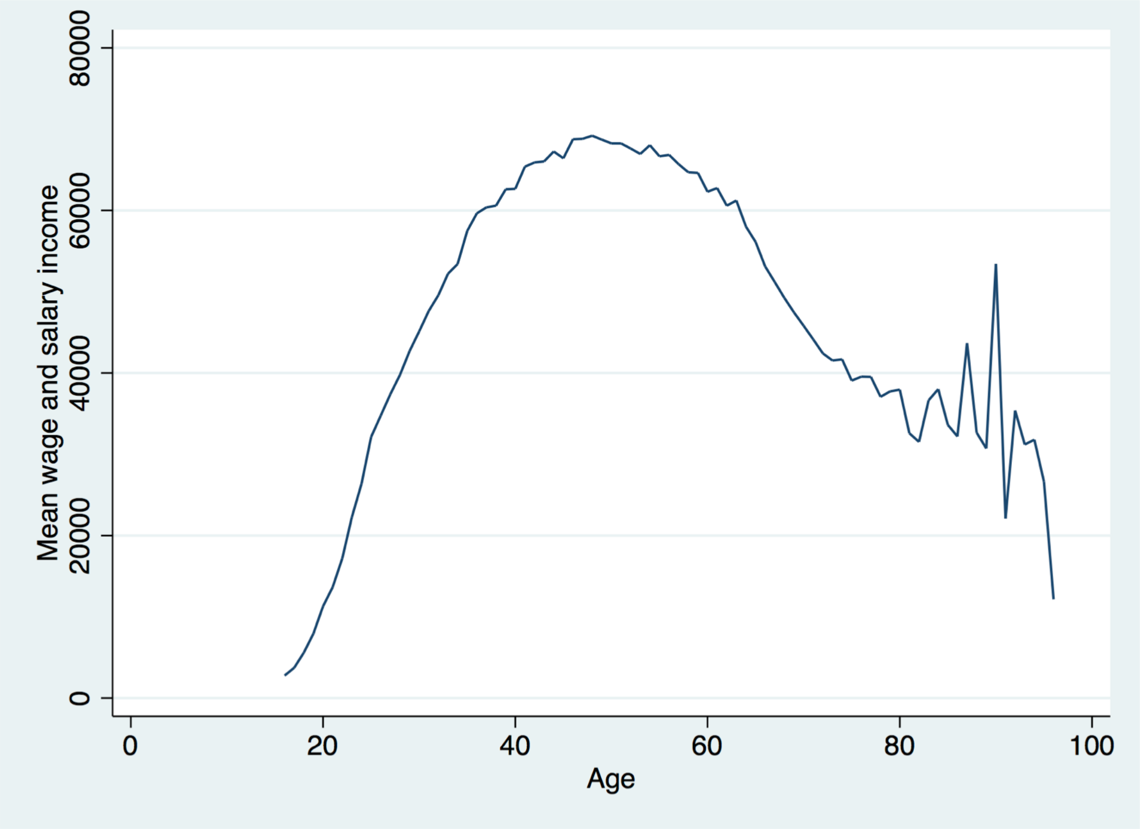
```
. svy, subpop(if income!=. & income!=0): reg income age
(running regress on estimation sample)

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.

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).

Income = F(Age, Age squared)

```
***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.

Mean 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

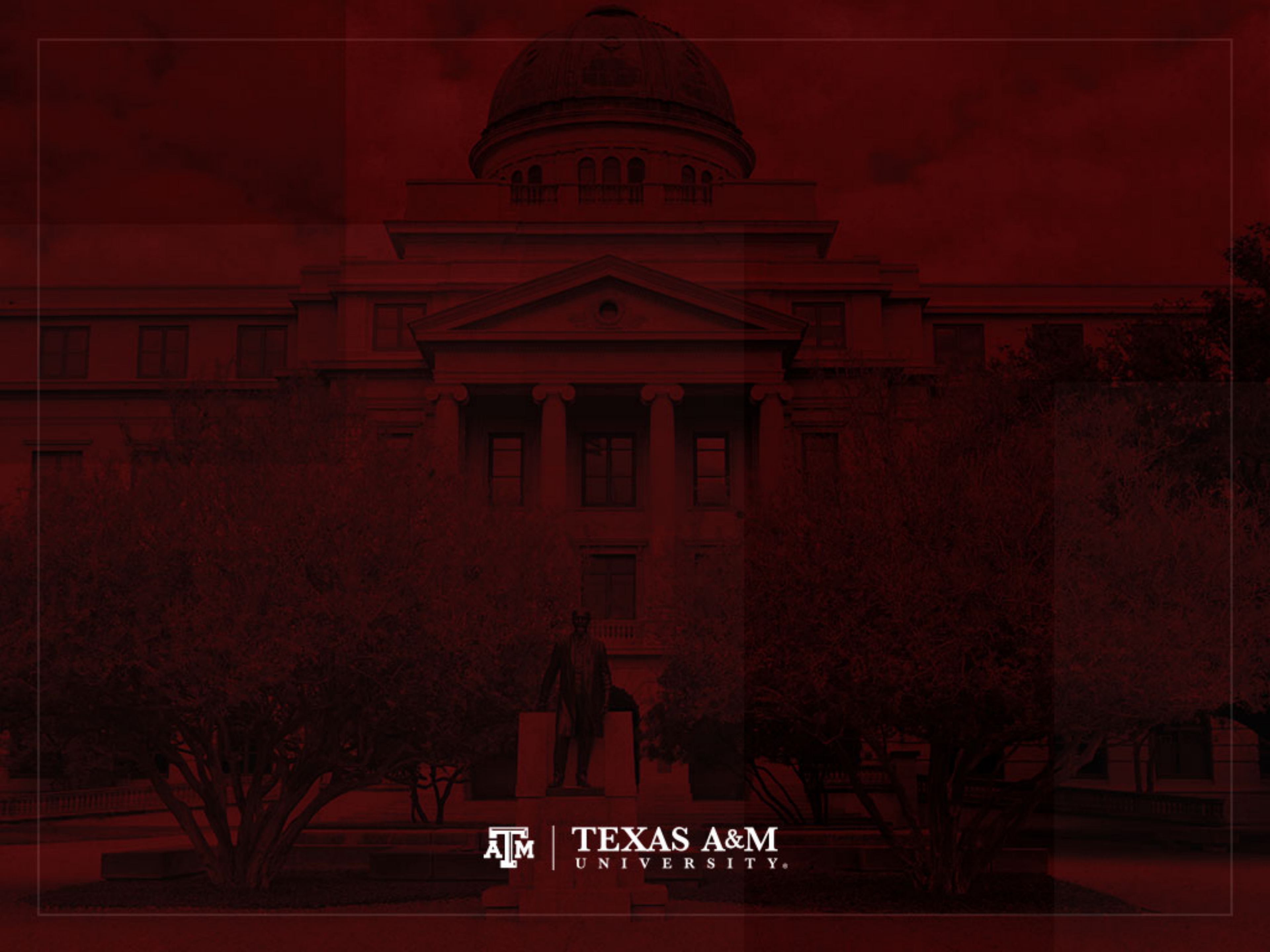
Source: 2018 American Community Survey.

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

ACS example: Pearson's *r*

```
. ***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
```

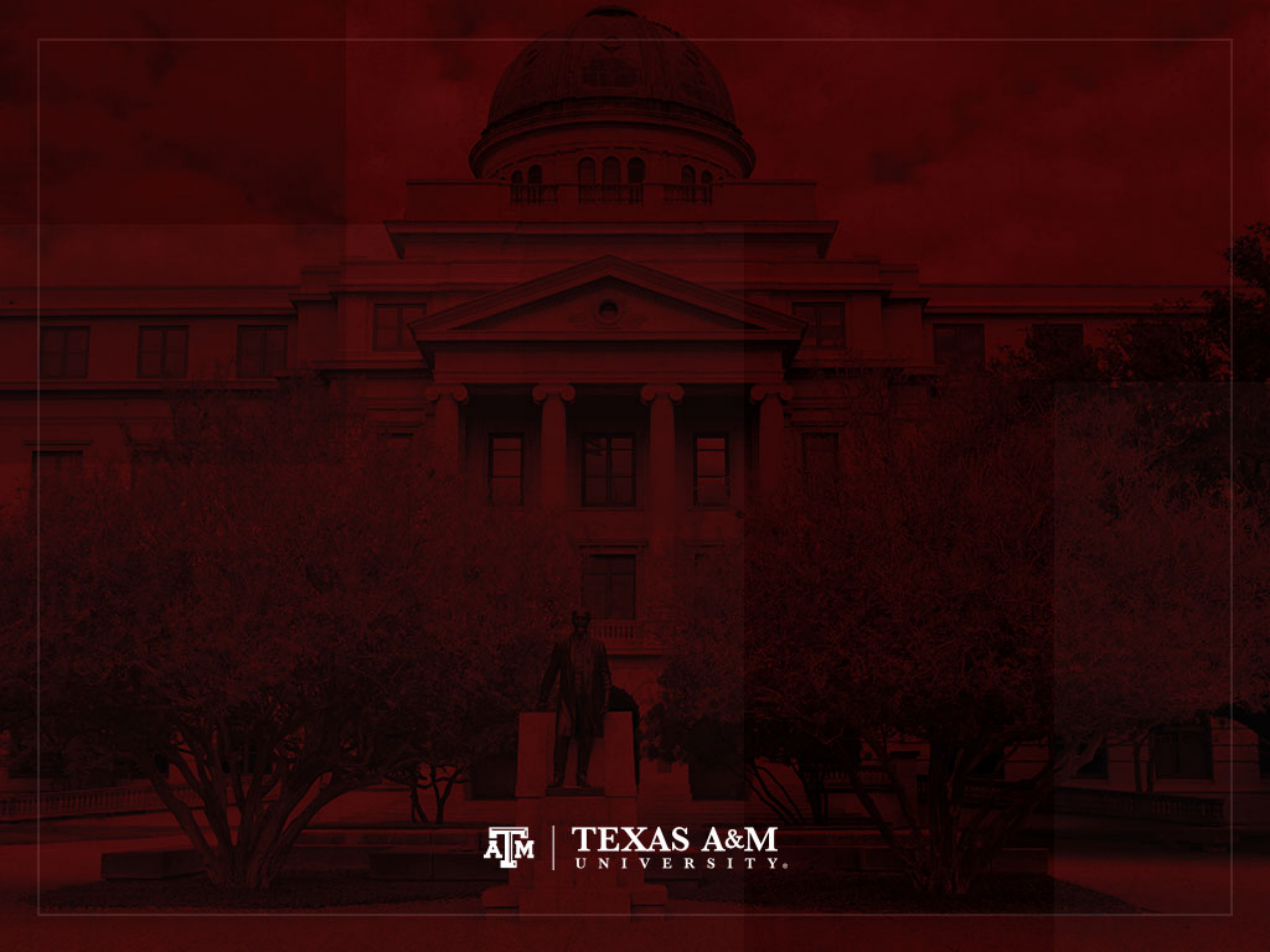
Source: 2018 American Community Survey.

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

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 example: ANOVA

- Does at least one category of the race/ethnicity variable have mean income different than the others?
  - Not good example for ANOVA, because race/ethnicity variable does not have equal numbers of cases across its categories

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

raceth	mean(income)	sd(income)	N(income)
White	55289.18	67964.86	1079026
African American	37183.63	41141.3	138,827
Hispanic	36236.16	40343.66	218,441
Asian	64154.23	75930.09	93,409
Native American	34851.55	38132.45	11,393
Ohter races	44162.79	56520.07	33,217

```
.
. ***Total number of cases
. count if raceth!=0 & income!=. & income!=0
  1,574,313
```

Source: 2018 American Community Survey.

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 in 2016

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

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      1.3178e+14      5   2.6356e+13   6975.87     0.0000
 Within groups      5.9480e+151574307   3.7782e+09
------------------------------------------------------------------------
    Total           6.0798e+151574312   3.8619e+09

Bartlett's test for equal variances:  chi2(5) =  7.3e+04  Prob>chi2 = 0.000
```

Source: 2016 General Social Survey.

Edited table

Table 1. One-way analysis of variance for wage and salary income by race/ethnicity, United States, 2018

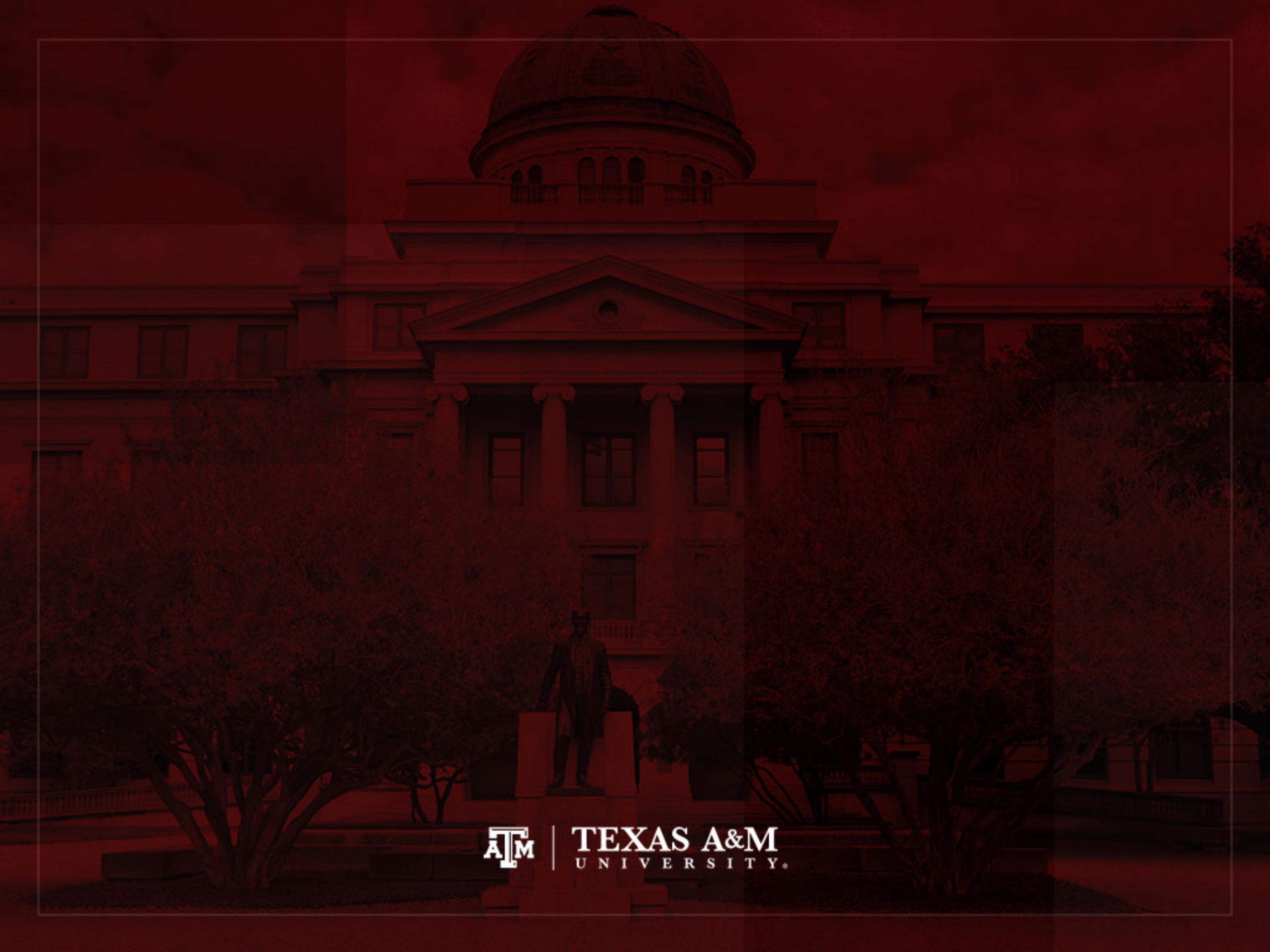
Source	Sum of Squares	Degrees of Freedom	Mean of Squares	F-test	Prob > F
Between groups	1.32e+14	5	2.64e+13	6,975.87	0.0000
Within groups	5.95e+15	1,574,307	3.78e+09		
Total	6.08e+15	1,574,312	3.86e+09		

Source: 2018 American Community Survey.

Stata practice time

- Let’s run the Stata command file

  http://www.ernestoamaral.com/docs/Stata2020a/Stata04.txt


TEXAS A&M
UNIVERSITY