Ordinary least squares regression

Ernesto F. L. Amaral

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

Outline

- **Introduction**
  - Partial slopes (*β*)
  - Standardized coefficients (*b**)
  - Statistical significance (*t*-test)
  - Multiple correlation (R^2)
  - Gauss-Markov theorem
- **Meaning of linear regression**
  - Example: Income = F(age, education)
- **Determining normality**
  - Example: ln(income) = F(age, education)
- **Predicted values**
- **Residual analysis with graphs**
  - Example: OLS with age and age squared
- **Dummy variables**
  - Example: Full OLS model
- All lecture examples: Texas (2018 American Community Survey)

Introduction

- Ordinary least squares (OLS) regression (linear regression)
  - Important technique to estimate associations of several independent variables (x_1, x_2, ..., x_k) with a dependent variable (y) at the interval-ratio level of measurement
  - Variables are at the interval-ratio level, but we can include ordinal and nominal variables as dummy variables
  - Each independent variable has a linear relationship with the dependent variable
  - Independent variables are uncorrelated with each other
  - When these and other requirements are violated (as they often are), this technique will produce biased and/or inefficient estimates

Bivariate and multivariate models

- Bivariate (simple) regression equation

$$y = a + bx = \beta_0 + \beta_1 x$$

  - $a = \beta_0 = y$ intercept (constant)
  - $b = \beta_1 =$ slope

- Multivariate (multiple) regression equation

$$y = a + b_1 x_1 + b_2 x_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

  - $b_1 = \beta_1 =$ partial slope of the linear relationship between the first independent variable (x_1) and y
  - $b_2 = \beta_1 =$ partial slope of the linear relationship between the second independent variable (x_2) and y

Multiple regression

$$y = a + b_1 y_1 + b_2 y_2 = \beta_0 + \beta_1 y_1 + \beta_2 y_2$$

- $a = \beta_0$ = the y intercept (constant), 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 (β)

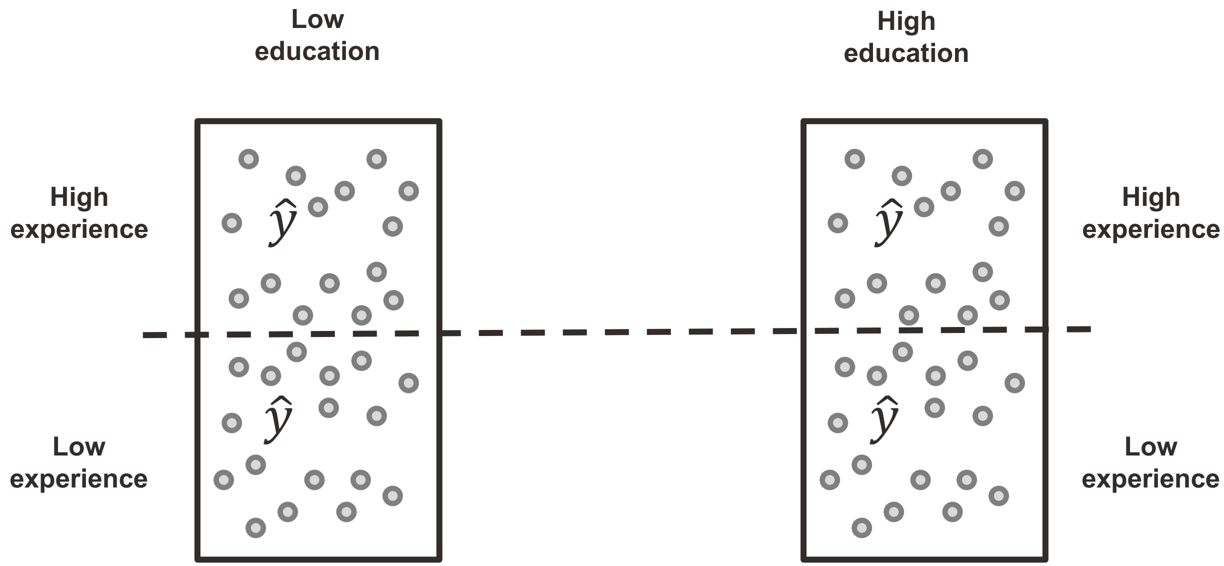
- The partial slopes (β) indicate the effect of each independent variable on y

- 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

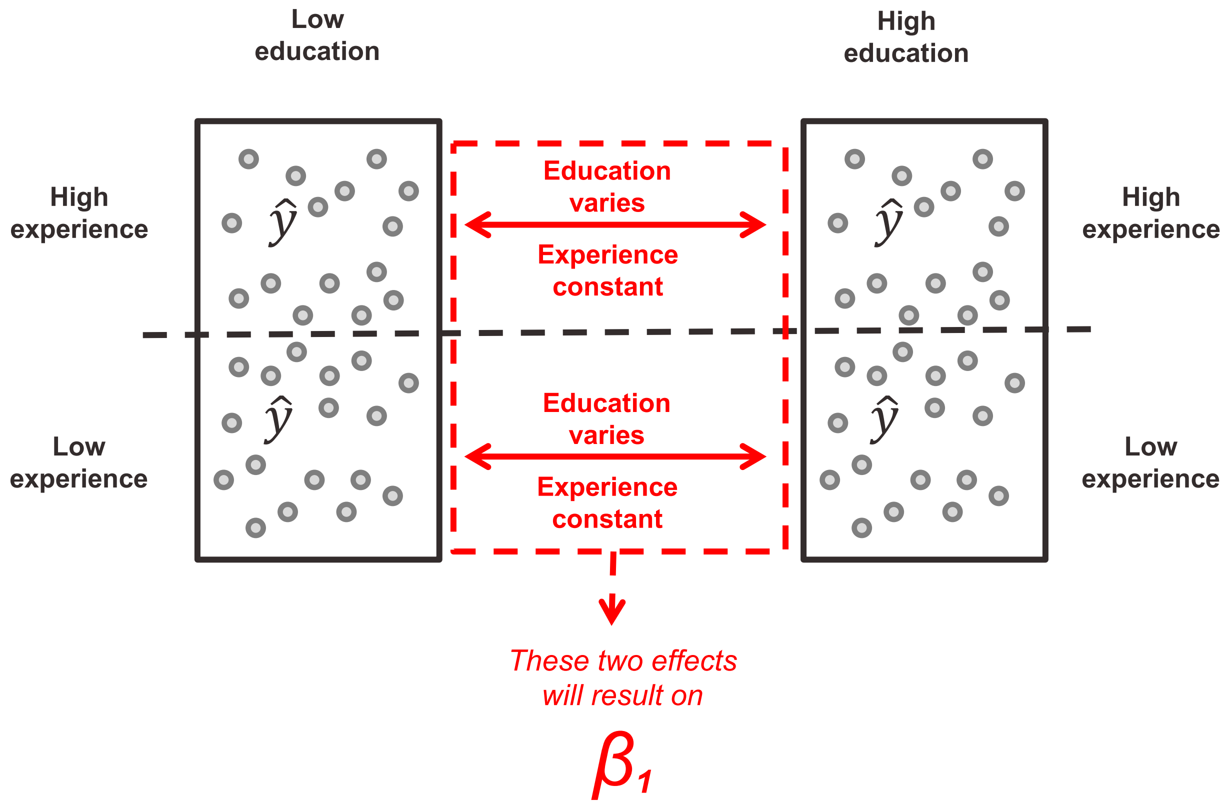
Ceteris paribus

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



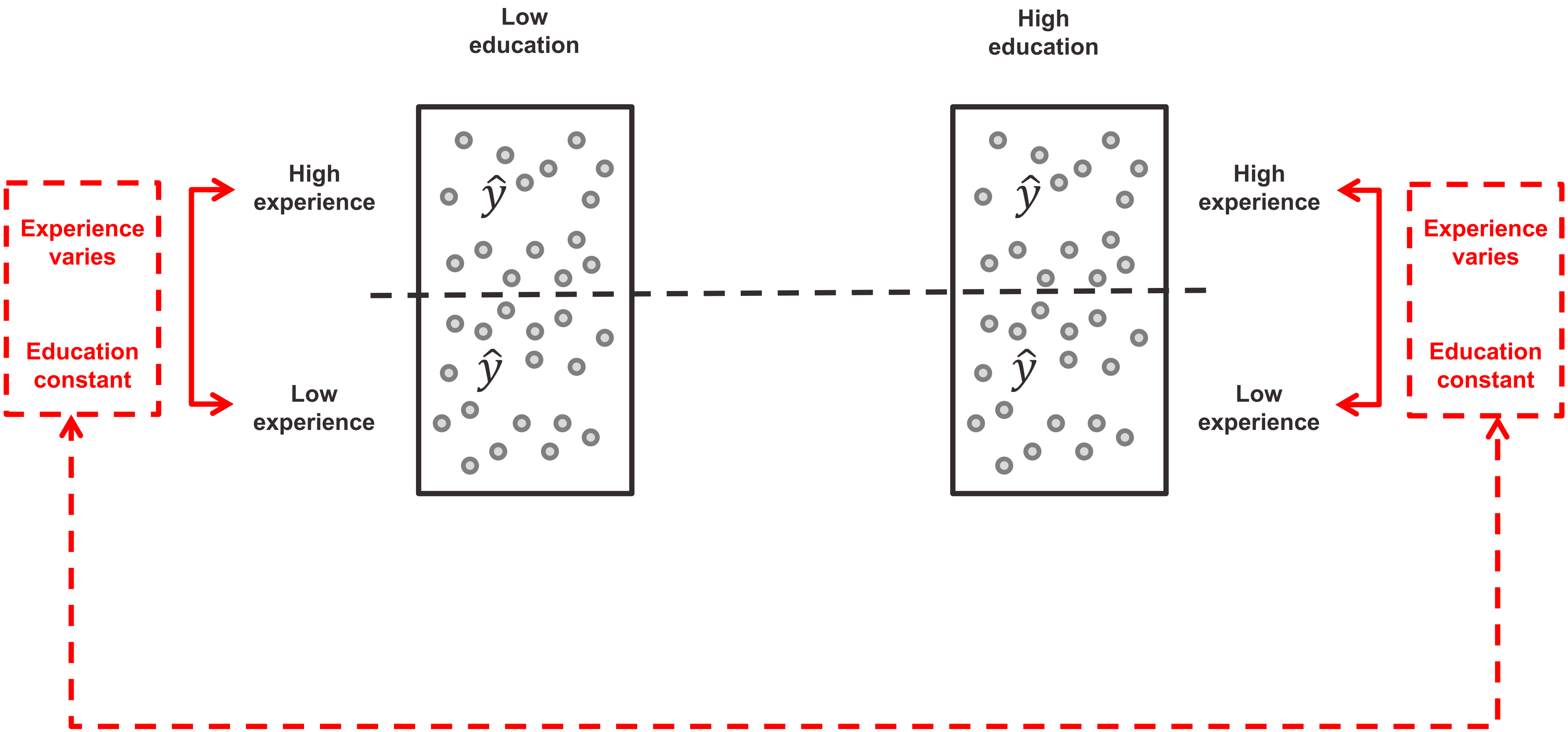
Ceteris paribus

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



Ceteris paribus

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

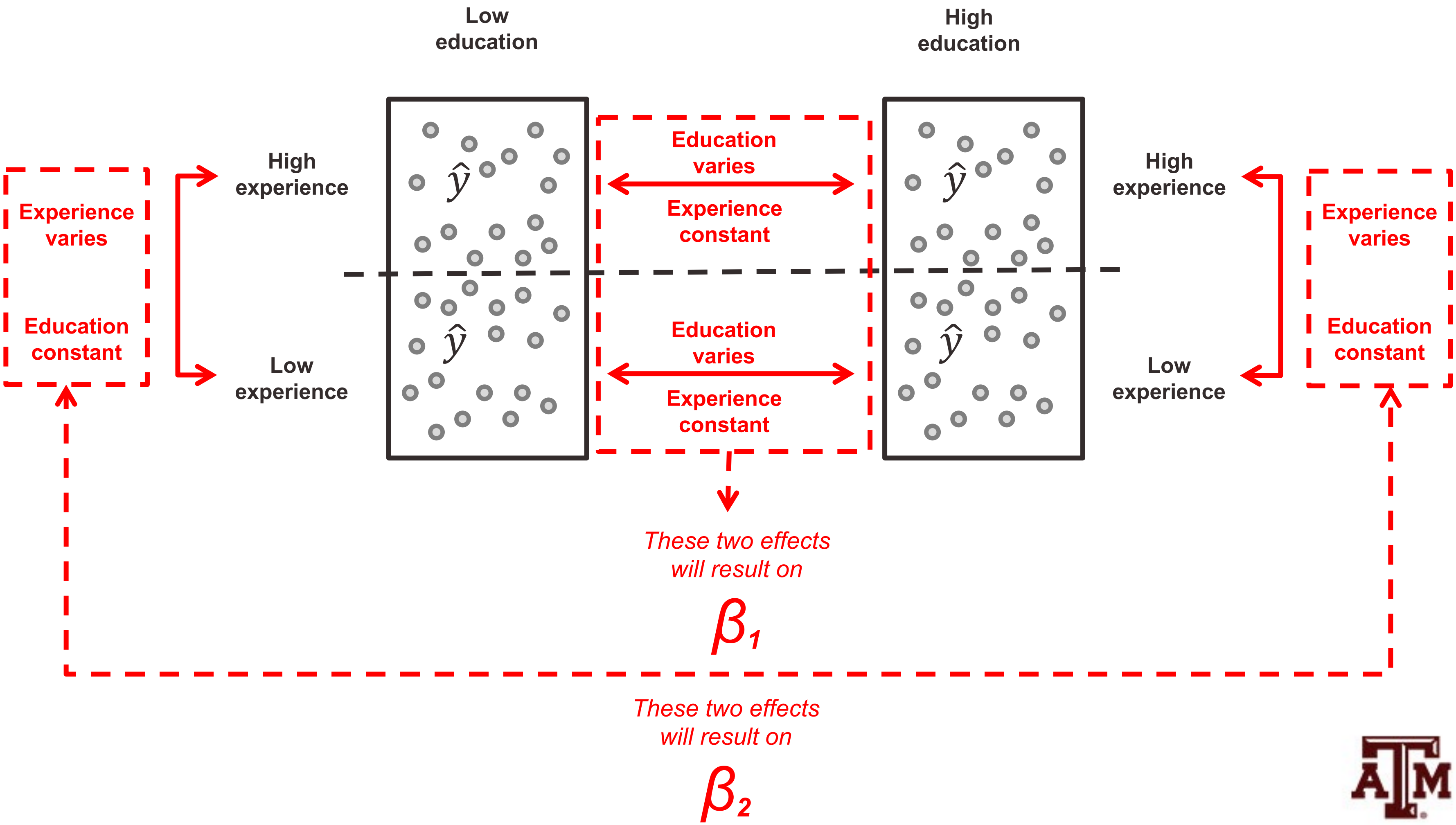


These two effects will result on

β_2

Ceteris paribus

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



Interpretation of partial slopes

- The partial slopes show the effects of the independent variables (x_1, x_2) in their original units

- These values can be used to predict scores on the dependent variable (y)

- Partial slopes must be computed before computing the y intercept (β_0)

Formulas of partial slopes

$$b_1 = \beta_1 = \left(\frac{s_y}{s_1}\right)\left(\frac{r_{y1} - r_{y2}r_{12}}{1 - r_{12}^2}\right)$$

$$b_2 = \beta_2 = \left(\frac{s_y}{s_2}\right)\left(\frac{r_{y2} - r_{y1}r_{12}}{1 - r_{12}^2}\right)$$

$b_1 = \beta_1$ = partial slope of x_1 on y

$b_2 = \beta_2$ = 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 (β_0)

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2$$

$$\beta_0 = \bar{y} - \beta_1\bar{x}_1 - \beta_2\bar{x}_2$$

Income = F(age, education)

```
. ***No weights
. reg income age educgr
```

Source	SS	df	MS
Model	8.2170e+13	2	4.1085e+13
Residual	4.5425e+14	127,782	3.5549e+09
Total	5.3642e+14	127,784	4.1979e+09

Number of obs	=	127,785
F(2, 127782)	=	11557.33
Prob > F	=	0.0000
R-squared	=	0.1532
Adj R-squared	=	0.1532
Root MSE	=	59623

income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	724.3054	11.11857	65.14	0.000	702.5132	746.0976
educgr	18177.19	140.4437	129.43	0.000	17901.92	18452.45
_cons	-32363.61	614.972	-52.63	0.000	-33568.95	-31158.28

Source: 2018 American Community Survey.

Summary of Stata weights

WEIGHTS IN FREQUENCY DISTRIBUTIONS		
Weight unit of measurement	**Expand to population size**	**Maintain sample size**
Discrete	**fweight**	**aweight**
Continuous	**iweight**	

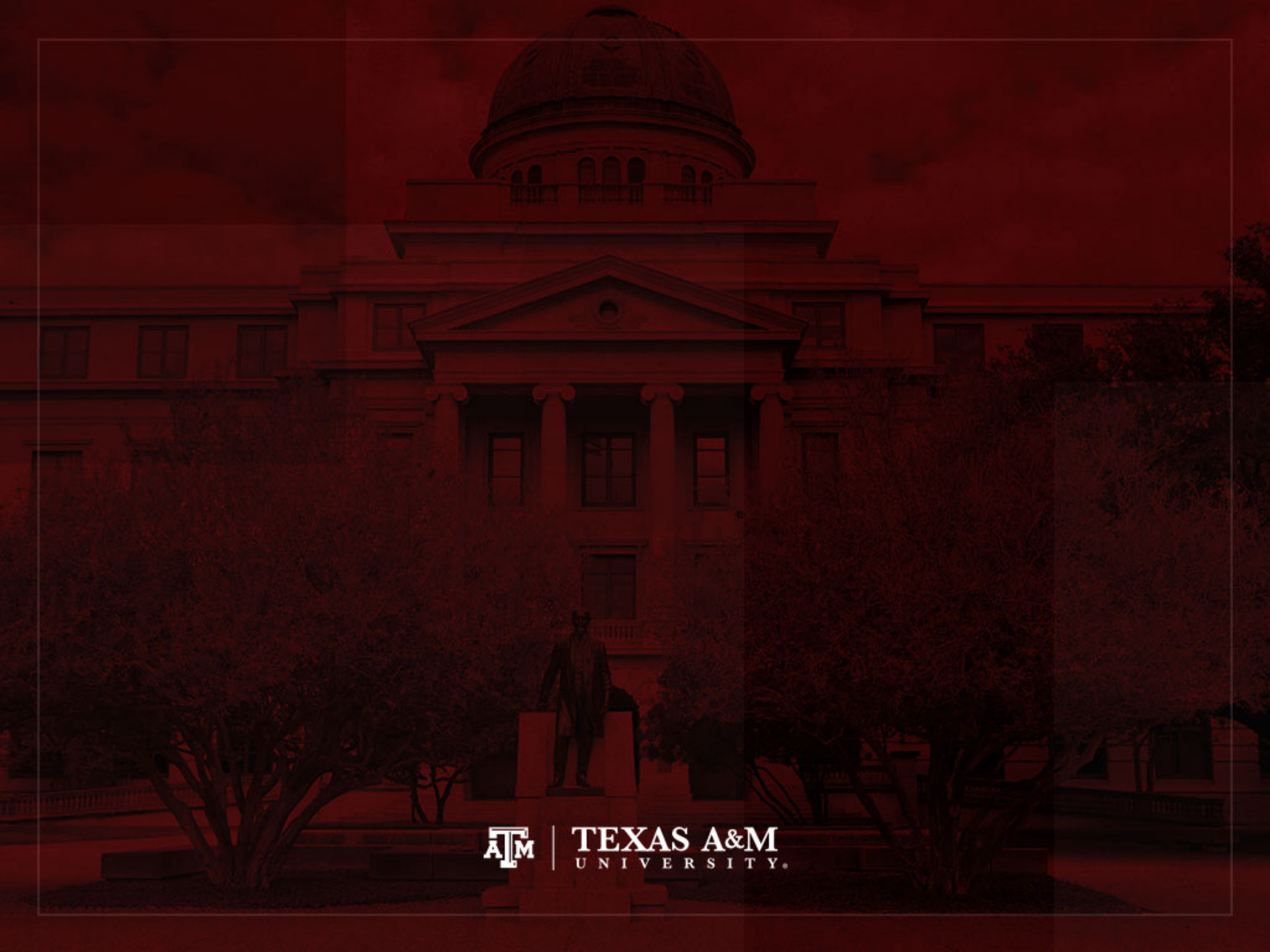
WEIGHTS IN STATISTICAL REGRESSIONS should maintain sample size	
Robust standard error	**Adjusted R^2, TSS, ESS, RSS**
pweight	**aweight**
reg y x, robust	**outreg2**

Example: Coefficients (β)

```
. ***Complex survey design
. svyset cluster [pweight=perwt], strata(strata)


. ***Use complex survey design
. svy: reg income age educgr
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        212                 Number of obs     =    127,785
Number of PSUs     =     79,499                 Population size   = 13,849,398
                                                Design df         =     79,287
                                                F(   2,   79286)  =    5751.26
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1652
```

income	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	796.3443	11.73077	67.89	0.000	773.3521	819.3366
educgr	16863.33	179.705	93.84	0.000	16511.11	17215.55
_cons	-31880.99	661.937	-48.16	0.000	-33178.38	-30583.59

Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Standardized coefficients (*b**)

- Partial slopes (b_1=β_1 ; b_2=β_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

$$b_1^* = b_1\left(\frac{s_1}{s_y}\right) = \beta_1^* = \beta_1\left(\frac{s_1}{s_y}\right)$$

$$b_2^* = b_2\left(\frac{s_2}{s_y}\right) = \beta_2^* = \beta_2\left(\frac{s_2}{s_y}\right)$$

Standardized coefficients

- Standardized regression equation

$$Z_y = a_z + b_1^* Z_1 + b_2^* Z_2$$

- *Z* indicates that all scores have been standardized to the normal curve

$$Z_i = \frac{x_i - \bar{x}}{s}$$

- The *y* intercept will always equal zero once the equation is standardized

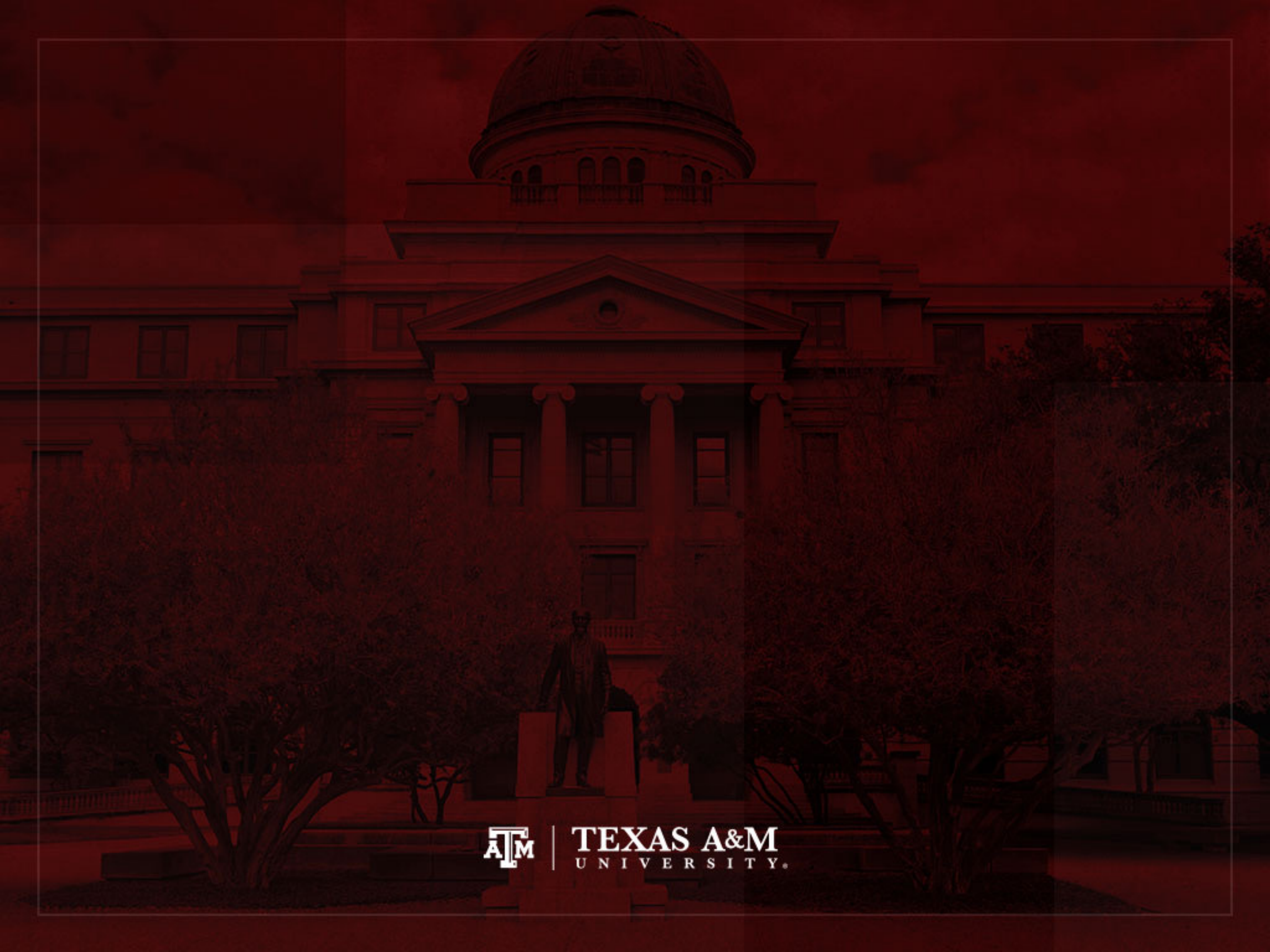
$$Z_y = b_1^* Z_1 + b_2^* Z_2$$

Example: Standardized beta (*b**)

```
. ***Standardized regression coefficients
. ***(i.e., standardized partial slopes, beta-weights)
. ***It does not allow the use of complex survey design
. ***Use pweight to maintain sample size and estimate robust standard errors
. reg income age educgr [pweight=perwt], beta
(sum of wgt is 13,849,398)

Linear regression                               Number of obs     =    127,785
                                                F(2, 127782)      =    5873.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1652
                                                Root MSE          =      54147
```

income	Coef.	Robust Std. Err.	t	P>|t|	Beta
age	796.3443	11.46129	69.48	0.000	.1943233
educgr	16863.33	177.6256	94.94	0.000	.3368842
_cons	-31880.99	649.8899	-49.06	0.000	.

Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Statistical significance (*t*-test)

- In a simple linear regression, the test of statistical significance for a *β* coefficient (*t*-test) is estimated as

$$t = \frac{\hat{\beta}}{SE_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{RSS}{df * S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \bar{x})^2}}}$$

  - SE_β: standard error of β
  - *MSE*: mean squared error = *RSS* / *df*
  - *RSS*: residual sum of squares = $\sum_i (y_i - \hat{y}_i)^2$ = $\sum_i \hat{e}_i{}^2$
  - *df*: degrees of freedom = *n*–2 for simple linear regression
    - 2 statistics (slope and intercept) are estimated to calculate sum of squares
  - S_{xx}: corrected sum of squares for x (total sum of squares)

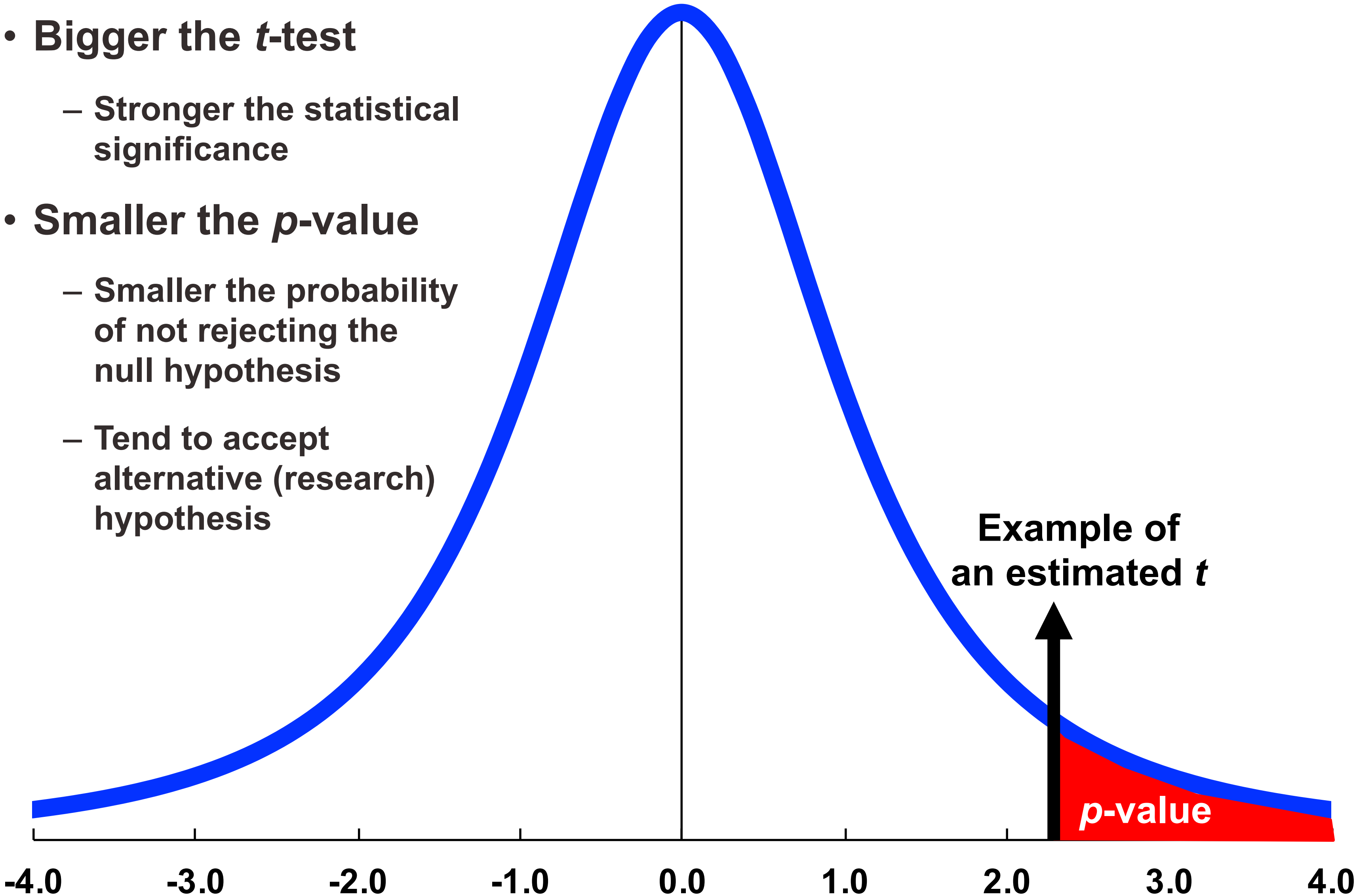
Statistical power

- Statistical power for regression analysis is the probability of finding a significant coefficient ($\hat{\beta} \neq 0$), when there is a significant relationship in the population ($\beta \neq 0$)
  - Power is dependent on the confidence level, size of coefficient (magnitude), and sample size
  - Small samples might not capture enough variation among observations
  - If we have large samples, we tend to have statistical significance (as measured by *t*-test), even for coefficients ($\hat{\beta}$) with small magnitude

$$\uparrow t = \frac{\hat{\beta}}{SE_{\hat{\beta}}} = \frac{\hat{\beta}}{\sqrt{\frac{MSE}{S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{RSS}{df * S_{xx}}}} = \frac{\hat{\beta}}{\sqrt{\frac{\sum_i (y_i - \hat{y}_i)^2}{\uparrow(n-2)\sum_i (x_i - \bar{x})^2}}}\downarrow$$

t distribution (*df* = 2)

- **Bigger the *t*-test**
  - **Stronger the statistical significance**
- **Smaller the *p*-value**
  - **Smaller the probability of not rejecting the null hypothesis**
  - **Tend to accept alternative (research) hypothesis**



Decisions about hypotheses

Hypotheses	$p < \alpha$	$p > \alpha$
Null hypothesis (H_0)	Reject	Do not reject
Alternative hypothesis (H_1)	Accept	Do not accept

- ***p*-value** is the probability of not rejecting the null hypothesis
- If a statistical software gives only the two-tailed *p*-value, divide it by 2 to obtain the one-tailed *p*-value

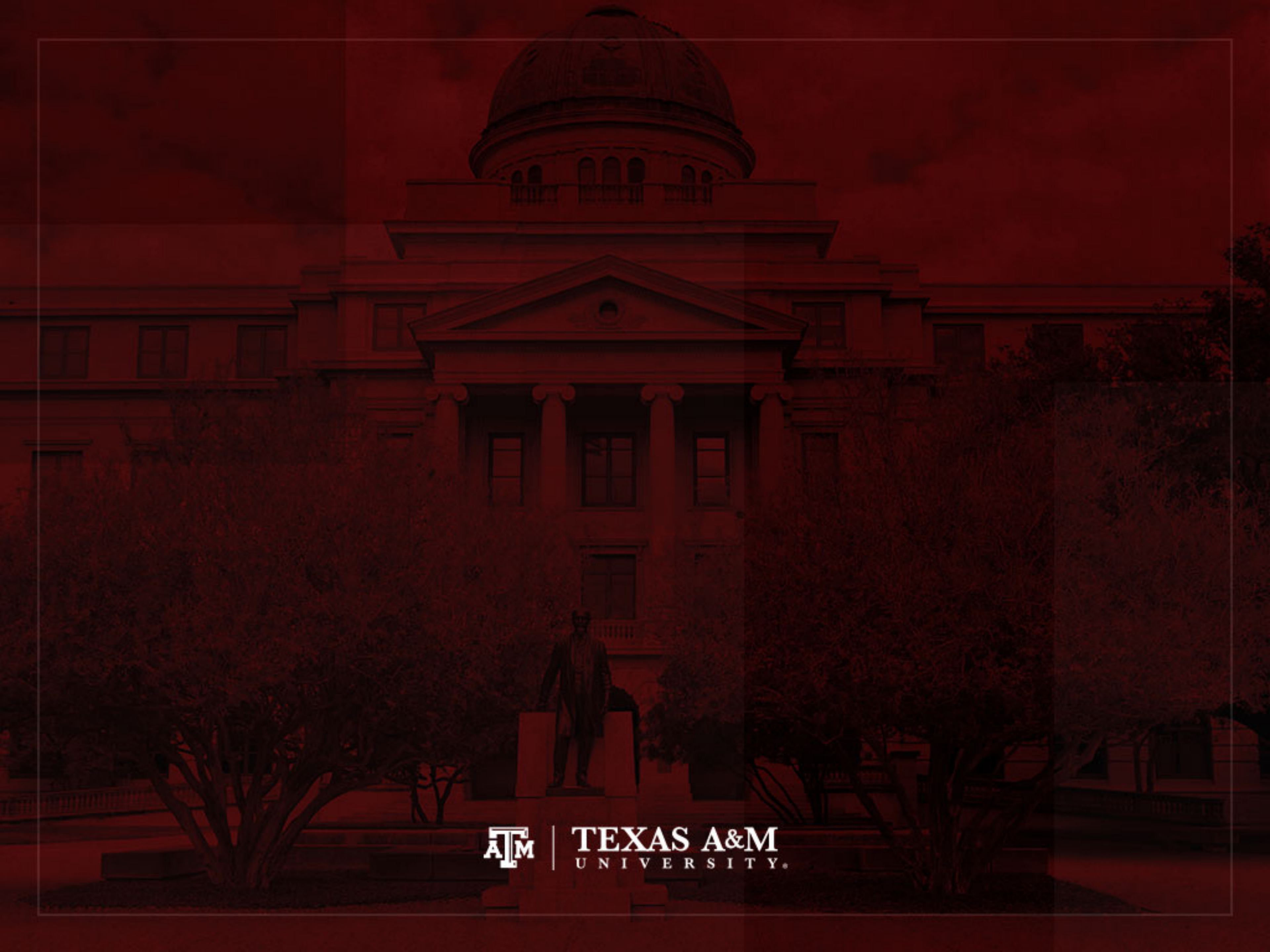
Significance level (α)	Confidence level (success rate)
0.10 (10%)	90%
0.05 (5%)	95%
0.01 (1%)	99%
0.001 (0.1%)	99.9%

Example: Statistical significance

```
. ***Use complex survey design
. svy: reg income age educgr
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        212                Number of obs      =    127,785
Number of PSUs     =     79,499                Population size    = 13,849,398
                                               Design df          =     79,287
                                               F(    2,   79286)  =    5751.26
                                               Prob > F           =     0.0000
                                               R-squared          =     0.1652
```

income	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	796.3443	11.73077	67.89	0.000	773.3521	819.3366
educgr	16863.33	179.705	93.84	0.000	16511.11	17215.55
_cons	-31880.99	661.937	-48.16	0.000	-33178.38	-30583.59

Source: 2018 American Community Survey.



TEXAS A&M
UNIVERSITY

Multiple correlation (R^2)

- The coefficient of multiple determination (R^2) measures how much of the dependent variable (y) is explained by all independent variables (x_1, x_2, x_3, ..., x_k) combined
- R^2 is an estimation of the percentage of the variation in y that is explained by variations in all independent variables in the population
- The coefficient of multiple determination is an indicator of the strength of the entire regression equation

R^2 estimation

- For a regression with two independent variables, this is the equation to estimate R^2

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

  - R^2 = coefficient of multiple determination
  - r_{y1}^2 = coefficient of determination for y and x_1 (or amount of variation in y explained by x_1)
  - $r_{y2.1}^2$ = partial correlation of y and x_2, while controlling for x_1 (or amount of variation in y explained by x_2, after x_1 is controlled)
  - $\left(1 - r_{y1}^2\right)$ = amount of variation remaining in y, after 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, while controlling for x_1 ($r_{y2.1}$)

$$r_{y2.1} = \frac{r_{y2} - (r_{y1})(r_{12})}{\sqrt{1 - r_{y1}^2}\sqrt{1 - r_{12}^2}}$$

- We need three correlations
  - Bivariate correlation between *y* and x_1 (r_{y1})
  - Bivariate correlation between *y* and x_2 (r_{y2})
  - Bivariate correlation between x_1 and x_2 (r_{12})

Explaining R^2 estimation

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

- If the partial correlation of y and x_2, while controlling for x_1 ($r_{y2.1}$), is not equal to zero
  - R^2 will necessarily increase by adding x_2
  - Any variable x will have a non-zero correlation with y
  - In real databases, y and any x don't have correlation exactly equal to zero
- Thus, more independent variables (even if not related to theory) will generate higher R^2

R^2 and independent variables

- Selection of independent variables based on R^2 size might generate unreasonable models
- There is nothing in the hypotheses of linear models that require a minimum value for R^2
- Models with small R^2 might mean that we didn't include important independent variables
  - It doesn't mean necessarily that non-observed factors (residuals) are correlated with independent variables
- R^2 size doesn't have influence on the mean of residuals being equal to zero

R^2 in terms of variance

- R^2 can also be written in terms of variance of y in the population (σ_y^2) and variance of error term (residual u) in the population (σ_u^2)

$$R^2 = 1 - \sigma_u^2 / \sigma_y^2$$

- R^2 is the proportion of variation in y explained by all independent variables

$$R^2 = \text{ESS} / \text{TSS}$$

$$R^2 = 1 - \text{RSS} / \text{TSS}$$

$$R^2 = 1 - (\text{RSS}/n) / (\text{TSS}/n)$$

  - Explained sum of squares (ESS), model sum of squares
  - Residual sum of squares (RSS)
  - Total sum of squares (TSS)

Adjusted R^2

- We can replace RSS/*n* and TSS/*n* by non-biased terms for σ_u^2 and σ_y^2

  Adjusted $R^2 = 1 - [RSS/(n-k-1)] / [TSS/(n-1)]$

  - Adjusted R^2 doesn't correct for possible bias of R^2 estimating the true population R^2
  - But it penalizes for the inclusion of redundant independent variables
  - *k* is the number of independent variables
  - Negative adjusted R^2 indicates a poor overall fit

$$\downarrow Adjusted\ R^2 = 1 - \frac{\frac{1-R^2}{n-1}}{\downarrow n - \uparrow k - 1} \uparrow$$

Comparing models

- We can compare adjusted R^2 of models with different forms of independent variables

$$y = \beta_0 + \beta_1 \log(x) + u$$

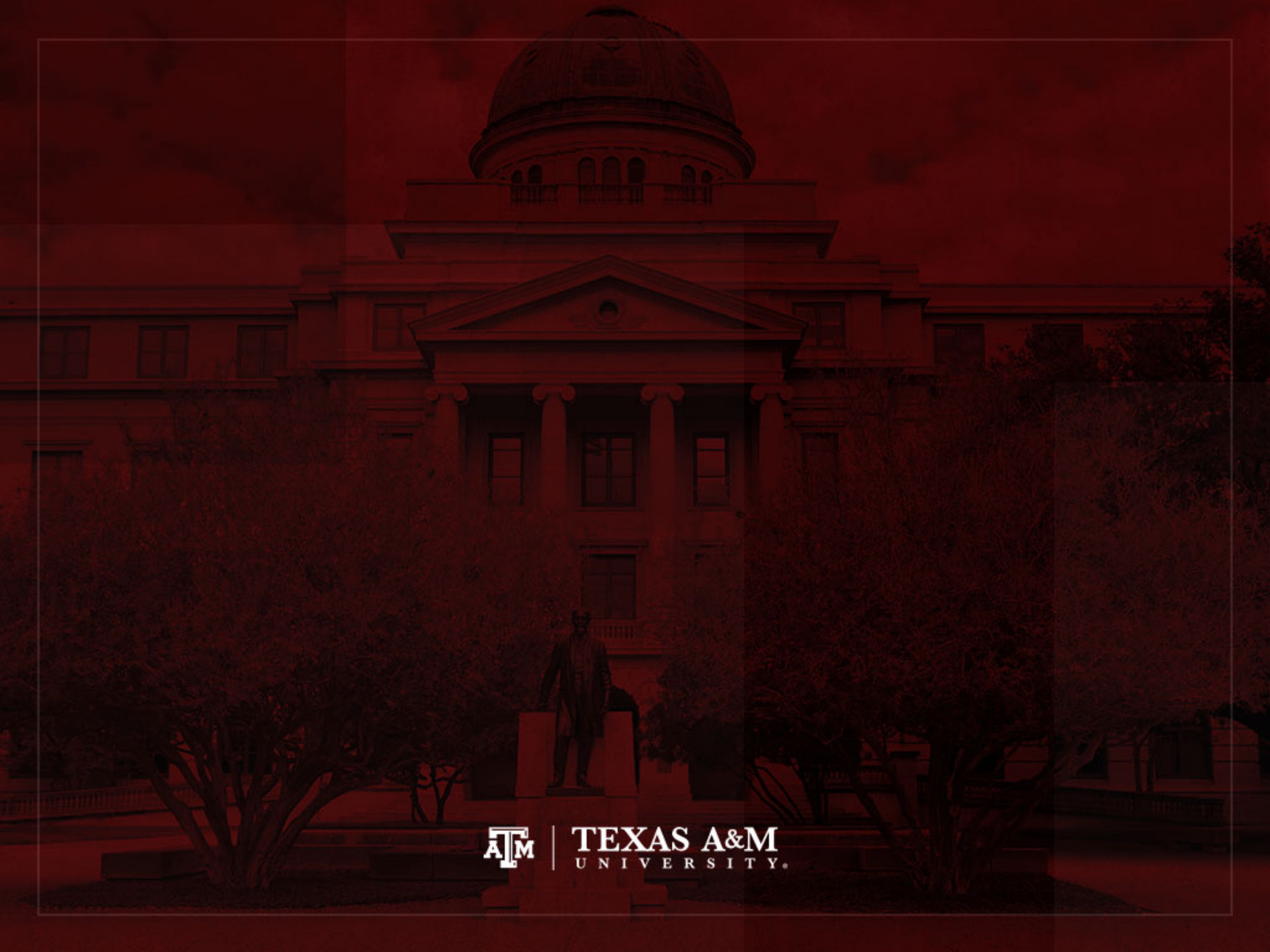
$$y = \beta_0 + \beta_1 \mathrm{x} + \beta_2 x^2 + u$$

  – We cannot use R^2 or adjusted R^2 to choose between different forms of dependent variable

  – Different forms of *y* have different amounts of variation to be explained

Example: R^2, Adjusted R^2

```
. ***Use aweight to estimate adjusted R-squared
. ***pweight and complex survey design omit sum of squares and adjusted R-squared
. reg income age educgr [aweight=perwt]
(sum of wgt is 13,849,398)

      Source |       SS           df       MS      Number of obs   =   127,785
-------------+----------------------------------   F(2, 127782)    =  12641.17
       Model |  7.4126e+13         2  3.7063e+13   Prob > F        =    0.0000
    Residual |  3.7465e+14   127,782  2.9319e+09   R-squared       =    0.1652
-------------+----------------------------------   Adj R-squared   =    0.1652
       Total |  4.4877e+14   127,784  3.5120e+09   Root MSE        =     54147

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   796.3443   10.53436    75.59   0.000     775.6972    816.9915
      educgr |   16863.33   128.6752   131.05   0.000     16611.13    17115.53
       _cons |  -31880.99   554.2213   -57.52   0.000    -32967.25   -30794.72
------------------------------------------------------------------------------
```


TEXAS A&M
UNIVERSITY

Gauss-Markov theorem

- The Gauss-Markov theorem states that if the linear regression model satisfies classical assumptions
  - Then ordinary least squares (OLS) regression produces unbiased estimates that have the smallest variance of all possible linear estimators
  - *Best Linear Unbiased Estimators (BLUEs)*

Linear in parameters

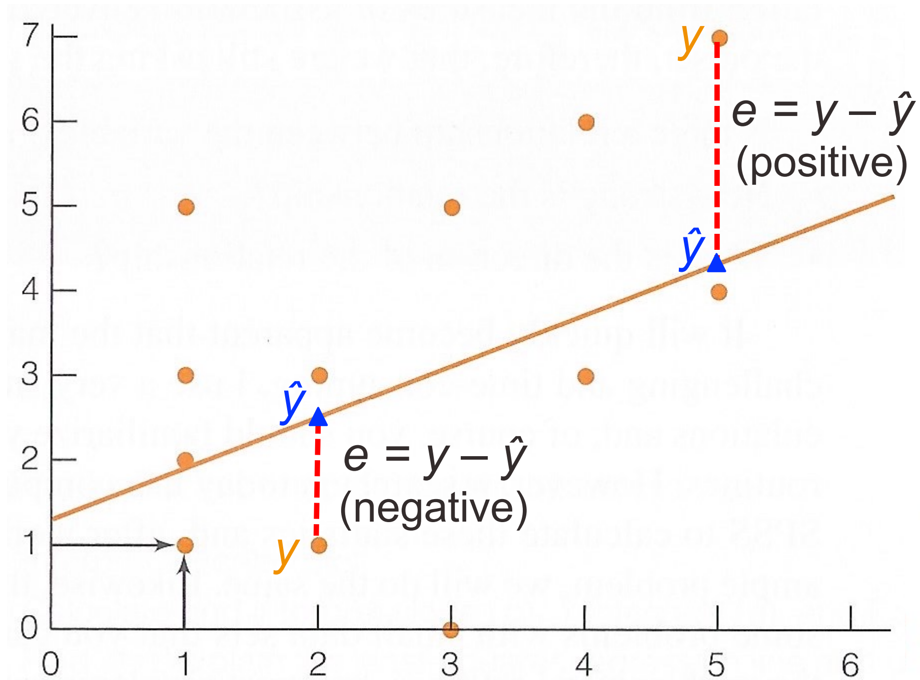
- The regression model is linear in the coefficients and the error term
  - All terms in the model are either the constant or a parameter multiplied by an independent variable
  - The population model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots + \beta_k x_k + e$$

  - β_0, β_1,..., β_k represent unknown parameters
  - Error term is known as the residual (e, ϵ, or u)
    - It is an unobserved random error
    - It is the variation in *y* that the model doesn't explain
- We should have a random sample of *n* observations for the population model

Conditional mean equals zero

- The error term has as population mean of zero
  - The expected value (mean) of the unobserved random error (*e*) is zero, given any values of the independent variables
  - $E(e|x_1,x_2,\ldots,x_k) = 0$
- Residuals = $e = y_i - \hat{y}_i$
  - Observed minus fitted
  - Observed minus predicted
  - Sum of residuals (population mean) should be zero



Source: Healey 2015, p.344.

All x are uncorrelated with e

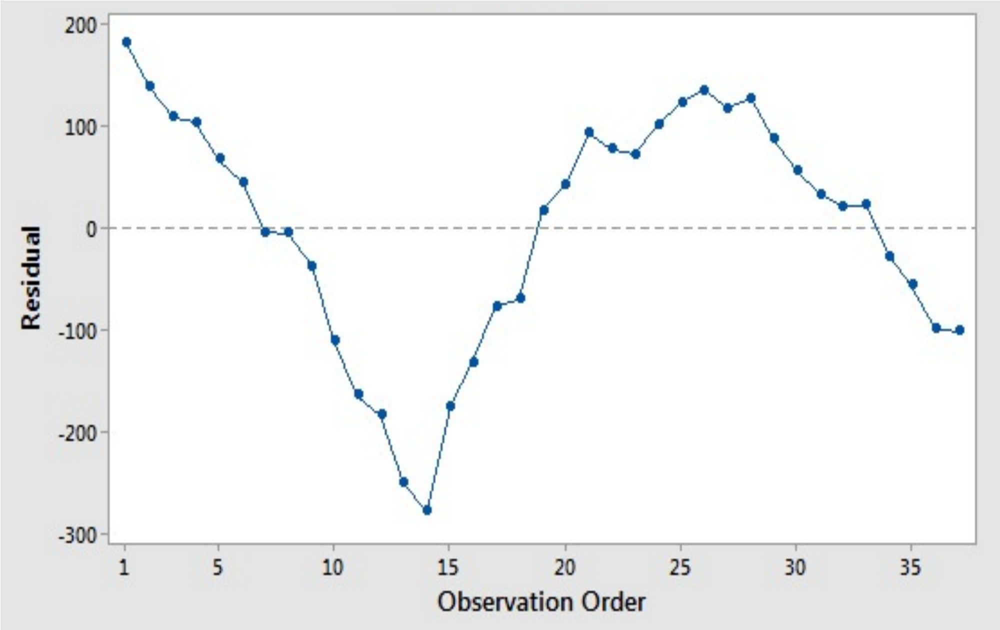
- All independent variables (*x*) are uncorrelated with the error term (*e*)
  - If an independent variable is correlated with the error term, the independent variable can be used to predict the error term
  - This violates the notion that the error term represents unpredictable random error
- This assumption is referred to as exogeneity
  - When this type of correlation exists, there is endogeneity
  - There is reverse causality between independent and dependent variables, omitted variable bias, or measurement error

Source: https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/

Uncorrelated observations of *e*

- Observations of the error term (*e*) are uncorrelated with each other
  - One observation of the error term should not predict the next observation
- Verify by graphing the residuals in the order that the data was collected
  - We want to see randomness in the plot

Example: observations of *e* are correlated



Source: https://statisticsbyjim.com/regression/ols-linear-regression-assumptions/

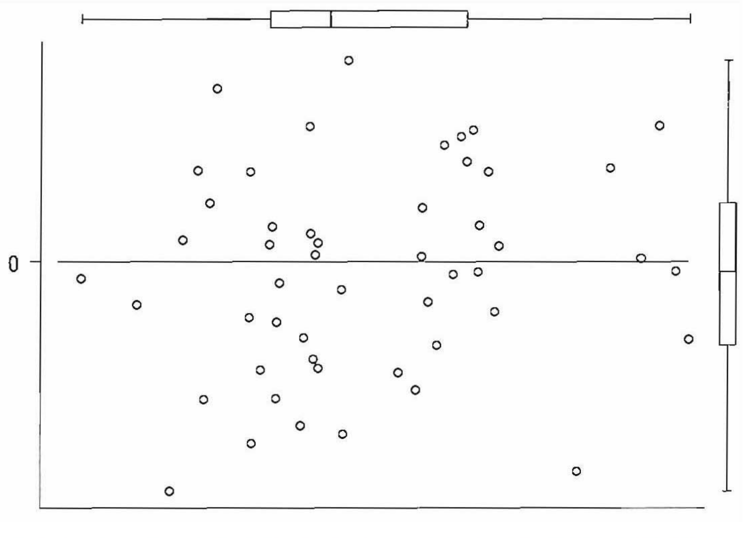
No perfect collinearity

- No independent variable is a perfect linear function of other independent variables
  - No independent variable is constant and there are no exact linear relations among independent variables
- Independent variables should be associated among themselves, but there should be **no perfect collinearity**
  - e.g., one variable should not be the multiple of another one
- High levels of correlation among independent variables and small sample size increase standard errors of β
  - This decreases statistical significance: $t = \beta / SE_{\beta}$
- High correlation (but not perfect) among independent variables is not desirable (**multicollinearity**)

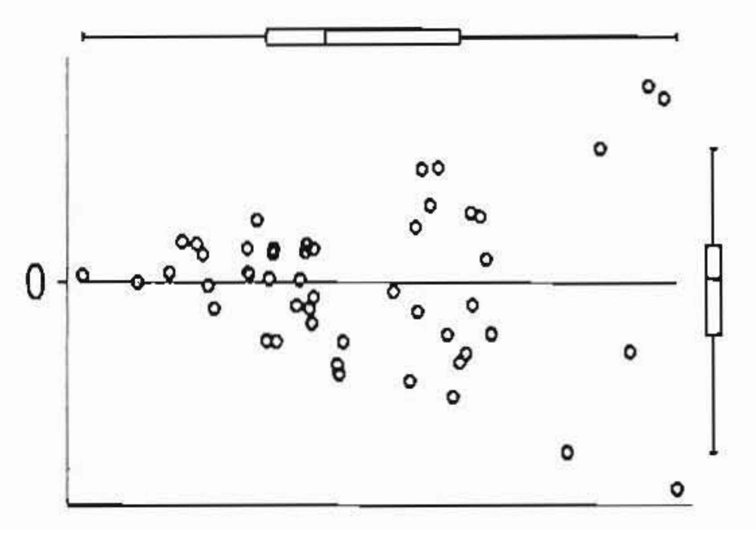
Homoscedasticity

- The error term has a constant variance (no heteroscedasticity)
  - Variance of errors (*e*) should be consistent for all observations
  - Variance does not change for each observation or range of observations
  - If this assumption is violated, the model has heteroscedasticity
- Optional: Error terms should be normally distributed

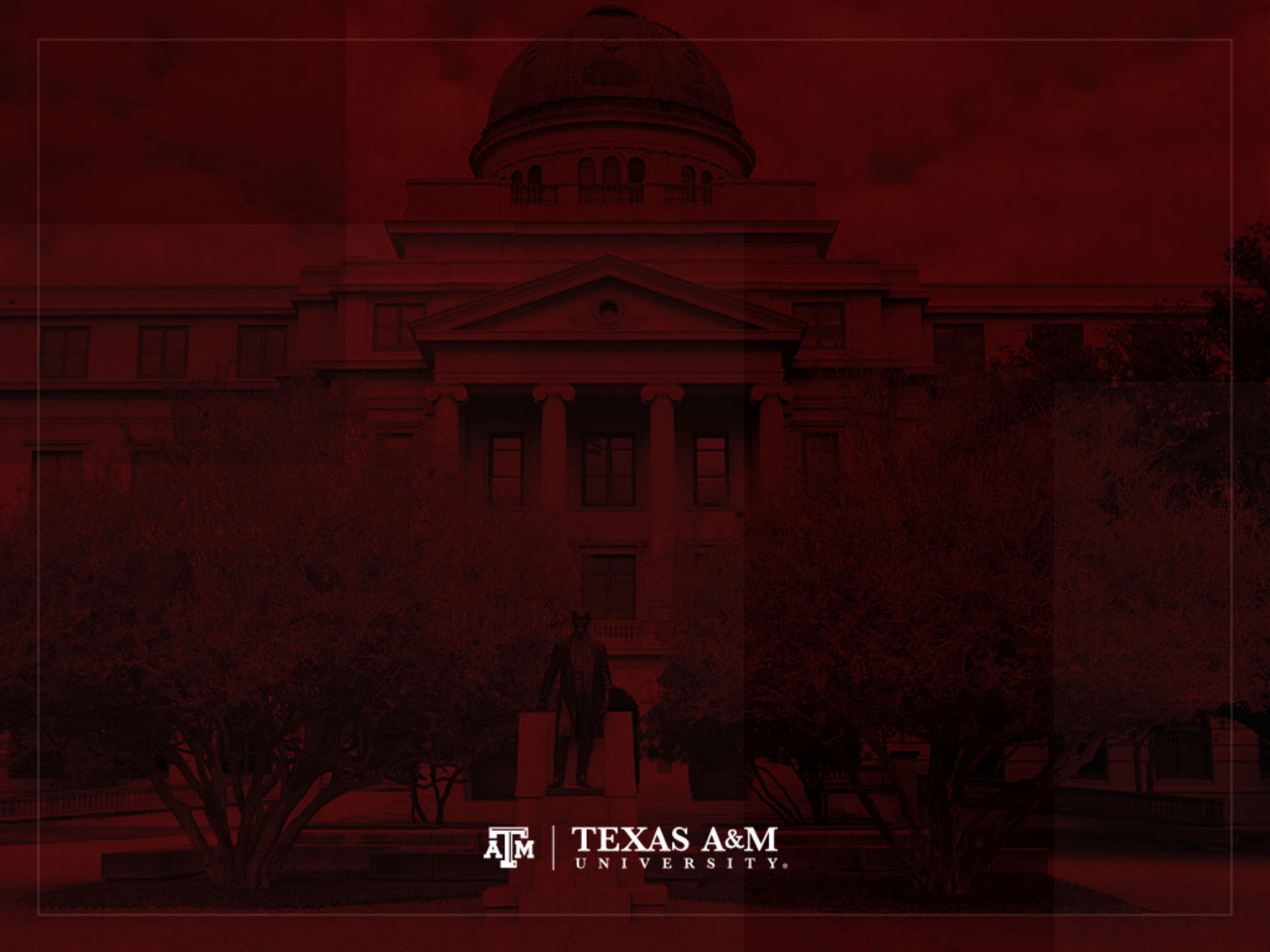
Homoscedasticity



Heteroscedasticity



Source: Hamilton, 1992, p.52–53.


TEXAS A&M
UNIVERSITY

Meaning of linear regression

- Ordinary least squares regression is commonly named linear regression
  - But it allows us to include non-linear associations
  - The model is linear in the parameters: β_0, β_1...
- There are no restrictions of how *y* and *x* are associated with the original dependent and independent variables
  - We can use natural logarithm, squared values, squared root, dummy independent variables...
  - The **interpretation** of coefficients depends of how *y* and *x* are estimated and included in the regression

Interpretation of coefficients

- An increase of one unit in x increases y by β_1 units

$$y = \beta_0 + \beta_1 x + e$$

- An increase of 1% in x increases y by (β_1/100) units

$$y = \beta_0 + \beta_1 log(x) + e$$

- An increase of one unit in x increases y by (100*β_1)%
  - Exact percentual change with semi-elasticity {[exp(β_1) – 1]*100}

$$log(y) = \beta_0 + \beta_1 x + e$$

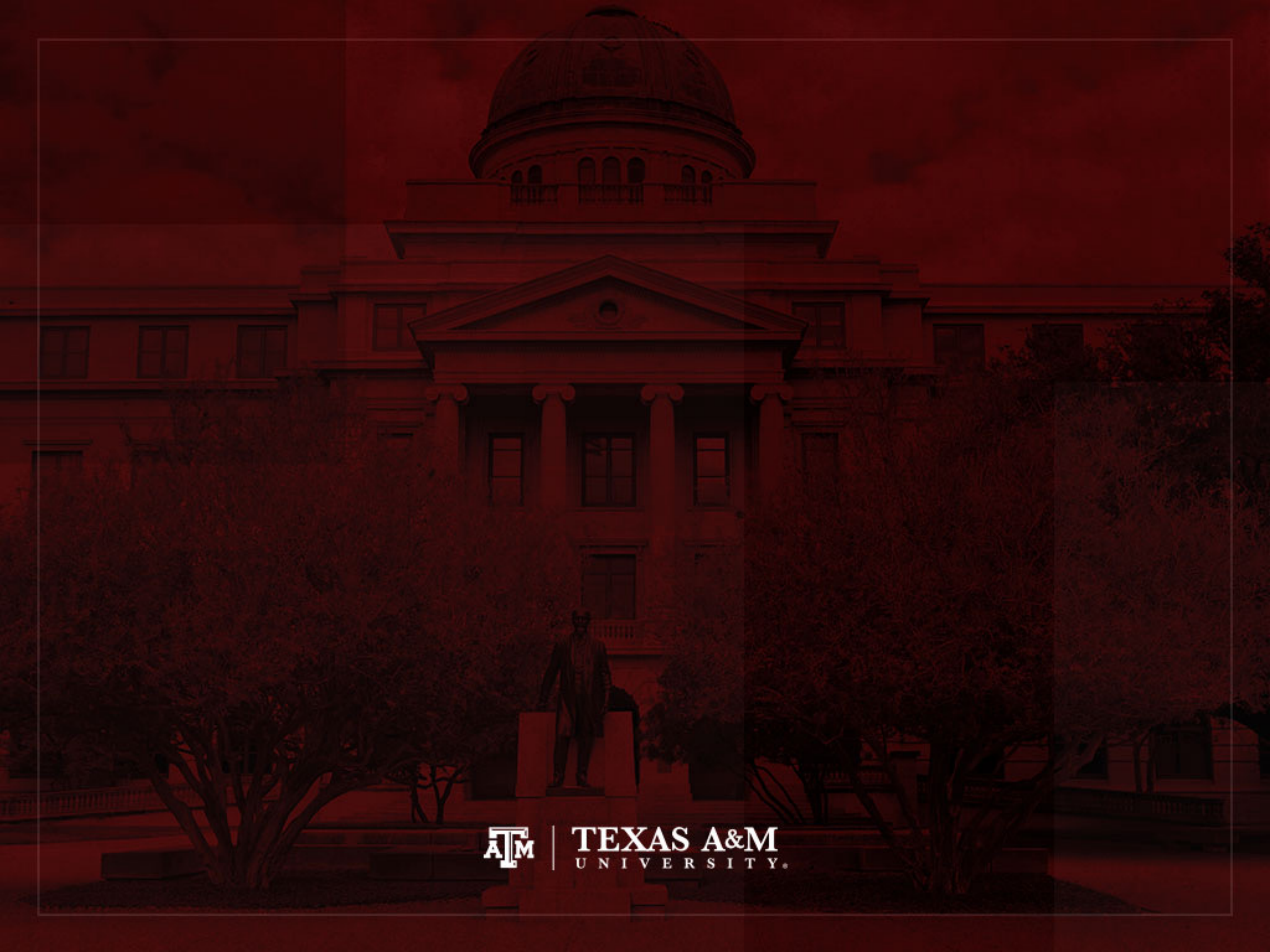
- An increase of 1% in x increases y by β_1%
  - Constant elasticity model
  - Elasticity is the ratio of the percentage change in y to the percentage change in x

$$log(y) = \beta_0 + \beta_1 log(x) + e$$

Logarithm functional forms

Model	Dependent variable	Independent variable	Interpretation of $\boldsymbol{\beta_1}$
linear	y	x	$\Delta y=\beta_1\Delta x$
linear-log	y	log(x)	$\Delta y=(\beta_1/100)\%\Delta x$
log-linear (semi-log)	log(y)	x	$\%\Delta y=(100\beta_1)\Delta x$
log-log	log(y)	log(x)	$\%\Delta y=\beta_1\%\Delta x$

Source: Wooldridge, 2008.


TEXAS A&M
UNIVERSITY

Income = F(age, education)

```
. ***Use complex survey design
. svy: reg income age educgr
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        212          Number of obs      =    127,785
Number of PSUs     =     79,499          Population size    = 13,849,398
                                         Design df          =     79,287
                                         F(    2,  79286)   =    5751.26
                                         Prob > F           =     0.0000
                                         R-squared          =     0.1652
```

income	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	796.3443	11.73077	67.89	0.000	773.3521	819.3366
educgr	16863.33	179.705	93.84	0.000	16511.11	17215.55
_cons	-31880.99	661.937	-48.16	0.000	-33178.38	-30583.59

Source: 2018 American Community Survey.

Interpretation of coefficients

(income with continuous independent variables)

- Coefficient for **age** equals 796.34
  - When age increases by one unit, income increases on average by **796.34 dollars**, controlling for education
- Coefficient for **education** equals 16,863.33
  - When education increases by one unit, income increases on average by **16,863.33 dollars**, controlling for age

Standardized coefficients

```
. ***Standardized regression coefficients
. ***(i.e., standardized partial slopes, beta-weights)
. ***It does not allow the use of complex survey design
. ***Use pweight to maintain sample size and estimate robust standard errors
. reg income age educgr [pweight=perwt], beta
(sum of wgt is 13,849,398)

Linear regression                               Number of obs     =    127,785
                                                F(2, 127782)      =    5873.56
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1652
                                                Root MSE          =      54147
```

income	Coef.	Robust Std. Err.	t	P>|t|	Beta
age	796.3443	11.46129	69.48	0.000	.1943233
educgr	16863.33	177.6256	94.94	0.000	.3368842
_cons	-31880.99	649.8899	-49.06	0.000	.

Source: 2018 American Community Survey.

Interpretation of standardized
(income with continuous independent variables)

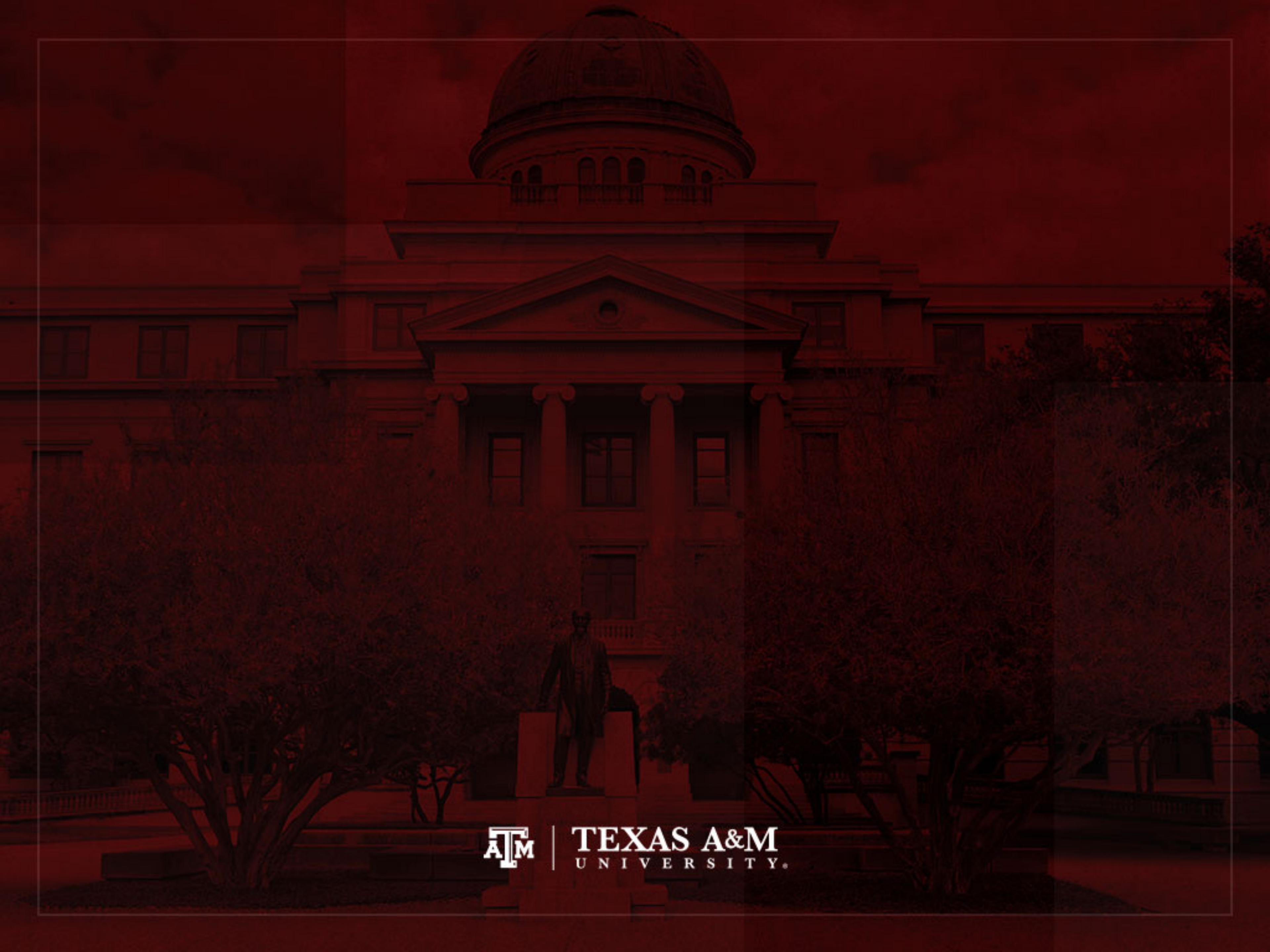
- Coefficient for **age** equals 0.1943
  - When age increases by one standard deviation, income increases on average by **0.1943 standard deviations**, controlling for education

- Coefficient for **education** equals 0.3369
  - When education increases by one standard deviation, income increases on average by **0.3369 standard deviations**, controlling for age

Adjusted R^2

```
. ***Use aweight to estimate adjusted R-squared
. ***pweight and complex survey design omit sum of squares and adjusted R-squared
. reg income age educgr [aweight=perwt]
(sum of wgt is 13,849,398)

      Source |       SS           df       MS      Number of obs   =   127,785
-------------+----------------------------------   F(2, 127782)    =  12641.17
       Model |  7.4126e+13         2  3.7063e+13   Prob > F        =    0.0000
    Residual |  3.7465e+14   127,782  2.9319e+09   R-squared       =    0.1652
-------------+----------------------------------   Adj R-squared   =    0.1652
       Total |  4.4877e+14   127,784  3.5120e+09   Root MSE        =     54147

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   796.3443   10.53436    75.59   0.000     775.6972    816.9915
      educgr |   16863.33   128.6752   131.05   0.000     16611.13    17115.53
       _cons |  -31880.99   554.2213   -57.52   0.000    -32967.25   -30794.72
------------------------------------------------------------------------------
```

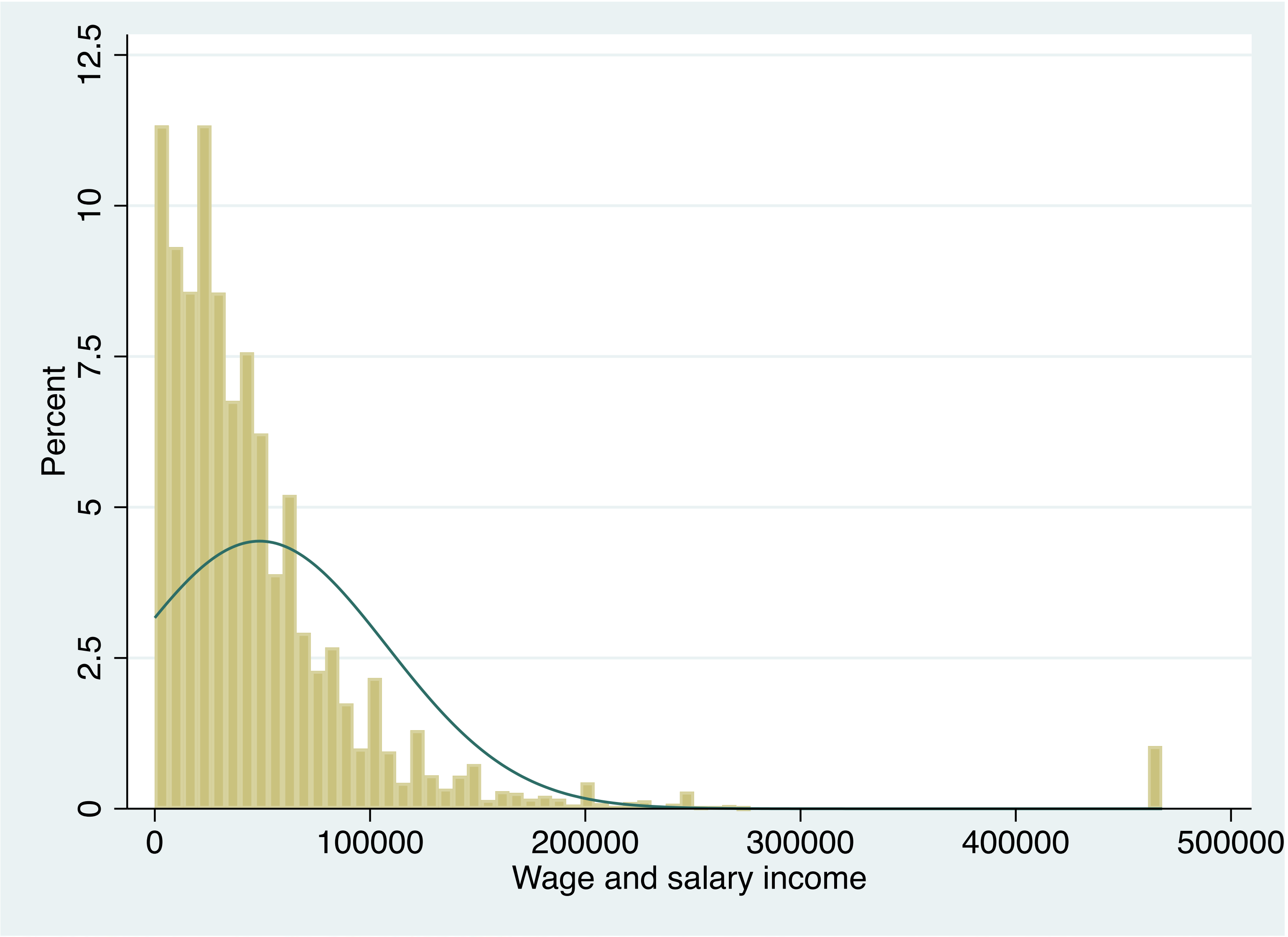
Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Determining normality

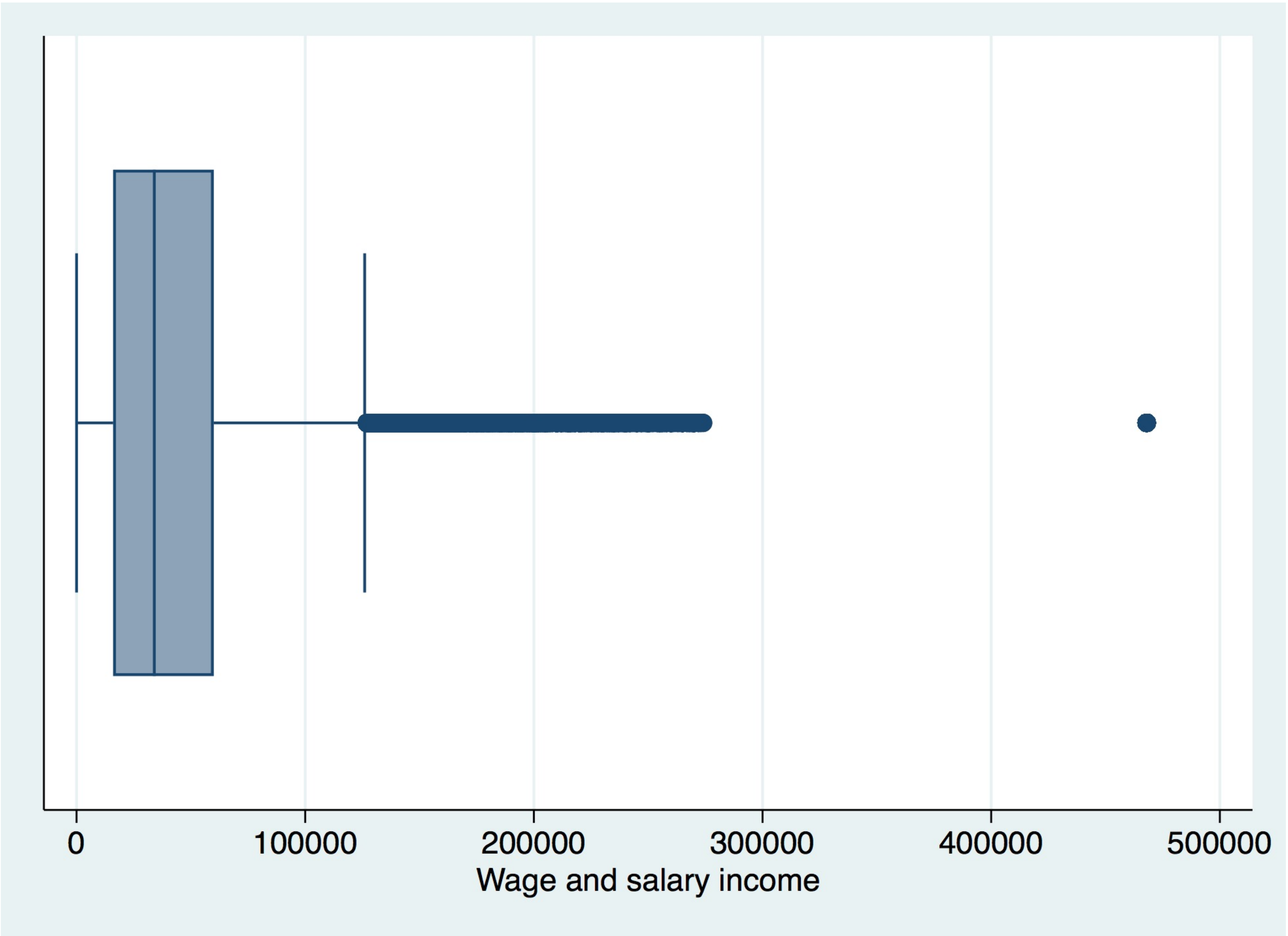
- Some statistical methods require random selection of respondents from a population with normal distribution for its variables
  - OLS regressions require normal distribution for its interval-ratio-level variables
  - We can analyze histograms, boxplots, outliers, quantile-normal plots, and measures of skewness and kurtosis to determine if variables have a normal distribution

Histogram of income



Source: 2018 American Community Survey.

Boxplot of income

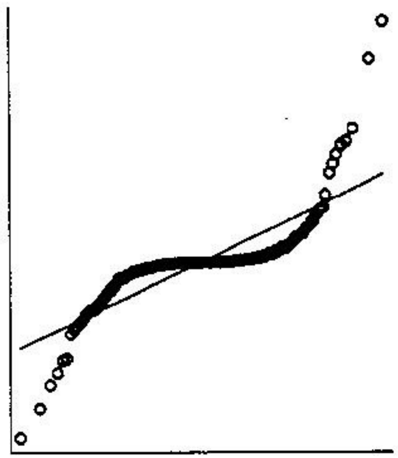


Source: 2018 American Community Survey.

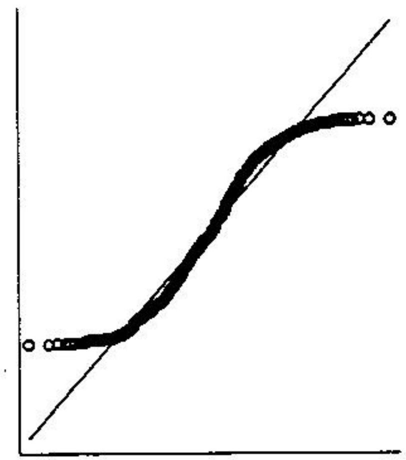
Quantile-normal plots

- A quantile-normal plot is a scatter plot
  - One axis has quantiles of the original data
  - The other axis has quantiles of the normal distribution
- If the points do not form a straight line or if the points have a non-linear symmetric pattern
  - The variable does not have a normal distribution
- If the pattern of points is roughly straight
  - The variable has a distribution close to normal
- If the variable has a normal distribution
  - The points would exactly overlap the diagonal line

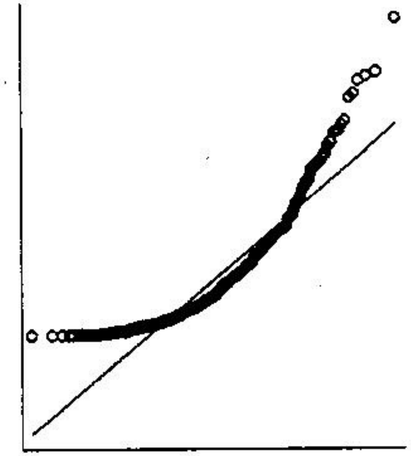
Quantile-normal plots reflect distribution shapes



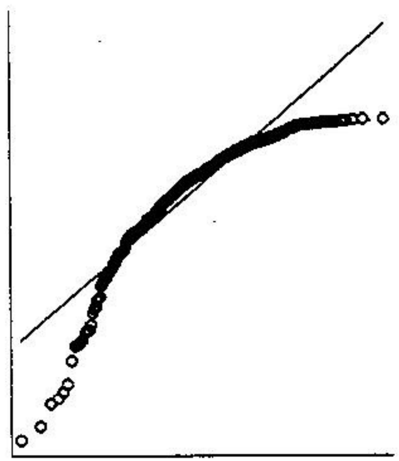
Heavy Tails, High and Low Outliers



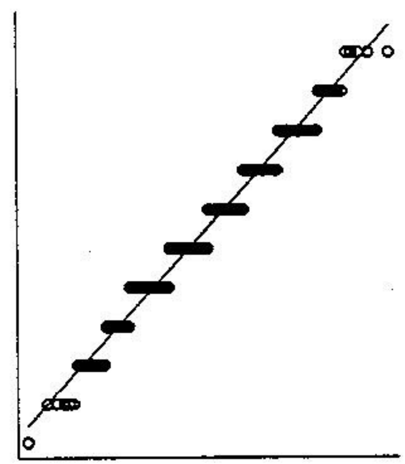
Light Tails, No Outliers



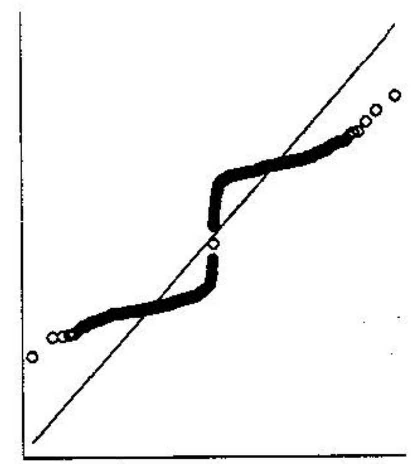
Positive Skew, High Outliers



Negative Skew, Low Outliers



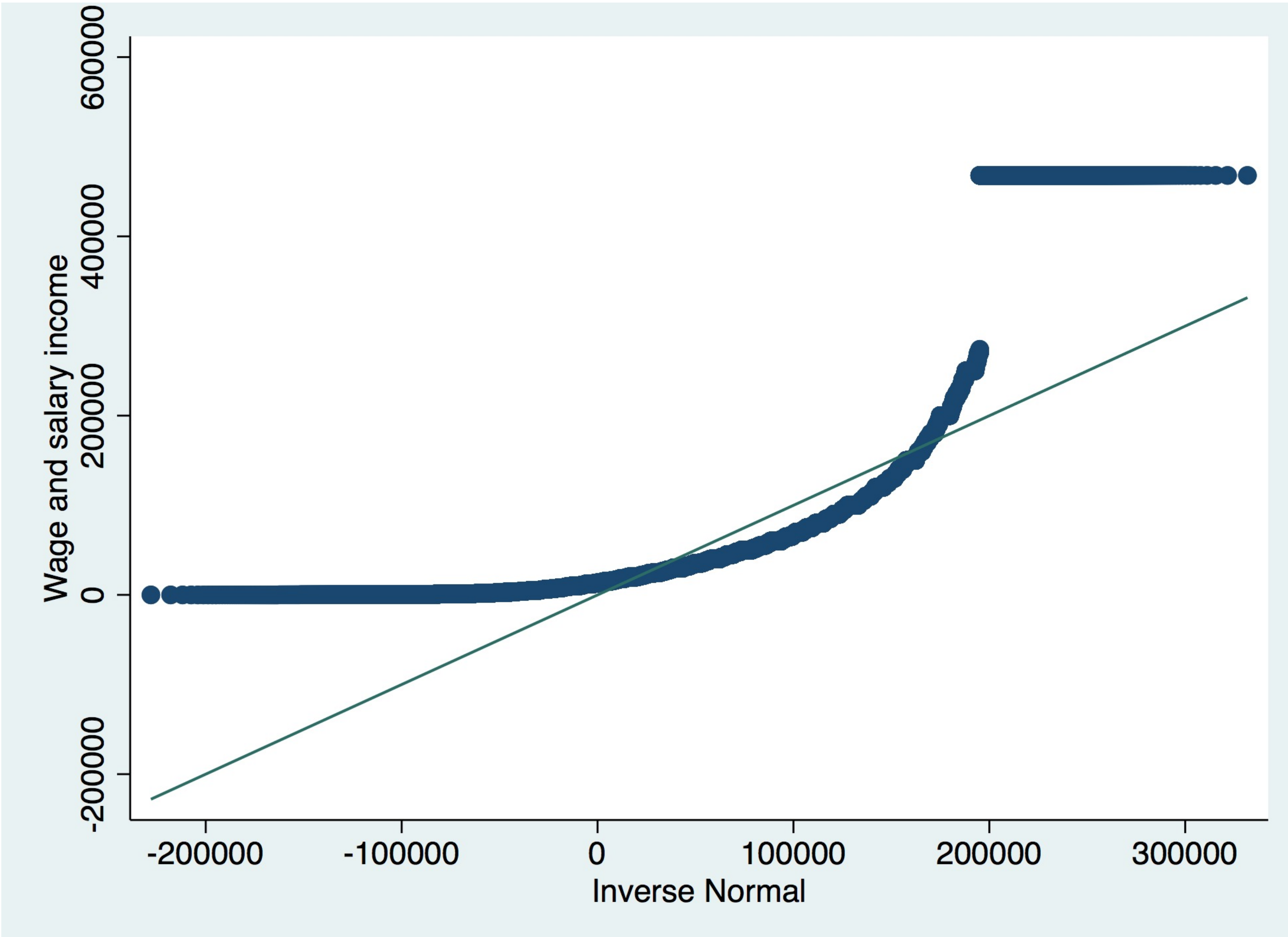
Granularity
(discrete values)



Two Peaks, Central Gap
(bimodal)

Source: Hamilton 1992, p.16.

Quantile-normal plot of income



Source: 2018 American Community Survey.

Skewness

- Skewness is a measure of symmetry
  - A distribution is symmetric if it looks the same to the left and right of the center point
  - Skewness for a normal distribution is zero
  - Negative values for the skewness indicate variable is skewed to the left (left tail is long relative to the right tail)
  - Positive values for the skewness indicate variable is skewed to the right (right tail is long relative to the left tail)
- Rule of thumb
  - Skewness between –0.5 and 0.5: variable is fairly symmetrical
  - Skewness between –1 and –0.5 or between 0.5 and 1: variable moderately skewed
  - Skewness less than –1 or greater than 1: variable is highly skewed

Source: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics

Kurtosis

- Kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution
  - Variables with high kurtosis tend to have heavy tails or outliers
  - Variables with low kurtosis tend to have light tails or lack of outliers
  - A uniform distribution would be the extreme case
  - **The kurtosis for a standard normal distribution is three**
- Excess kurtosis
  - Some sources subtract 3 from the kurtosis
  - The standard normal distribution has an excess kurtosis of zero
  - Positive excess kurtosis indicates a “heavy-tailed” distribution
  - Negative excess kurtosis indicates a “light tailed” distribution

Source: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm
https://www.spcforexcel.com/knowledge/basic-statistics/are-skewness-and-kurtosis-useful-statistics
https://www.stata-journal.com/sjpdf.html?articlenum=st0204

Skewness and Kurtosis

```
. sum income if income!=0 [fweight=perwt], d

                          income
-------------------------------------------------------------
      Percentiles      Smallest
 1%          500              4
 5%         2400              4
10%         5600              4       Obs          13,849,398
25%        16000              4       Sum of Wgt.  13,849,398

50%        34000                      Mean           48713.66
                        Largest       Std. Dev.      59261.63
75%        60000         468000
90%       100000         468000       Variance       3.51e+09
95%       136000         468000       Skewness        4.20286
99%       468000         468000       Kurtosis       27.61478
```

Source: 2018 American Community Survey.

Power transformation

- Lawrence Hamilton ("Regression with Graphics", 1992, p.18–19)

$$y^3 \longrightarrow q = 3$$

$$y^2 \longrightarrow q = 2$$

$$y^1 \longrightarrow q = 1$$

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

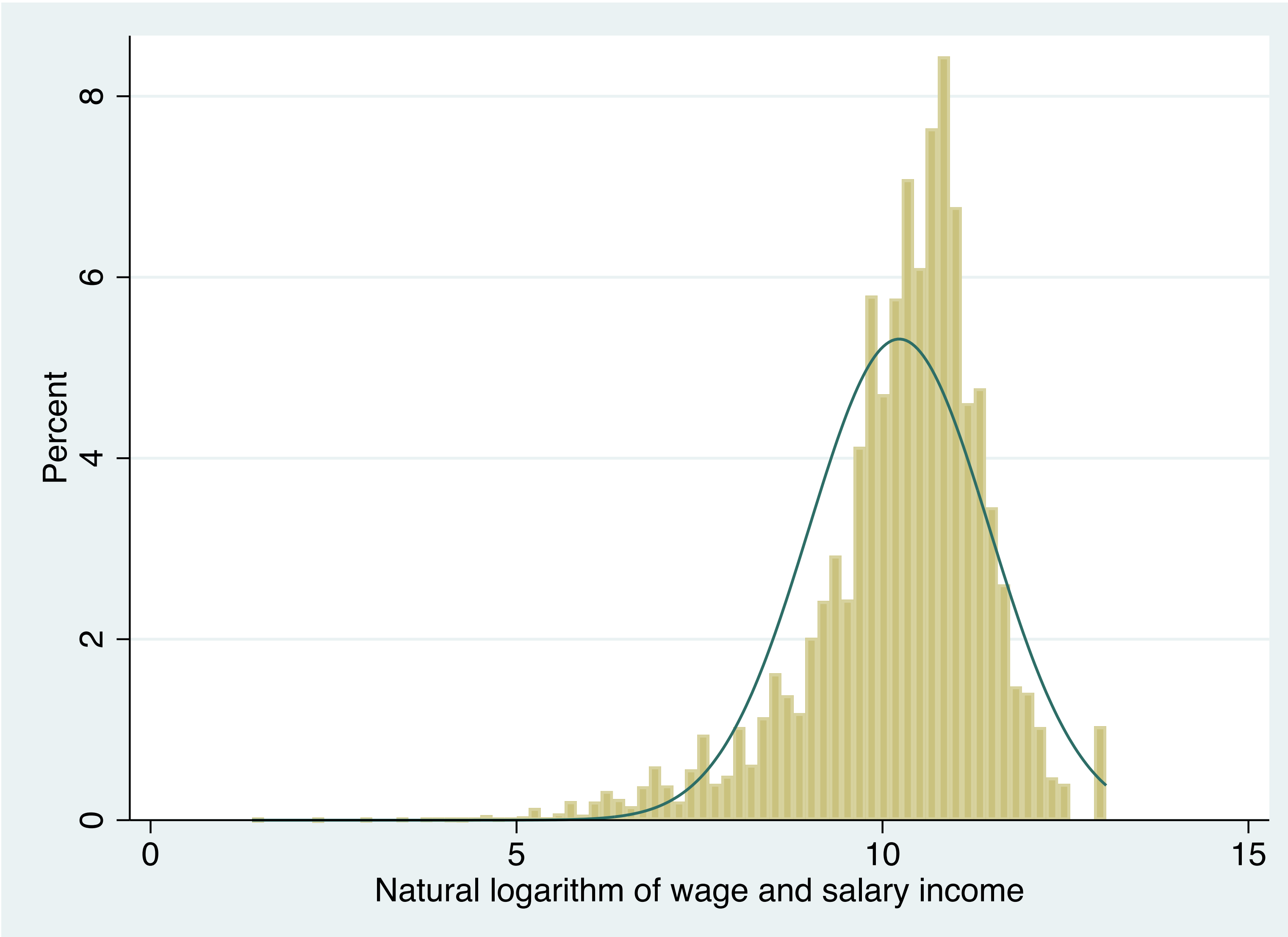
$$\log(y) \longrightarrow q = 0$$

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

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

- $q > 1$: 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

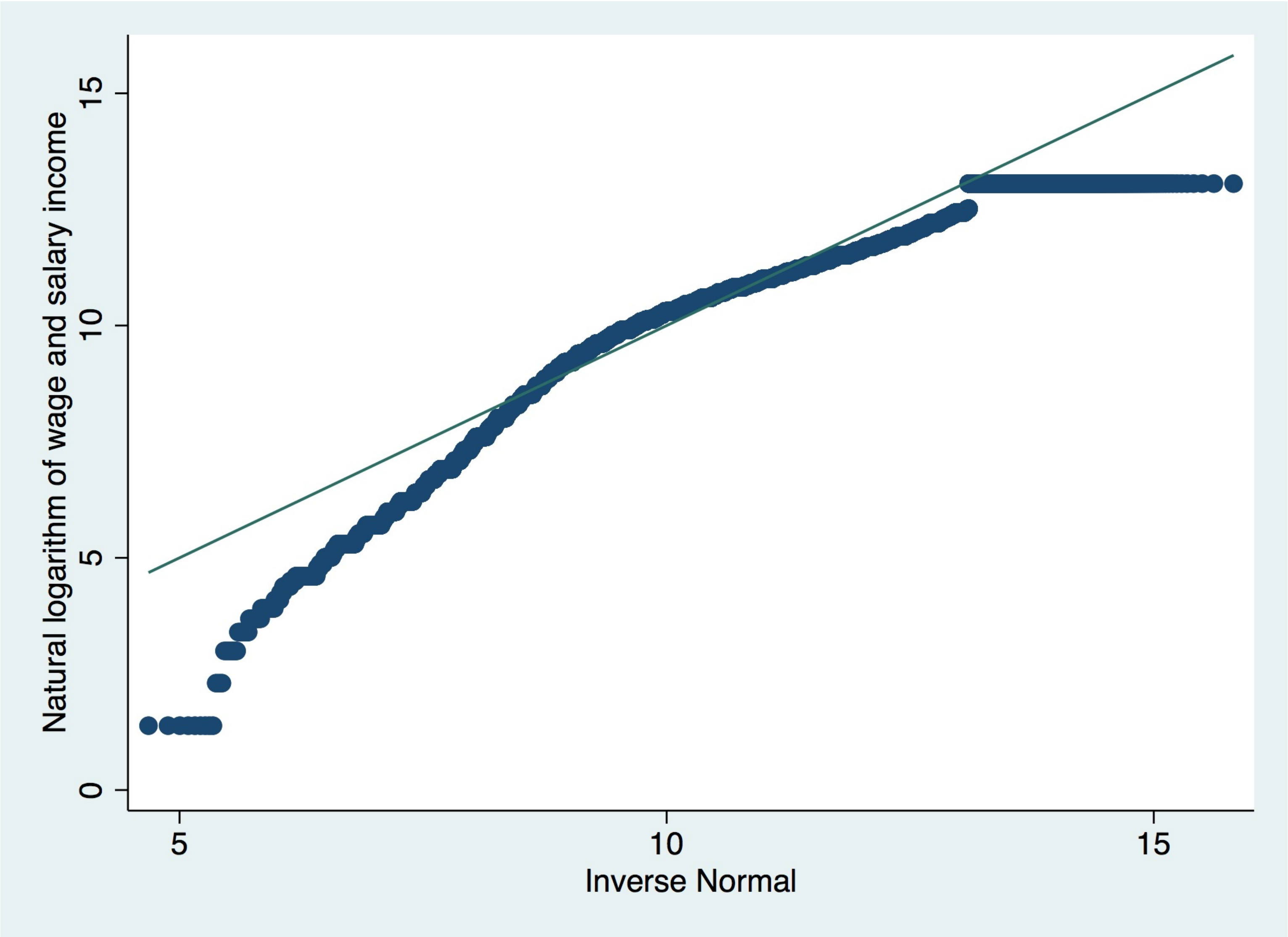


Source: 2018 American Community Survey.

Boxplot of log of income

Source: 2018 American Community Survey.

Quantile-normal plot of log of income



Source: 2018 American Community Survey.

Skewness and Kurtosis

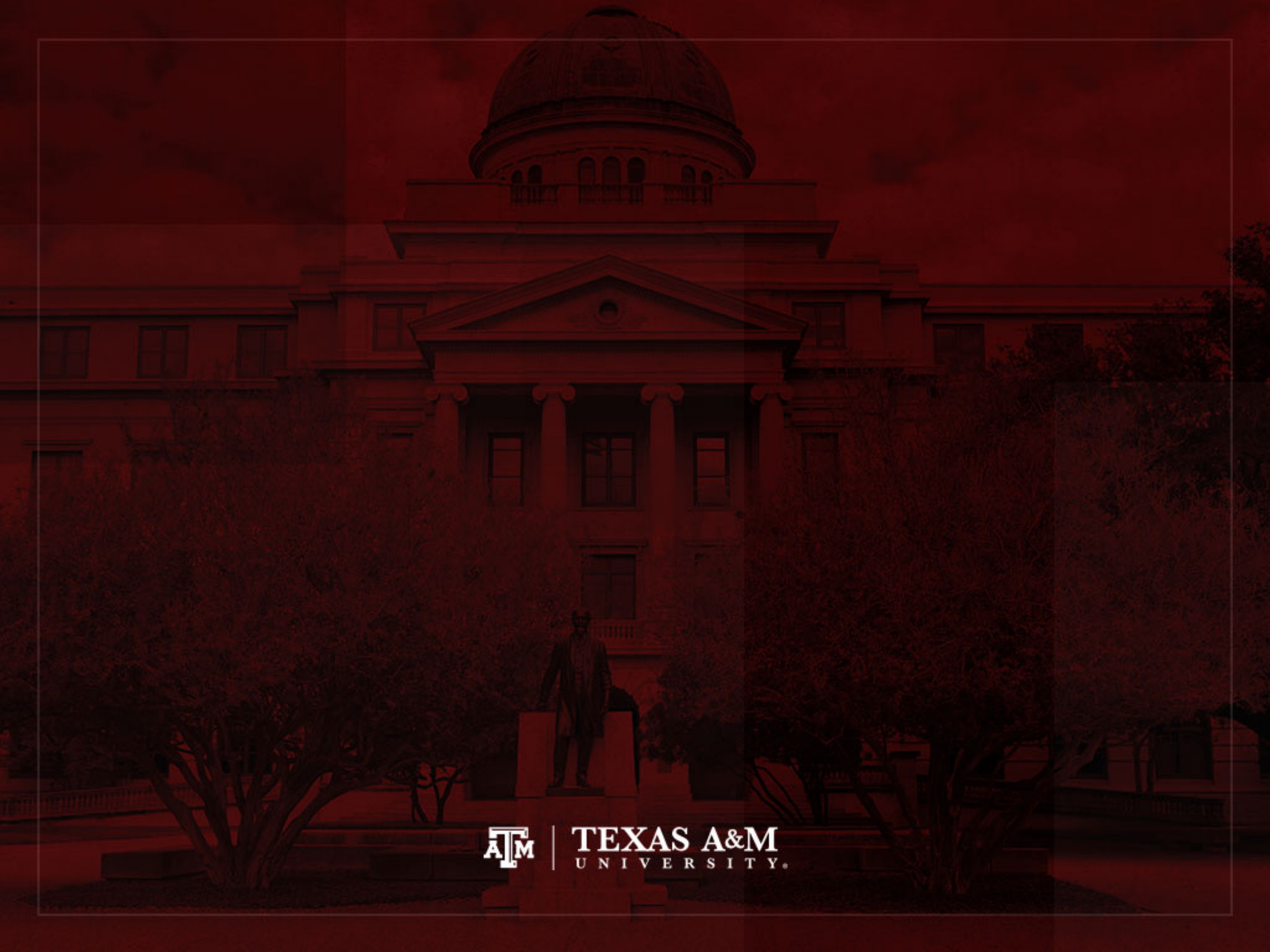
```
. sum lnincome [fweight=perwt], d

                          lnincome
-------------------------------------------------------------
      Percentiles      Smallest
 1%     6.214608       1.386294
 5%     7.783224       1.386294
10%     8.630522       1.386294       Obs          13,849,398
25%     9.680344       1.386294       Sum of Wgt.  13,849,398

50%     10.43412                      Mean           10.22871
                        Largest       Std. Dev.      1.233225
75%      11.0021       13.05622
90%     11.51293       13.05622       Variance       1.520844
95%     11.82041       13.05622       Skewness      -1.123294
99%     13.05622       13.05622       Kurtosis       5.349345
```

Source: 2018 American Community Survey.

Interpretation of ln(income)

(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(β_1) times
    - x_1 increases by one unit, y increases on average **1.13 times**, controlling for other independent variables
  - 100*[exp(β_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*β percent
    - x_1 increases by one unit, y increases on average **approximately by 12%**, controlling for other independents


ATM
TEXAS A&M
UNIVERSITY.

ln(income) = F(age, education)

```
. ***Use complex survey design
. svy: reg lnincome age educgr
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =       212        Number of obs     =     127,785
Number of PSUs     =    79,499        Population size   =  13,849,398
                                      Design df         =      79,287
                                      F(   2,  79286)   =     7451.80
                                      Prob > F          =      0.0000
                                      R-squared         =      0.1932
```

lnincome	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0224959	.0003153	71.35	0.000	.0218779	.0231139
educgr	.3381717	.0032453	104.20	0.000	.331811	.3445324
_cons	8.34881	.0175456	475.84	0.000	8.31442	8.383199

Source: 2018 American Community Survey.

Exponential of coefficients

```
. ***Automatically see exponential of coefficients
. svy: reg lnincome age educgr, eform(Exp. Coef.)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        212        Number of obs     =    127,785
Number of PSUs     =     79,499        Population size   = 13,849,398
                                       Design df         =     79,287
                                       F(   2,  79286)   =    7451.80
                                       Prob > F          =     0.0000
                                       R-squared         =     0.1932
```

lnincome	Exp. Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	1.022751	.0003225	71.35	0.000	1.022119	1.023383
educgr	1.402381	.0045511	104.20	0.000	1.393489	1.41133
_cons	4225.149	74.13273	475.84	0.000	4082.319	4372.976

Source: 2018 American Community Survey.

Interpretation of age

(income with continuous independent variables)

- Coefficient for **<u>age</u>** equals 0.0225
  - exp(β_1) times
    - When age increases by one unit, income increases on average by **<u>1.0228 times</u>**, controlling for education
  - 100*[exp(β_1)–1] percent
    - When age increases by one unit, income increases on average by **<u>2.28%</u>**, controlling for education
  - 100*β_1 percent
    - When age increases by one unit, income increases on average **<u>approximately by 2.25%</u>**, controlling for education

Interpretation of education

(income with continuous independent variables)

- Coefficient for **education** equals 0.3382
  - $\exp(\beta_1)$ times
    - When education increases by one unit, income increases on average by **1.4024 times**, controlling for age
  - $100*[\exp(\beta_1)–1]$ percent
    - When education increases by one unit, income increases on average by **40.24%**, controlling for age
  - $100*\beta_1$ percent
    - When education increases by one unit, income increases on average **approximately by 33.82%**, controlling for age

Standardized coefficients

```
. ***Standardized regression coefficients
. ***(i.e., standardized partial slopes, beta-weights)
. ***It does not allow the use of complex survey design
. ***Use pweight to maintain sample size and estimate robust standard errors
. reg lnincome age educgr [pweight=perwt], beta
(sum of wgt is 13,849,398)

Linear regression                               Number of obs     =    127,785
                                                F(2, 127782)      =    7996.52
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1932
                                                Root MSE          =     1.1077
```

lnincome	Coef.	Robust Std. Err.	t	P>|t|	Beta
age	.0224959	.0002969	75.76	0.000	.2637902
educgr	.3381717	.0031694	106.70	0.000	.3246429
_cons	8.34881	.0166508	501.41	0.000	.

Source: 2018 American Community Survey.

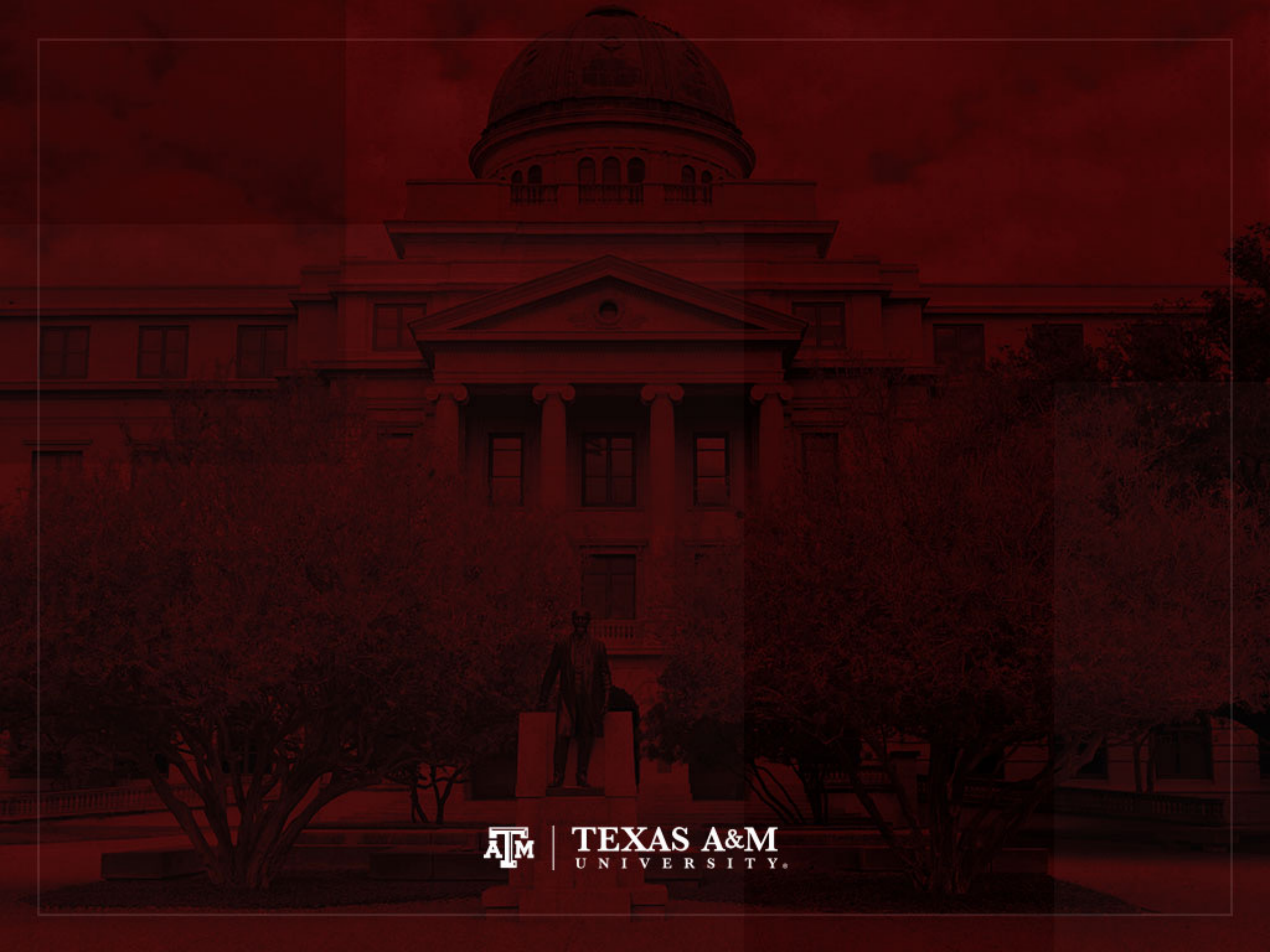
Interpretation of standardized
(income with continuous independent variables)

- Coefficient for **age** equals 0.2638
  - $\exp(\beta_1)$ times
    - When age increases by one standard deviation, income increases on average by **1.3019 times**, controlling for education
  - $100*[\exp(\beta_1)–1]$ percent
    - When age increases by one standard deviation, income increases on average by **30.19%**, controlling for education
  - $100*\beta_1$ percent
    - When age increases by one standard deviation, income increases on average **approximately by 26.38%**, controlling for education

Adjusted R^2

```
. ***Use aweight to estimate adjusted R-squared
. ***pweight and complex survey design omit sum of squares and adjusted R-squared
. reg lnincome age educgr [aweight=perwt]
(sum of wgt is 13,849,398)

      Source |       SS           df       MS      Number of obs   =   127,785
-------------+----------------------------------   F(2, 127782)    =  15298.69
       Model |  37544.8387         2  18772.4194   Prob > F        =    0.0000
    Residual |  156796.221   127,782  1.22706031   R-squared       =    0.1932
-------------+----------------------------------   Adj R-squared   =    0.1932
       Total |  194341.059   127,784  1.52085597   Root MSE        =    1.1077

------------------------------------------------------------------------------
    lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0224959   .0002155   104.39   0.000     .0220735    .0229183
      educgr |   .3381717   .0026324   128.47   0.000     .3330122    .3433311
       _cons |    8.34881   .0113381   736.35   0.000     8.326587    8.371032
------------------------------------------------------------------------------
```

Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Predicted values

- We can estimate the predicted values of the dependent variable for each individual in the dataset
- Use the estimated coefficients from the regression model

$$y_i' = \hat{y}_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$$

Predicted income

- Income = F(age, education)

income	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	796.3443	11.73077	67.89	0.000	773.3521	819.3366
educgr	16863.33	179.705	93.84	0.000	16511.11	17215.55
_cons	-31880.99	661.937	-48.16	0.000	-33178.38	-30583.59

- Use the regression equation to predict income for someone with 45 years of age and college education

$$\hat{y} = -31{,}880.99 + 796.34(\text{age}) + 16{,}863.33(\text{educgr})$$

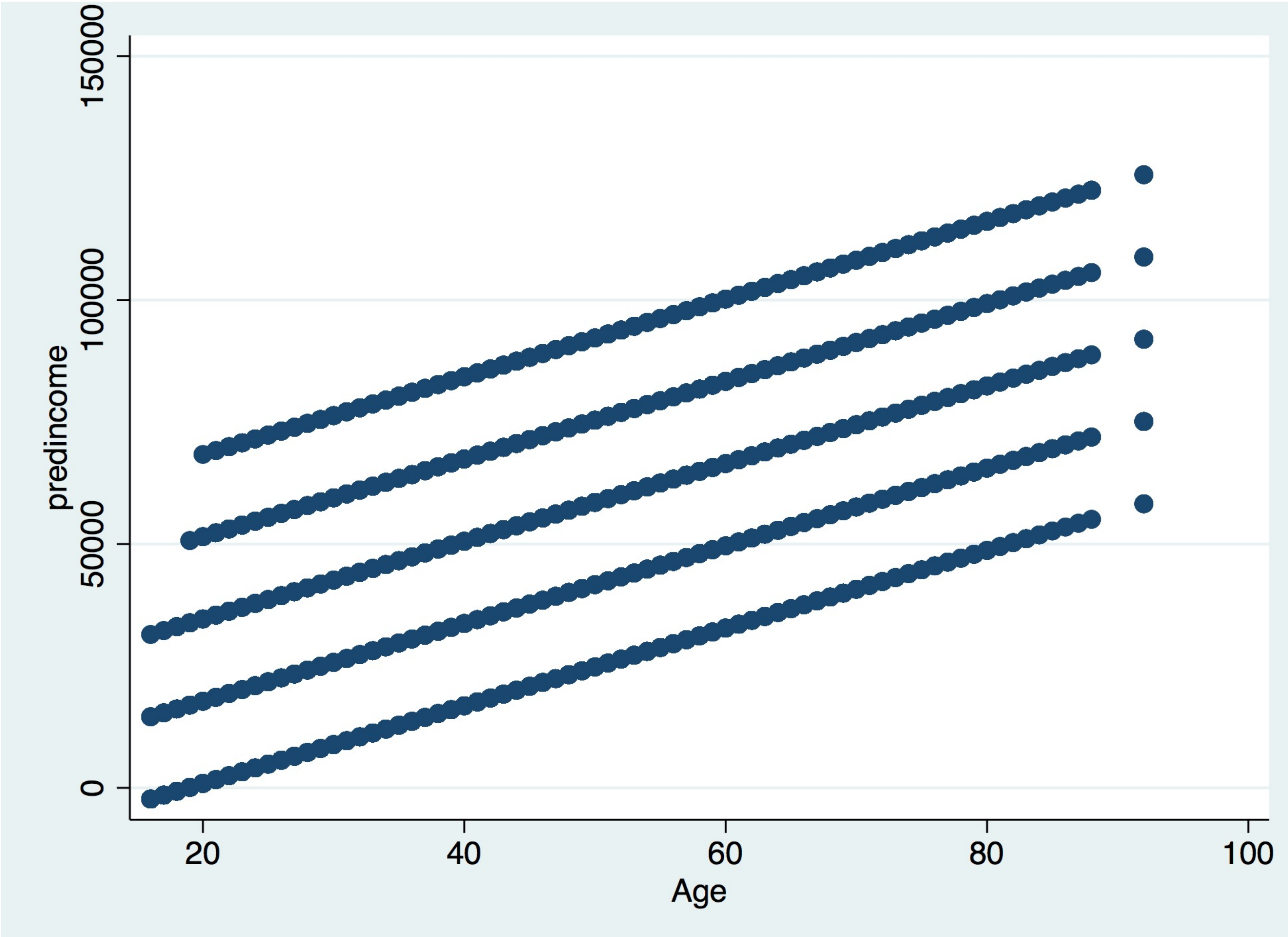
$$\hat{y} = -31{,}880.99 + (796.34)(45) + (16{,}863.33)(4)$$

$$\hat{y} = 71{,}407.63$$

- Under these conditions, we would predict 71,407.63 dollars for that individual

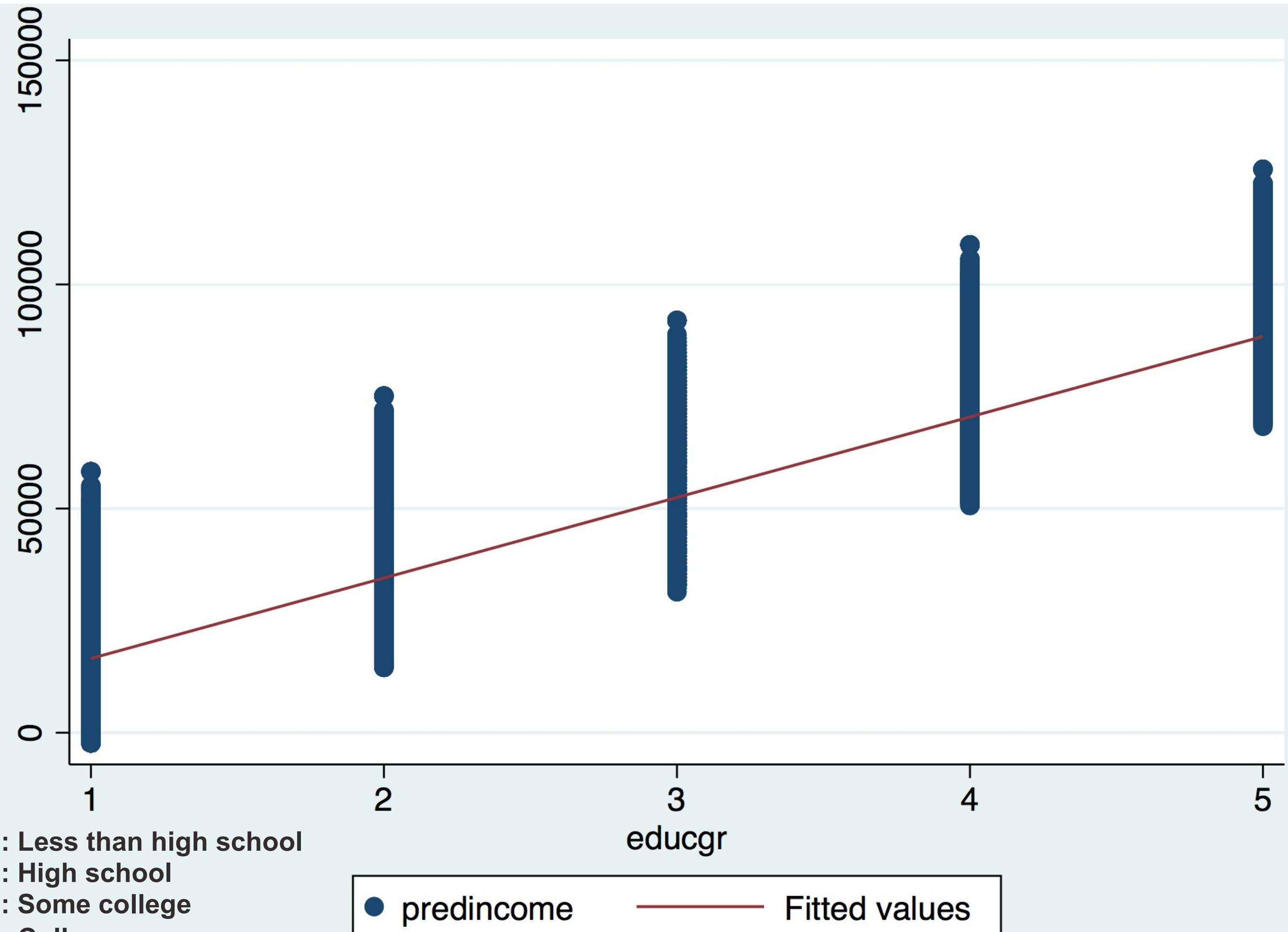
Source: 2018 American Community Survey.

Predicted income by age



Source: 2018 American Community Survey.

Predicted income by education



1: Less than high school
2: High school
3: Some college
4: College
5: Graduate school

Source: 2018 American Community Survey.

Predicted log of income

- ln(income) = F(age, education)

lnincome	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0224959	.0003153	71.35	0.000	.0218779	.0231139
educgr	.3381717	.0032453	104.20	0.000	.331811	.3445324
_cons	8.34881	.0175456	475.84	0.000	8.31442	8.383199

- Use the regression equation to predict log of income for someone with 45 years of age and college education

$$ln(\hat{y}) = 8.3488 + 0.0225(\text{age}) + 0.3382(\text{educgr})$$

$$ln(\hat{y}) = 8.3488 + (0.0225)(45) + (0.3382)(4)$$

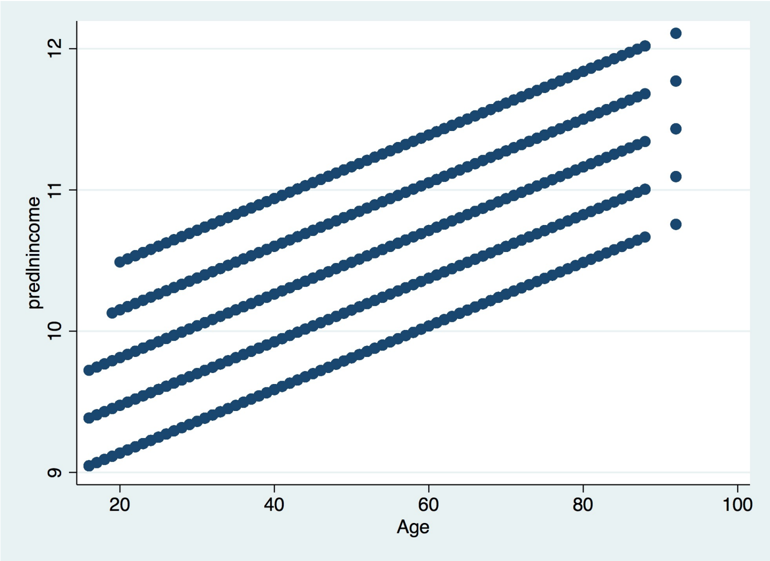
$$ln(\hat{y}) = 10.7141$$

$$\hat{y} = 44{,}985.70$$

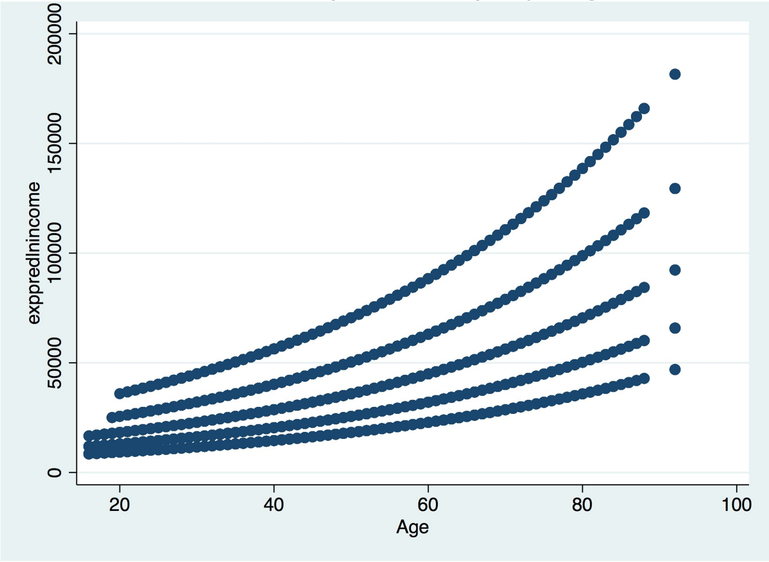
- Under these conditions, we would predict 44,985.70 dollars for that individual

Source: 2018 American Community Survey.

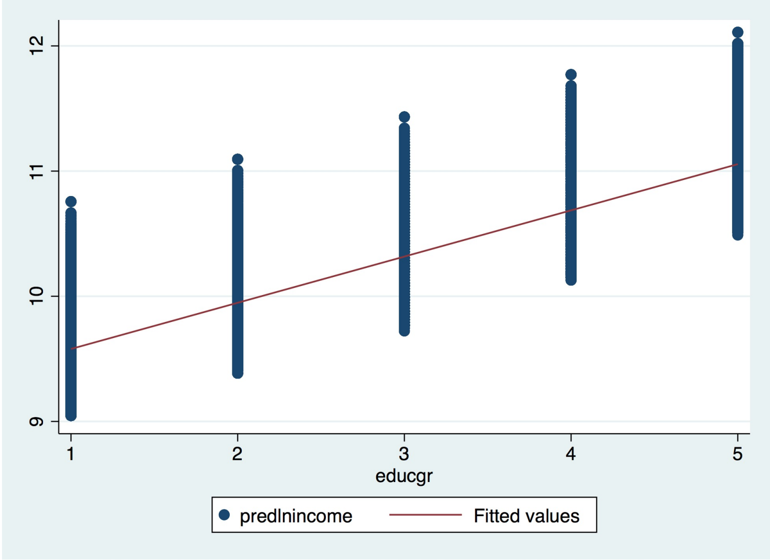
Predicted ln(income) by age



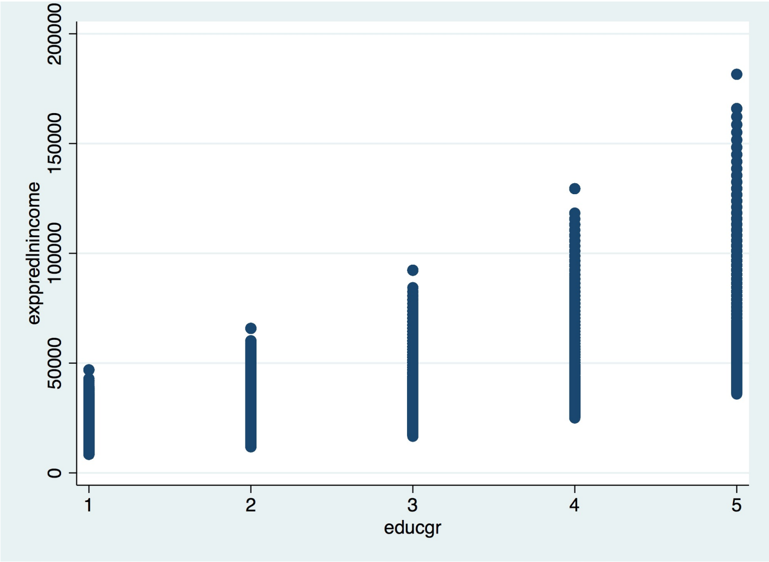
Exponential of predicted ln(income) by age



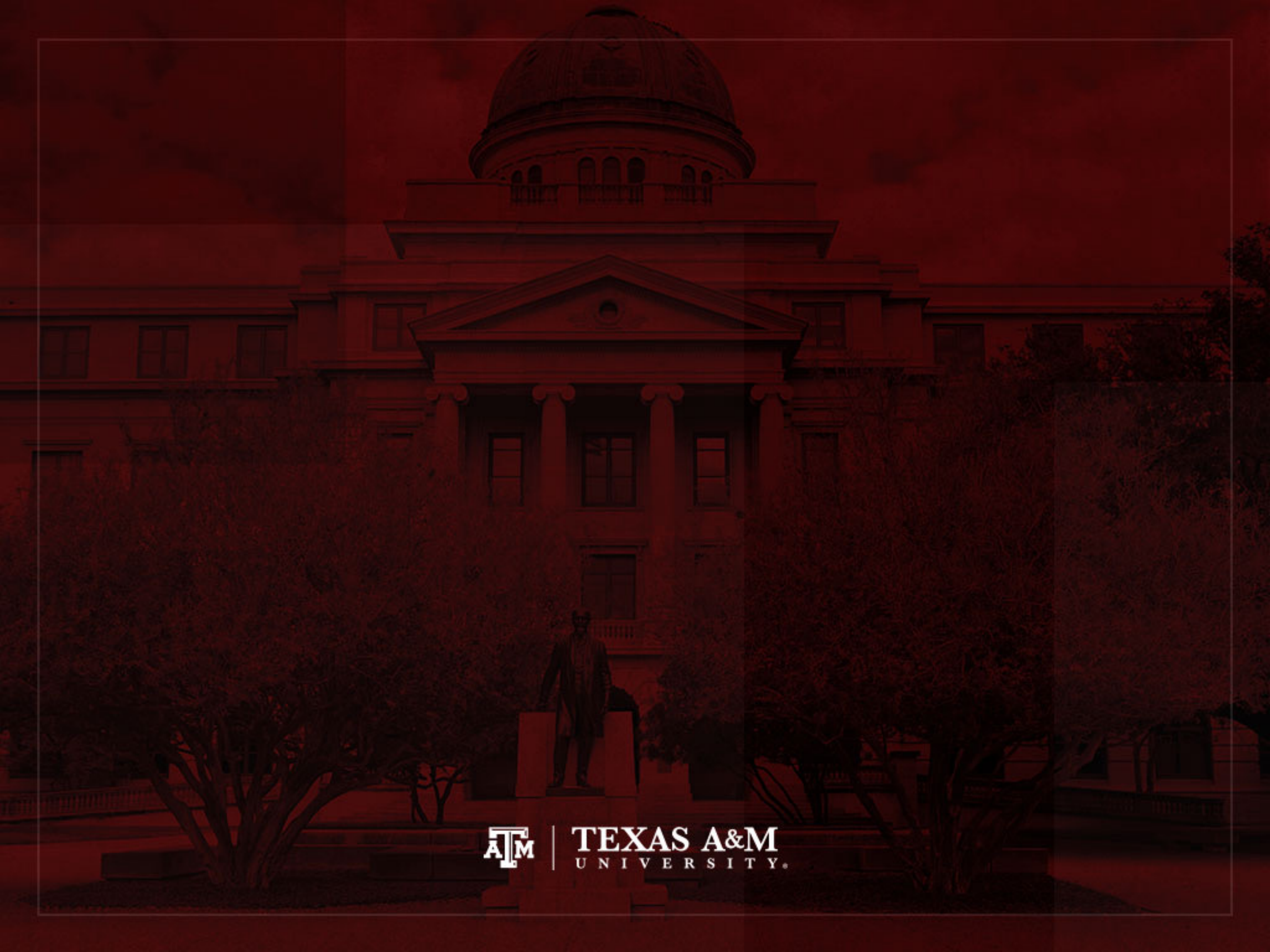
Predicted ln(income) by education



Exponential of predicted ln(income) by education



Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Residual analysis with graphs

- Homoscedasticity assumption
  - 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
- The same assumption applies to residuals
  - Difference between observed value (*y*) and predicted value ($\hat{y}$)
  - $e = y - \hat{y}$
  - We can plot residuals against predicted values $\hat{y}$ (which summarize all *x* variables)

Microdata

	age	educgr	income	predincome	resincome	lnincome	predlnincome	reslnincome
1	21	2	3200	18568.9	−15368.9	8.070906	9.497567	−1.426661
2	20	2	35000	17772.56	17227.44	10.4631	9.475071	.9880321
3	31	2	10000	26532.34	−16532.34	9.21034	9.722527	−.5121856
4	39	4	30000	66629.76	−36629.75	10.30895	10.57884	−.2698844
5	18	2	1500	16179.87	−14679.87	7.313221	9.430079	−2.116859
6	25	1	13000	4890.951	8109.049	9.472705	9.249379	.2233258
7	20	3	5600	34635.88	−29035.88	8.630522	9.813243	−1.182721
8	34	2	65000	28921.38	36078.62	11.08214	9.790014	1.292129
9	18	2	4000	16179.87	−12179.87	8.294049	9.430079	−1.13603
10	18	3	1400	33043.2	−31643.2	7.244227	9.768251	−2.524024
11	20	2	5000	17772.56	−12772.56	8.517193	9.475071	−.9578784
12	18	2	2300	16179.87	−13879.87	7.740664	9.430079	−1.689415
13	20	2	18000	17772.56	227.4432	9.798127	9.475071	.323056
14	19	3	14000	33839.54	−19839.54	9.546813	9.790747	−.243934
15	20	2	6000	17772.56	−11772.56	8.699514	9.475071	−.7755568
16	19	2	1800	16976.21	−15176.21	7.495542	9.452576	−1.957033
17	21	3	320	35432.23	−35112.23	5.768321	9.835739	−4.067418
18	22	3	1900	36228.57	−34328.57	7.549609	9.858234	−2.308625
19	46	2	28000	38477.51	−10477.51	10.23996	10.05997	.179995
20	20	3	5000	34635.88	−29635.88	8.517193	9.813243	−1.29605
21	23	3	1000	37024.92	−36024.92	6.907755	9.880731	−2.972975
22	19	2	10000	16976.21	−6976.212	9.21034	9.452576	−.2422348
23	19	3	600	33839.54	−33239.54	6.39693	9.790747	−3.393817
24	20	3	10000	34635.88	−24635.88	9.21034	9.813243	−.6029024
25	22	3	7000	36228.57	−29228.57	8.853665	9.858234	−1.004569
26	22	3	4000	36228.57	−32228.57	8.294049	9.858234	−1.564185
27	48	3	11000	56933.53	−45933.53	9.305651	10.44313	−1.137478
28	23	3	140	37024.92	−36884.92	4.941642	9.880731	−4.939088
29	21	3	2000	35432.23	−33432.23	7.600903	9.835739	−2.234836
30	21	3	3600	35432.23	−31832.23	8.188689	9.835739	−1.64705

Source: 2018 American Community Survey.

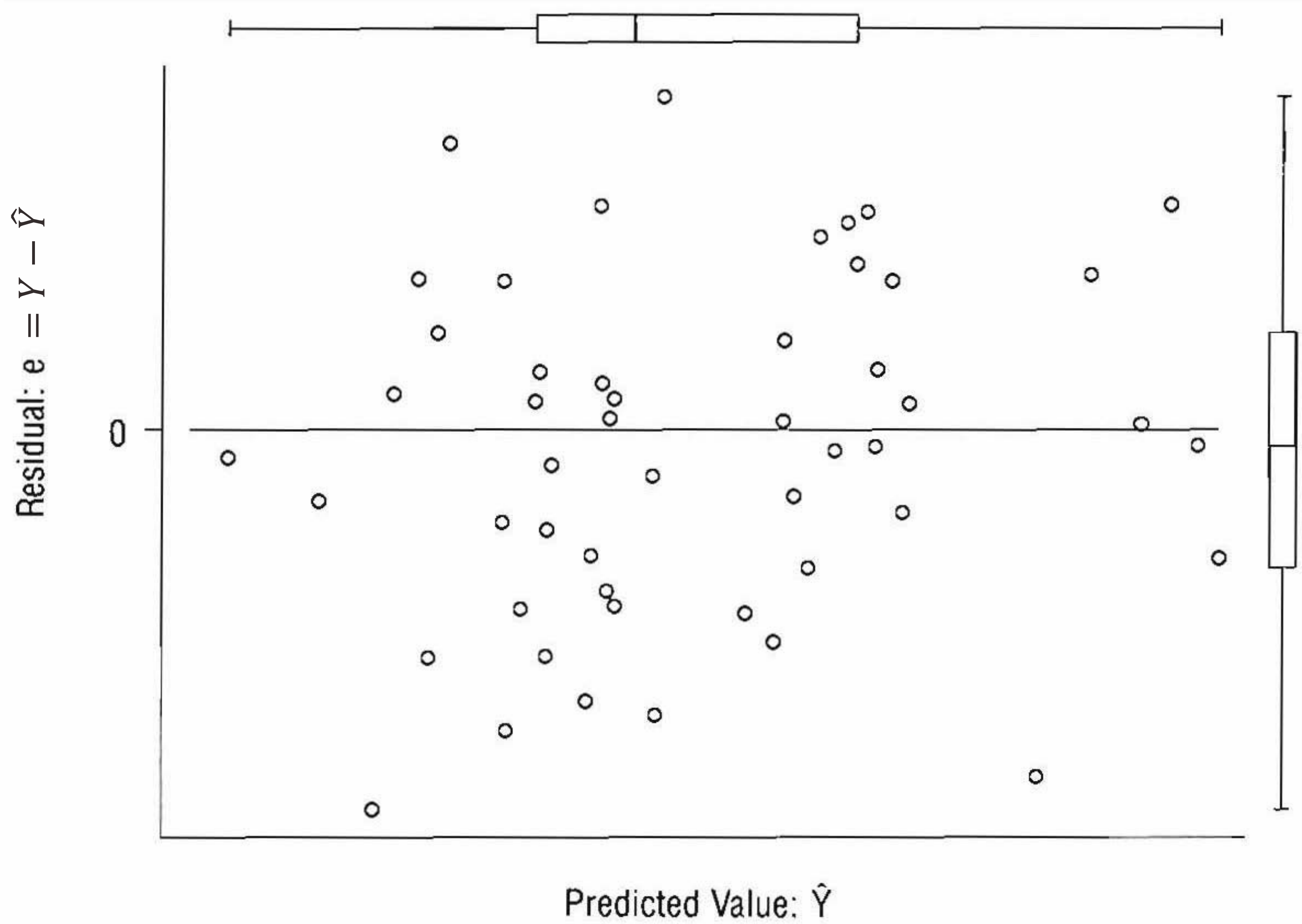


Figure 2.10 "All clear" e-versus-$\hat{Y}$ plot (artificial data).

Fonte: Hamilton, 1992: 52.

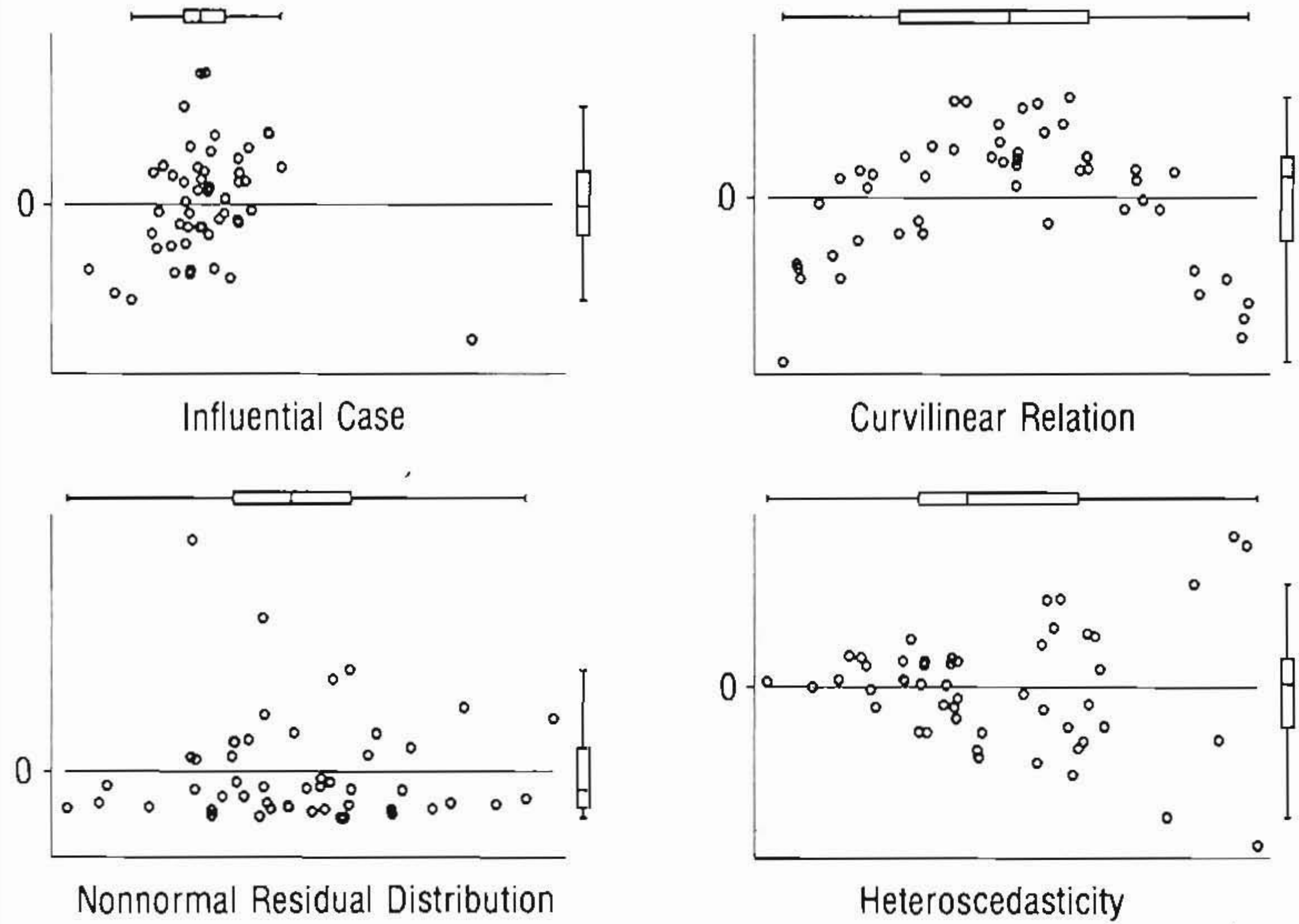
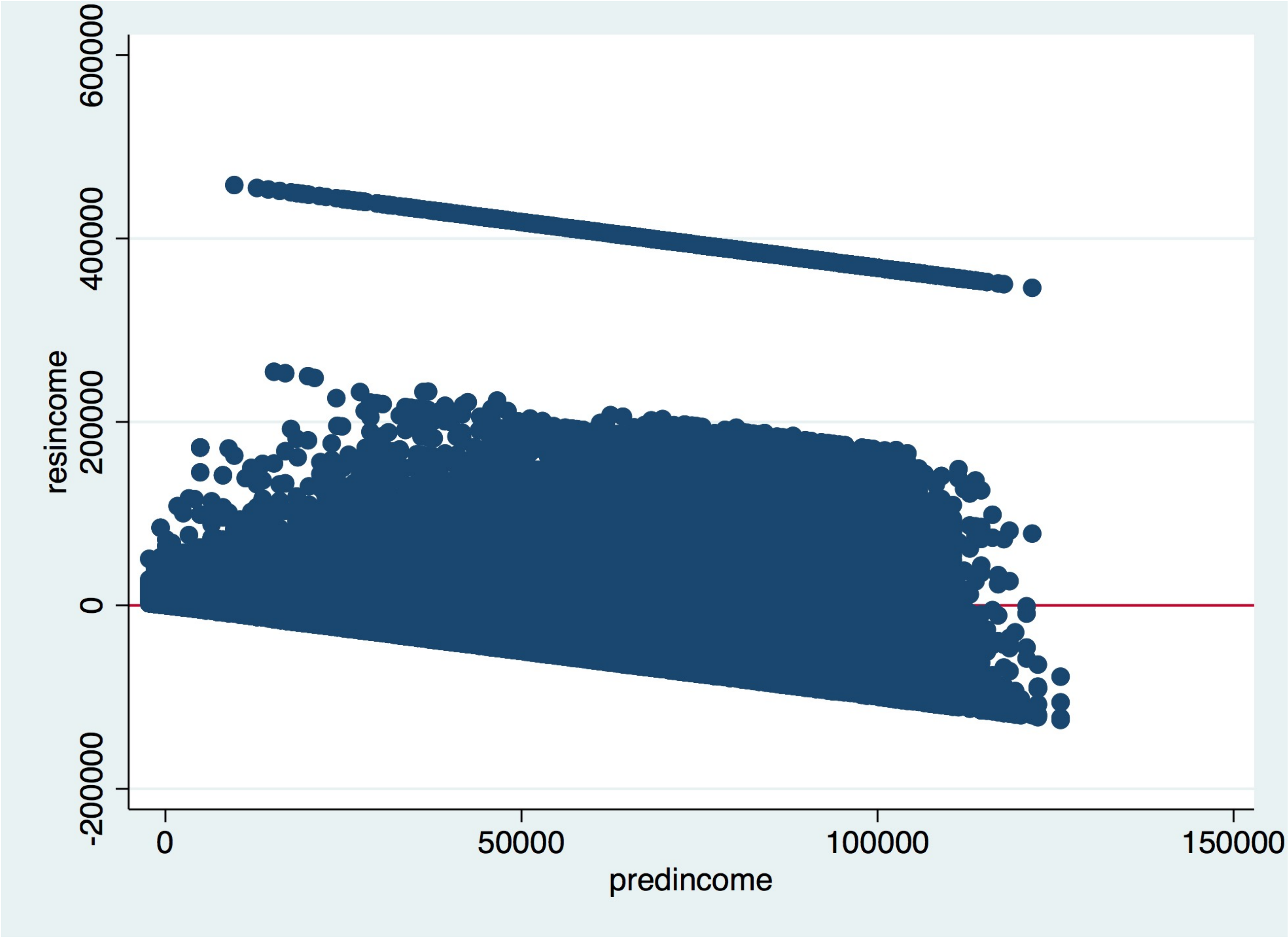


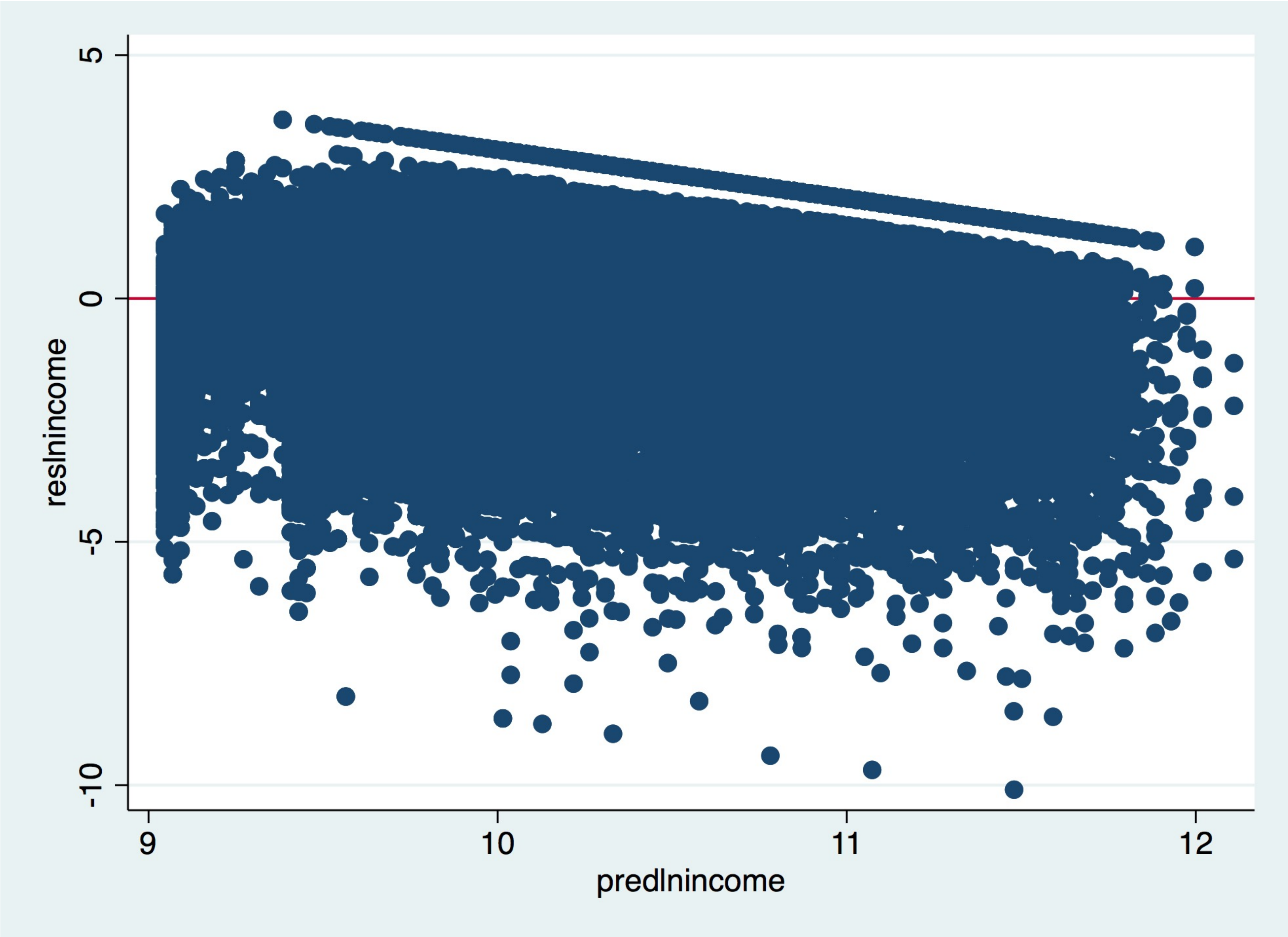
Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

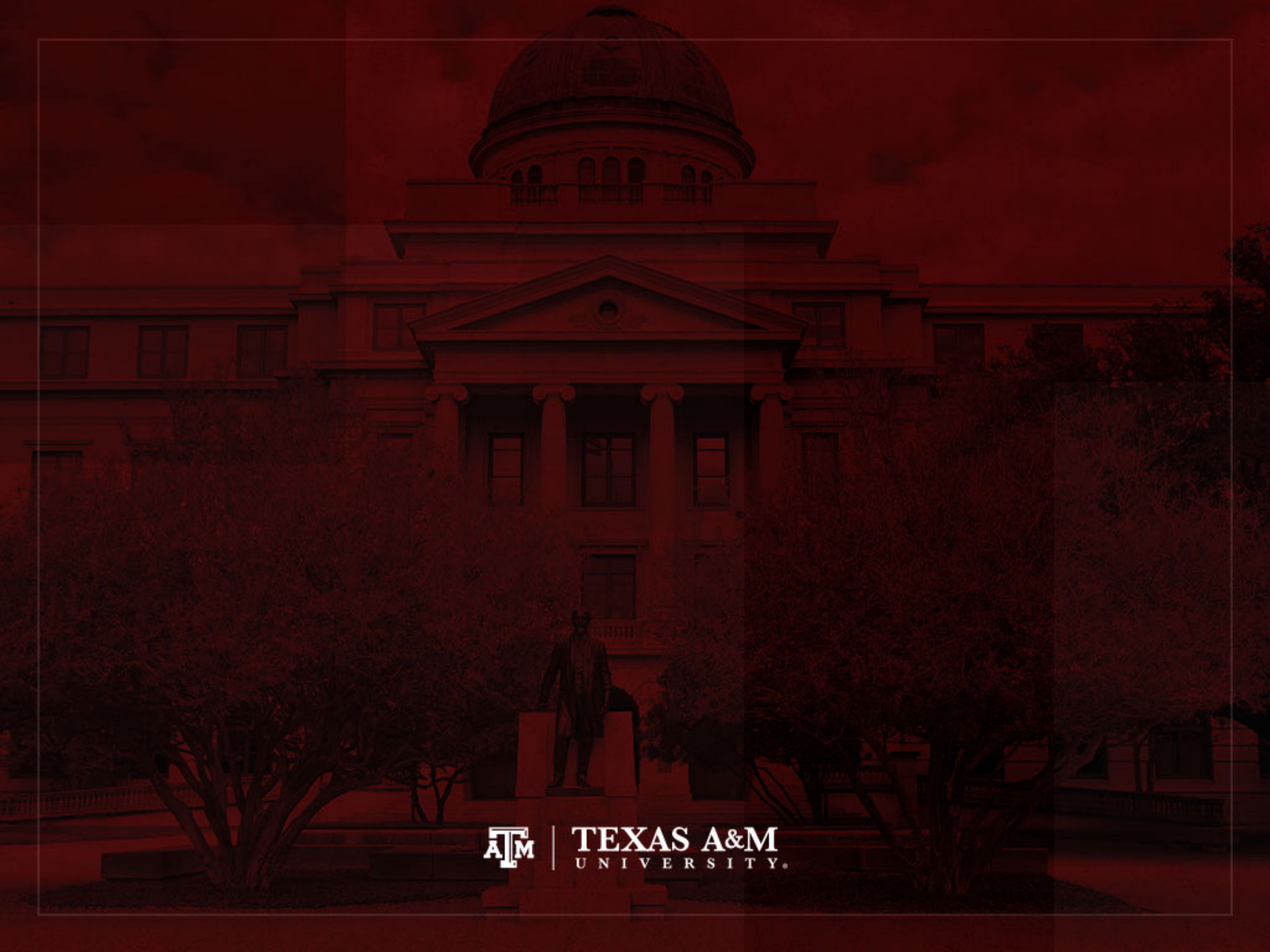
Fonte: Hamilton, 1992: 53.

Residuals: Income=F(age, education)



Residuals: ln(income)=F(age, education)




TEXAS A&M
UNIVERSITY

OLS with age and age squared

- ln(income) as a function of age and age squared

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

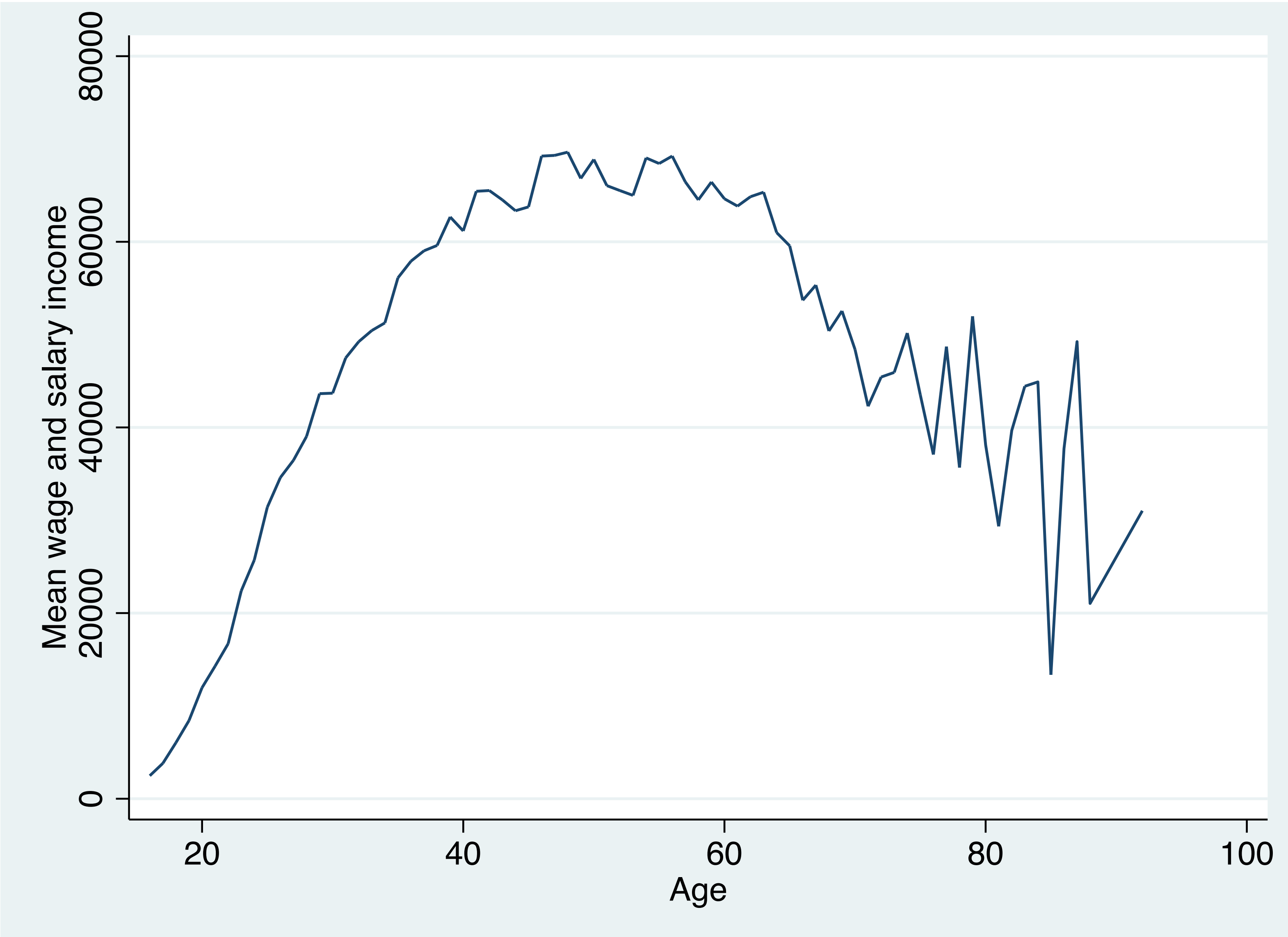
- Variation in income due to variation in age

$$\Delta y / \Delta x \approx \beta_1 + 2\beta_2 x$$

- Marginal effect of age on income depends on β_1, β_2, and specific age value (x)
- There is a positive value of x, in which the effect of x on y is zero, called the critical point (x^*)

$$x^* = |\beta_1/(2\beta_2)|$$

Mean income by age



Source: 2018 American Community Survey.

ln(income) = F(age, age squared)

```
. ***OLS with natural logarithm of income, age, and age squared
. svy: reg lnincome age agesq
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =       212              Number of obs     =    127,785
Number of PSUs     =    79,499              Population size   = 13,849,398
                                            Design df         =     79,287
                                            F(   2,  79286)   =    7983.37
                                            Prob > F          =     0.0000
                                            R-squared         =     0.2185
```

lnincome	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.1943162	.0017962	108.18	0.000	.1907956	.1978369
agesq	-.0019721	.0000205	-96.06	0.000	-.0020123	-.0019319
_cons	6.009389	.0368055	163.27	0.000	5.937251	6.081528

Source: 2018 American Community Survey.

Association of income with age

- Variation in income due to variation in age

$$\Delta\ln(\text{income}) / \Delta\text{age} \approx \beta_1 + 2\beta_2(\text{age})$$

$$\Delta\ln(\text{income}) / \Delta\text{age} \approx 0.1943 + 2(-0.0020)(\text{age})$$

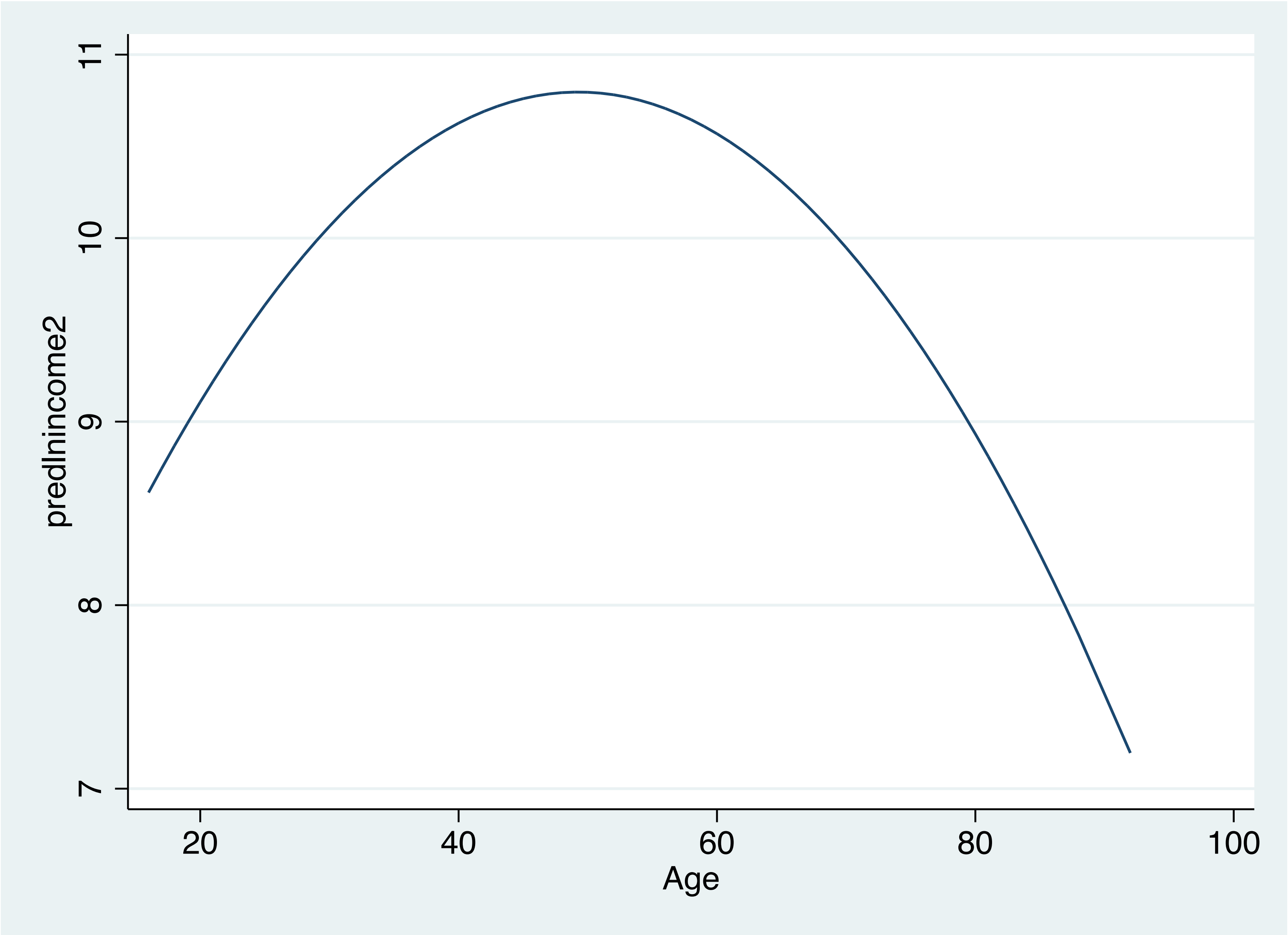
$$\Delta\ln(\text{income}) / \Delta\text{age} \approx 0.1943 - 0.0040(\text{age})$$

- Critical point (curve changes from upward to downward)

$$\text{age}^* = |\beta_1/(2\beta_2)| = |0.1943/(2^*-0.0020)|$$

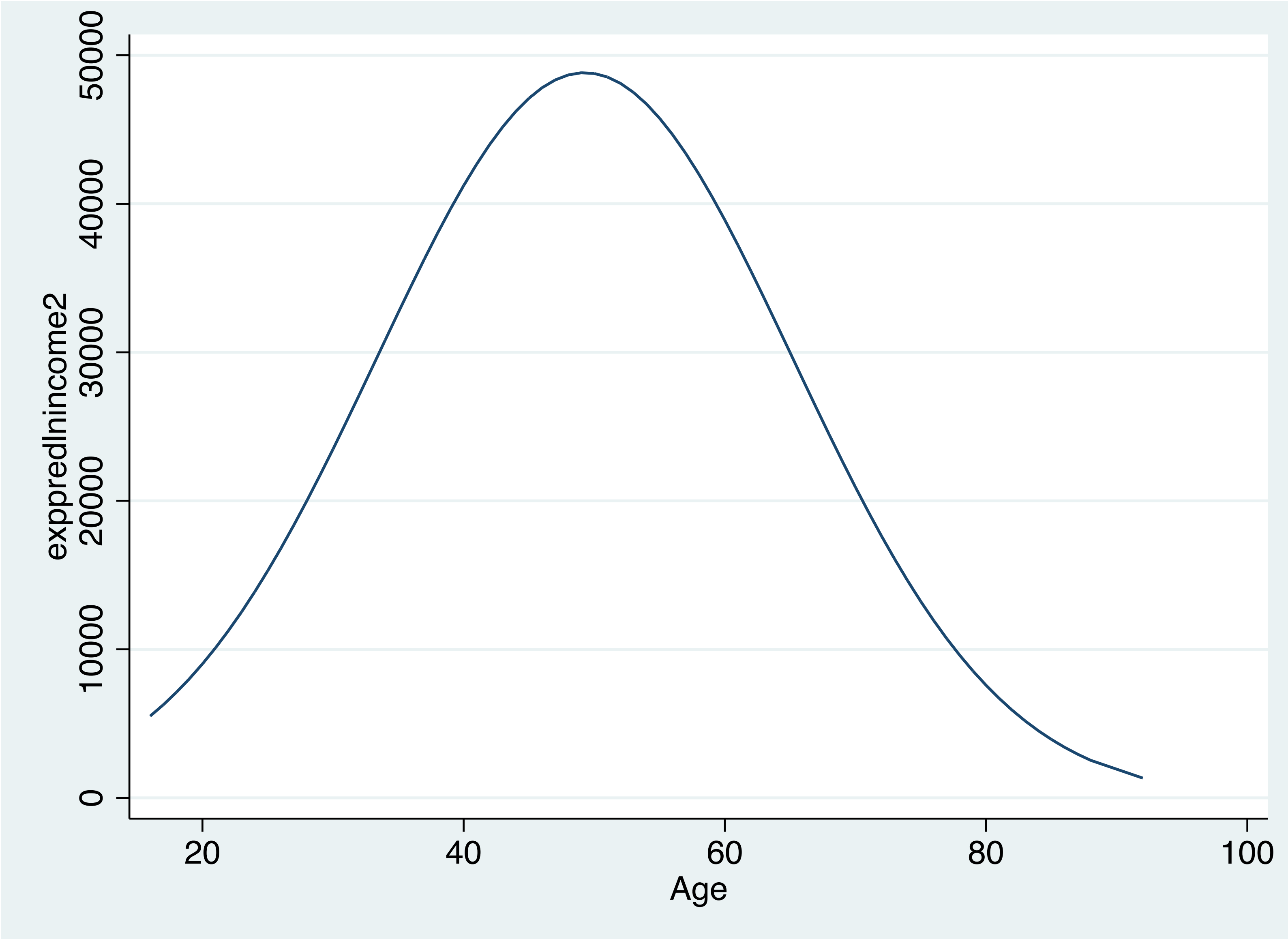
$$\text{age}^* = |-48.57| = 48.57$$

Predicted ln(income) by age, age^2



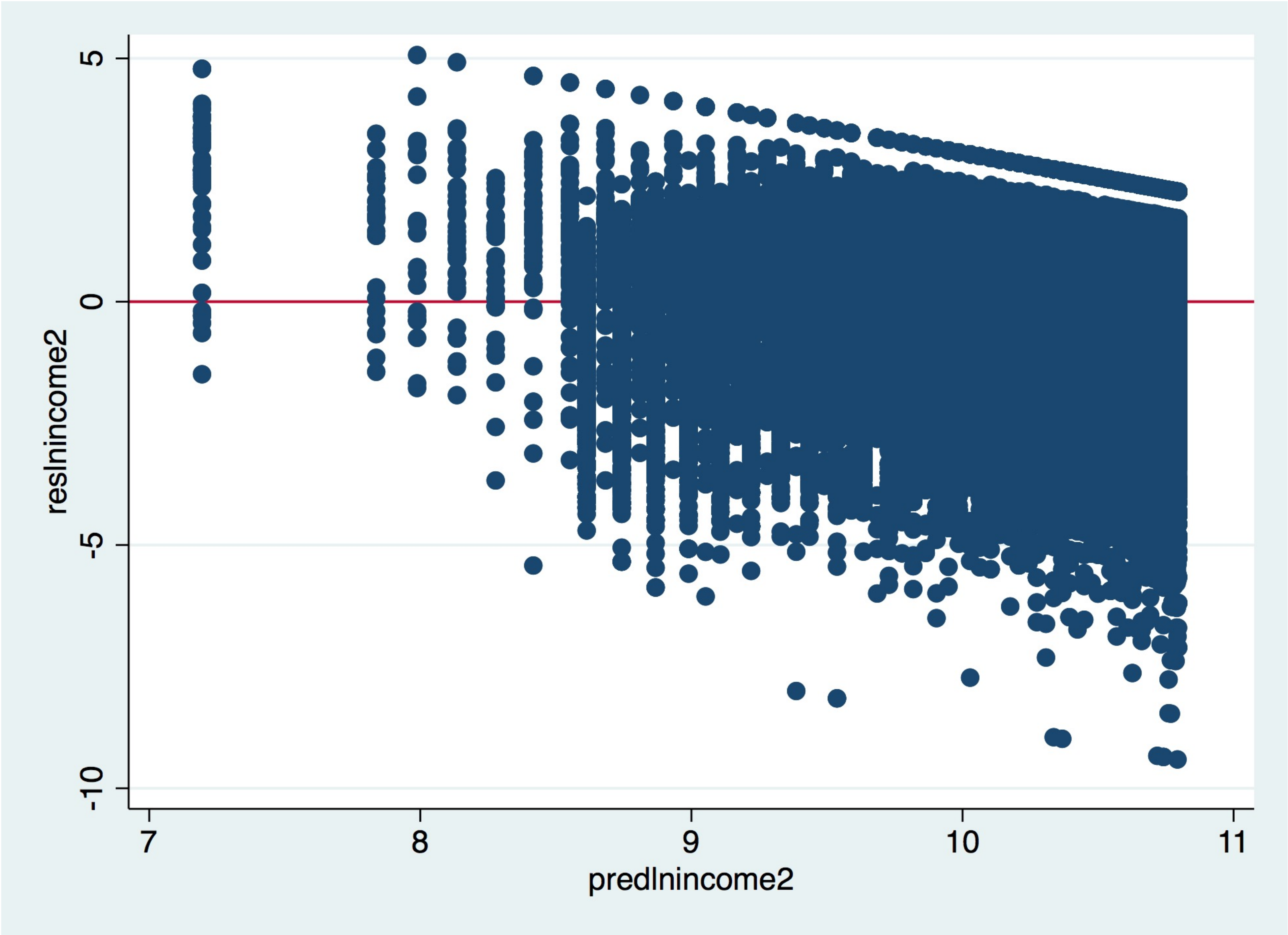
Source: 2018 American Community Survey.

Exponential of predicted ln(income) by age, age^2



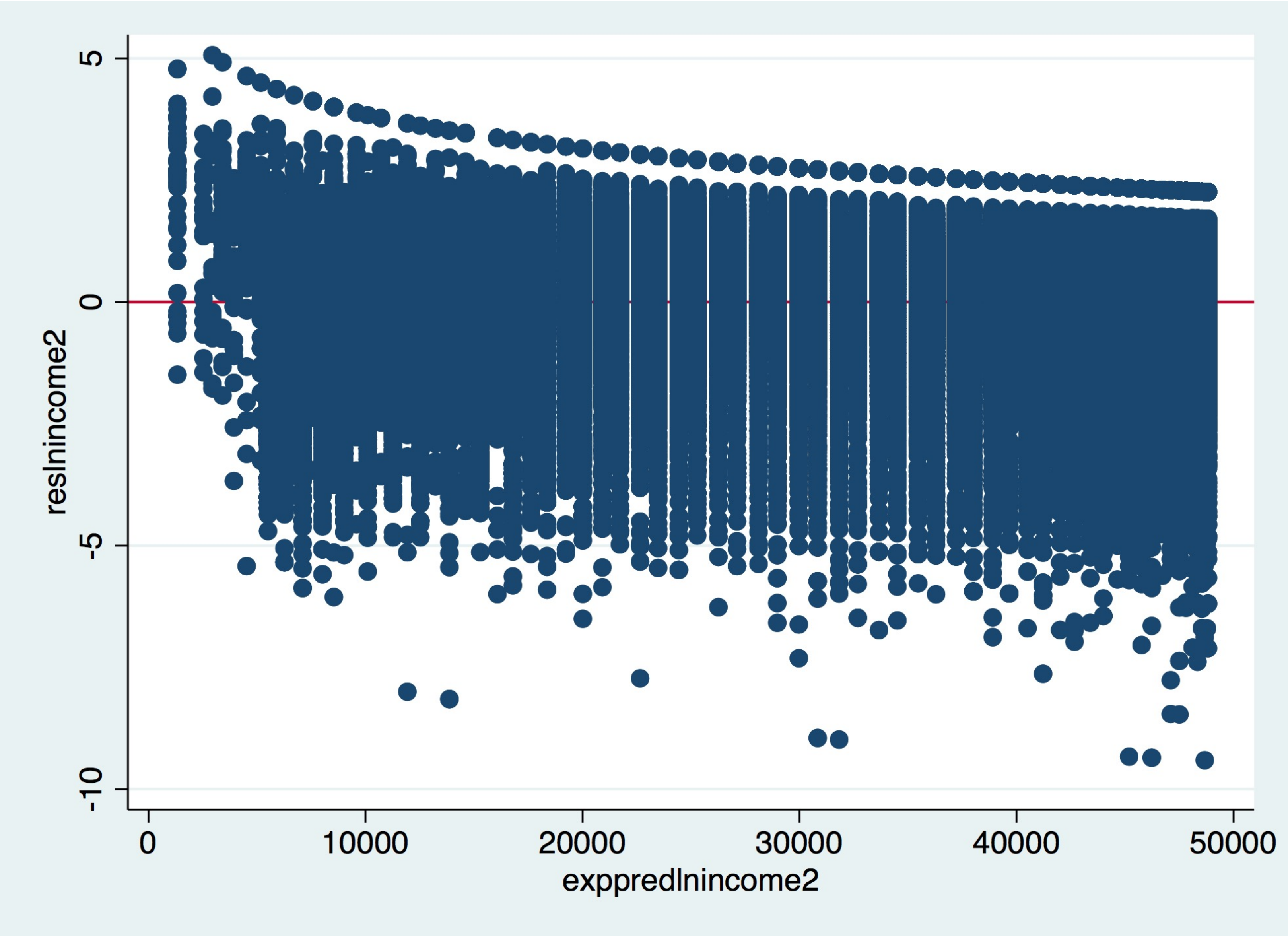
Source: 2018 American Community Survey.

Residuals: ln(income)=F(age,age^2)

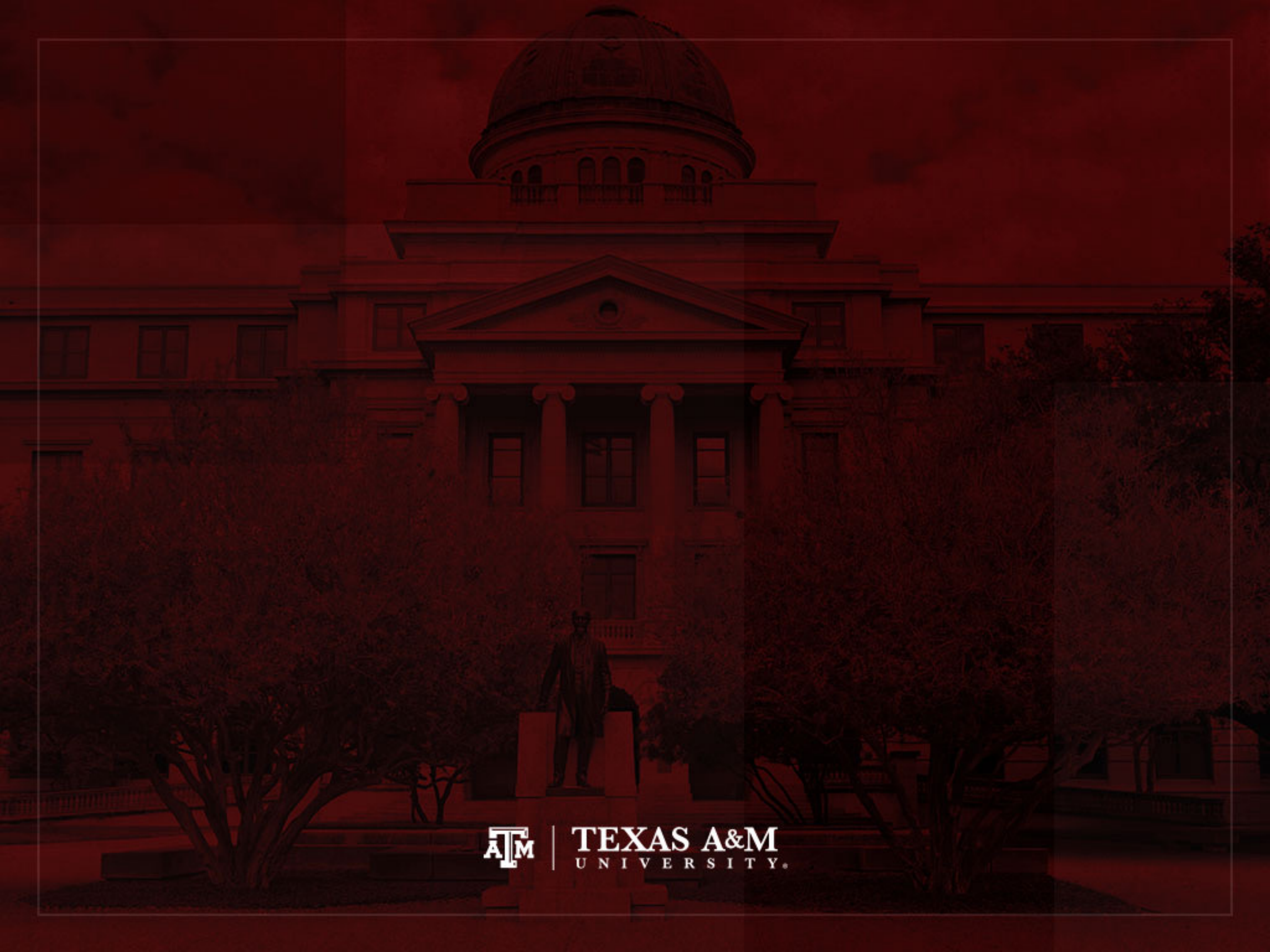


Source: 2018 American Community Survey.

Residuals: Exp. ln(income)=F(age, age^2)



Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Dummy variables

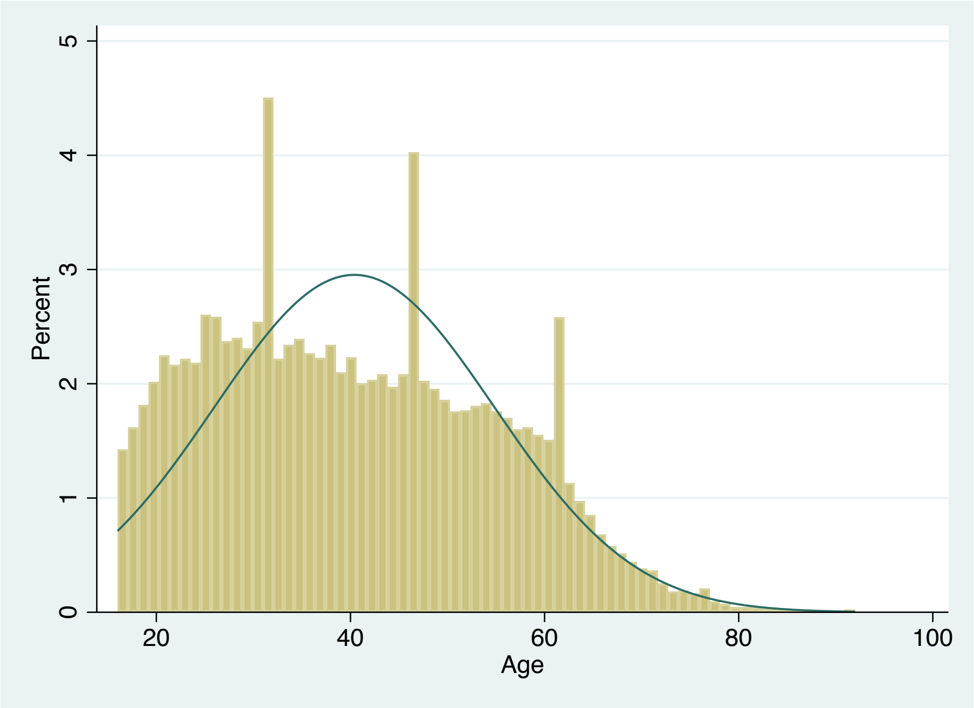
- Many variables that are important in social life are nominal-level variables
  - They cannot be included in a regression equation or correlational analysis (e.g., sex, race/ethnicity)
- We can create dummy variables
  - Two categories, one coded as 0 and the other as 1

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)
  - 16–19; 20–24; 25–34; 35–44; 45–54; 55–64; 65+

Source: 2018 American Community Survey.

Age in ordinal level

- Age has seven categories

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

agegr	min(age)	max(age)	N(age)
16	16	19	6,337
20	20	24	11,945
25	25	34	26,752
35	35	44	25,575
45	45	54	25,454
55	55	64	22,457
65	65	92	9,265

- Generate dummy variables for age...

Source: 2018 American Community Survey.

Dummies for age

- Generate dummy variables for age group

Age group	**Age 16–19**	**Age 20–24**	**Age 25–34**	**Age 35–44**	**Age 45–54**	**Age 55–64**	**Age 65+**
16–19	1	0	0	0	0	0	0
20–24	0	1	0	0	0	0	0
25–34	0	0	1	0	0	0	0
35–44	0	0	0	1	0	0	0
45–54	0	0	0	0	1	0	0
55–64	0	0	0	0	0	1	0
65+	0	0	0	0	0	0	1

Reference category

- Use the category with the largest sample size as the reference (25–34)

```
. tab agegr, m
```

agegr	Freq.	Percent	Cum.
16	6,337	4.96	4.96
20	11,945	9.35	14.31
25	26,752	20.94	35.24
35	25,575	20.01	55.26
45	25,454	19.92	75.18
55	22,457	17.57	92.75
65	9,265	7.25	100.00
Total	127,785	100.00	

- Or category with large sample and meaningful interpretation for your problem (age group with the highest average income: 45–54)

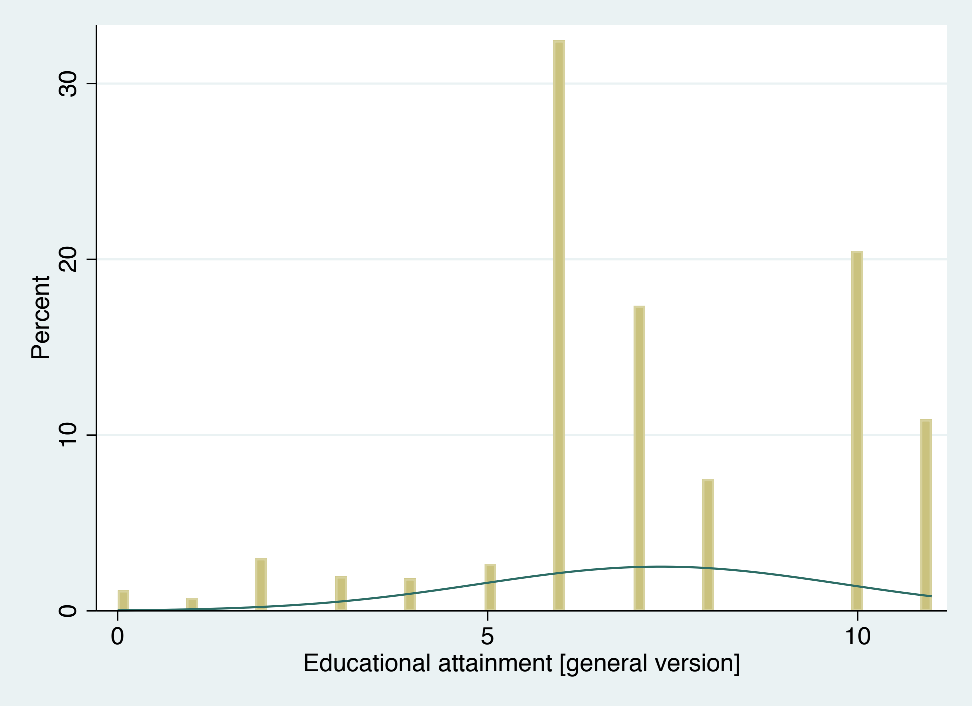
```
. table agegr, c(mean income)
```

agegr	mean(income)
16	6051.891
20	18397.36
25	42752.68
35	61426.85
45	67367.77
55	65728.8
65	50250.71

Source: 2018 American Community Survey.

Educational attainment

- Education does not have a normal distribution



- Generate education group variable (categorical)
  - Less than high school; high school; some college; college; graduate school

Source: 2018 American Community Survey.

Education in ordinal level

- Education has five categories

```
. tab educgr, m
```

educgr	Freq.	Percent	Cum.
Less than high school	12,719	9.95	9.95
High school	40,869	31.98	41.94
Some college	30,360	23.76	65.69
College	28,110	22.00	87.69
Graduate school	15,727	12.31	100.00
Total	127,785	100.00	

- Generate dummy variables for education...

Source: 2018 American Community Survey.

Dummies for education

- Generate dummy variables for education group

Education group	<High school	High school	Some College	College	Graduate school
Less than high school	1	0	0	0	0
High school	0	1	0	0	0
Some college	0	0	1	0	0
College	0	0	0	1	0
Graduate school	0	0	0	0	1

Reference group

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

```
. tab educgr, m
```

educgr	Freq.	Percent	Cum.
Less than high school	12,719	9.95	9.95
High school	40,869	31.98	41.94
Some college	30,360	23.76	65.69
College	28,110	22.00	87.69
Graduate school	15,727	12.31	100.00
Total	127,785	100.00	

Source: 2018 American Community Survey.

log income = F(age, education)

```
. svy: reg lnincome ib45.agegr ib2.educgr
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        212                Number of obs     =    127,785
Number of PSUs     =     79,499                Population size   = 13,849,398
                                               Design df         =     79,287
                                               F(  10,  79278)   =    2860.65
                                               Prob > F          =     0.0000
                                               R-squared         =     0.3129

------------------------------------------------------------------------------
                  |             Linearized
         lnincome |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------------+-----------------------------------------------------------
            agegr |
              16  |  -2.223012   .0227431   -97.74   0.000    -2.267588   -2.178435
              20  |  -1.151434   .0155642   -73.98   0.000     -1.18194   -1.120928
              25  |  -.3856507   .0104177   -37.02   0.000    -.4060693    -.365232
              35  |  -.0929935   .0104004    -8.94   0.000    -.1133781   -.0726089
              55  |   -.053233   .0111394    -4.78   0.000    -.0750662   -.0313998
              65  |  -.5928305   .0186409   -31.80   0.000    -.6293667   -.5562944
                  |
           educgr |
Less than high school  |  -.3066773   .0128821   -23.81   0.000    -.3319261   -.2814286
     Some college |   .1354166   .0097974    13.82   0.000     .1162138    .1546194
          College |   .5445375   .0101702    53.54   0.000      .524604     .564471
  Graduate school |   .8187744      .0121    67.67   0.000     .7950584    .8424904
                  |
            _cons |   10.41295   .0092523  1125.44   0.000     10.39482    10.43109
------------------------------------------------------------------------------
```

Source: 2018 American Community Survey.

Exponential of coefficients

```
. ***Automatically see exponential of coefficients
. svy: reg lnincome ib45.agegr ib2.educgr, eform(Exp. Coef.)
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        212           Number of obs     =    127,785
Number of PSUs     =     79,499           Population size   = 13,849,398
                                          Design df         =     79,287
                                          F(  10,  79278)   =    2860.65
                                          Prob > F          =     0.0000
                                          R-squared         =     0.3129

-----------------------------------------------------------------------------------------
                       |             Linearized
              lnincome | Exp. Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-----------------------+-----------------------------------------------------------------
                 agegr |
                    16 |   .1082825   .0024627   -97.74   0.000     .1035617    .1132186
                    20 |    .316183   .0049211   -73.98   0.000     .3066833     .325977
                    25 |    .680008   .0070841   -37.02   0.000      .666264    .6940356
                    35 |   .9111994   .0094768    -8.94   0.000      .892813    .9299645
                    55 |    .948159   .0105619    -4.78   0.000     .9276821     .969088
                    65 |   .5527605    .010304   -31.80   0.000     .5329292    .5733297
                       |
                educgr |
 Less than high school |    .735888   .0094797   -23.81   0.000     .7175404    .7547048
          Some college |   1.145014   .0112182    13.82   0.000     1.123236    1.167214
               College |   1.723811   .0175315    53.54   0.000      1.68979    1.758517
       Graduate school |   2.267719   .0274395    67.67   0.000      2.21457    2.322143
                       |
                 _cons |   33288.07   307.9918  1125.44   0.000     32689.85    33897.24
-----------------------------------------------------------------------------------------
```

Source: 2018 American Community Survey.

Interpretation of age

(log of income with dummies as independent variables)

- 45–54 age group is reference category for **age**
- Coefficient for 16–19 age group equals –2.2230
  - exp(β_1) times
    - People between 16–19 years of age have on average earnings **0.1083 times** the earnings of people between 45–54 years of age, controlling for the other independent variables
  - 100*[exp(β_1)–1] percent
    - People between 16–19 years of age have on average earnings **89.17% lower** than earnings of people between 45–54 years of age, controlling for the other independent variables
  - 100*β_1 percent: result is not good because β_1>0.3
    - People between 16–19 years of age have on average earnings **approximately 222.30% lower** than earnings of people between 45–54 years of age, controlling for the other independent variables

Interpretation of education

(log of income with dummies as independent variables)

- High school is reference category for **<u>education</u>**
- Coefficient for college equals 0.5445
  - $\exp(\beta_1)$ times
    - People with college degree have on average earnings **<u>1.7238 times higher</u>** than earnings of high school graduates, controlling for the other independent variables
  - $100*[\exp(\beta_1)–1]$ percent
    - People with college degree have on average earnings **<u>72.38% higher</u>** than earnings of high school graduates, controlling for the other independent variables
  - $100*\beta_1$ percent: result is not good because $\beta_1>0.3$
    - People with college degree have on average earnings **<u>approximately 54.45% higher</u>** than earnings of high school graduates, controlling for the other independent variables

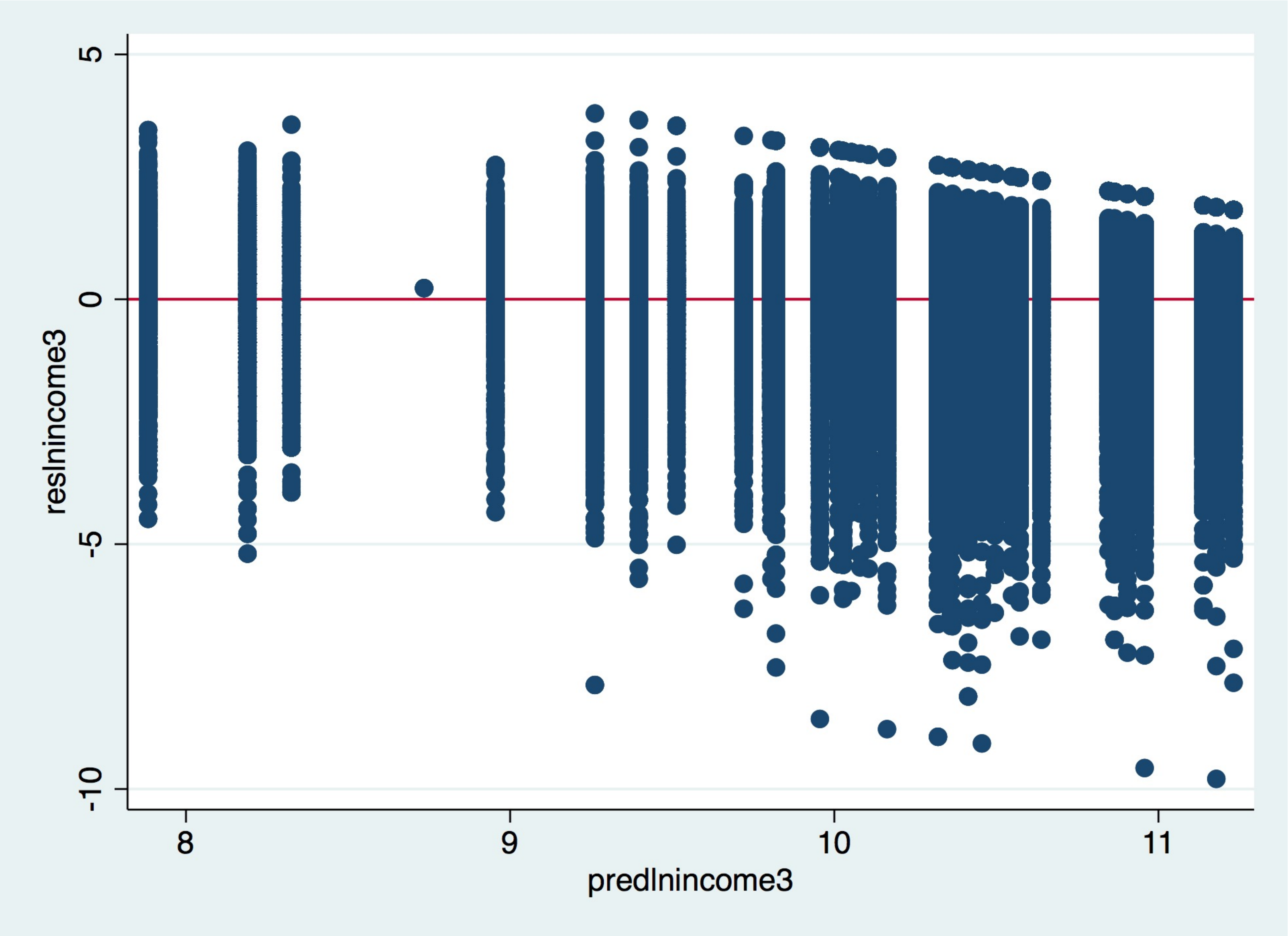
Standardized coefficients

```
. reg lnincome ib45.agegr ib2.educgr [pweight=perwt], beta
(sum of wgt is 13,849,398)

Linear regression                               Number of obs     =    127,785
                                                F(10, 127774)     =    3037.91
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3129
                                                Root MSE          =     1.0223
```

lnincome	Coef.	Robust Std. Err.	t	P>|t|	Beta
agegr					
16	-2.223012	.022166	-100.29	0.000	-.3875416
20	-1.151434	.0148555	-77.51	0.000	-.290206
25	-.3856507	.0103423	-37.29	0.000	-.1333188
35	-.0929935	.0103849	-8.95	0.000	-.0310561
55	-.053233	.0110966	-4.80	0.000	-.0151658
65	-.5928305	.018443	-32.14	0.000	-.107231
educgr					
Less than high school	-.3066773	.0125263	-24.48	0.000	-.0788327
Some college	.1354166	.0096013	14.10	0.000	.047455
College	.5445375	.0100048	54.43	0.000	.1781623
Graduate school	.8187744	.0120082	68.18	0.000	.2068187
_cons	10.41295	.0091286	1140.69	0.000	.

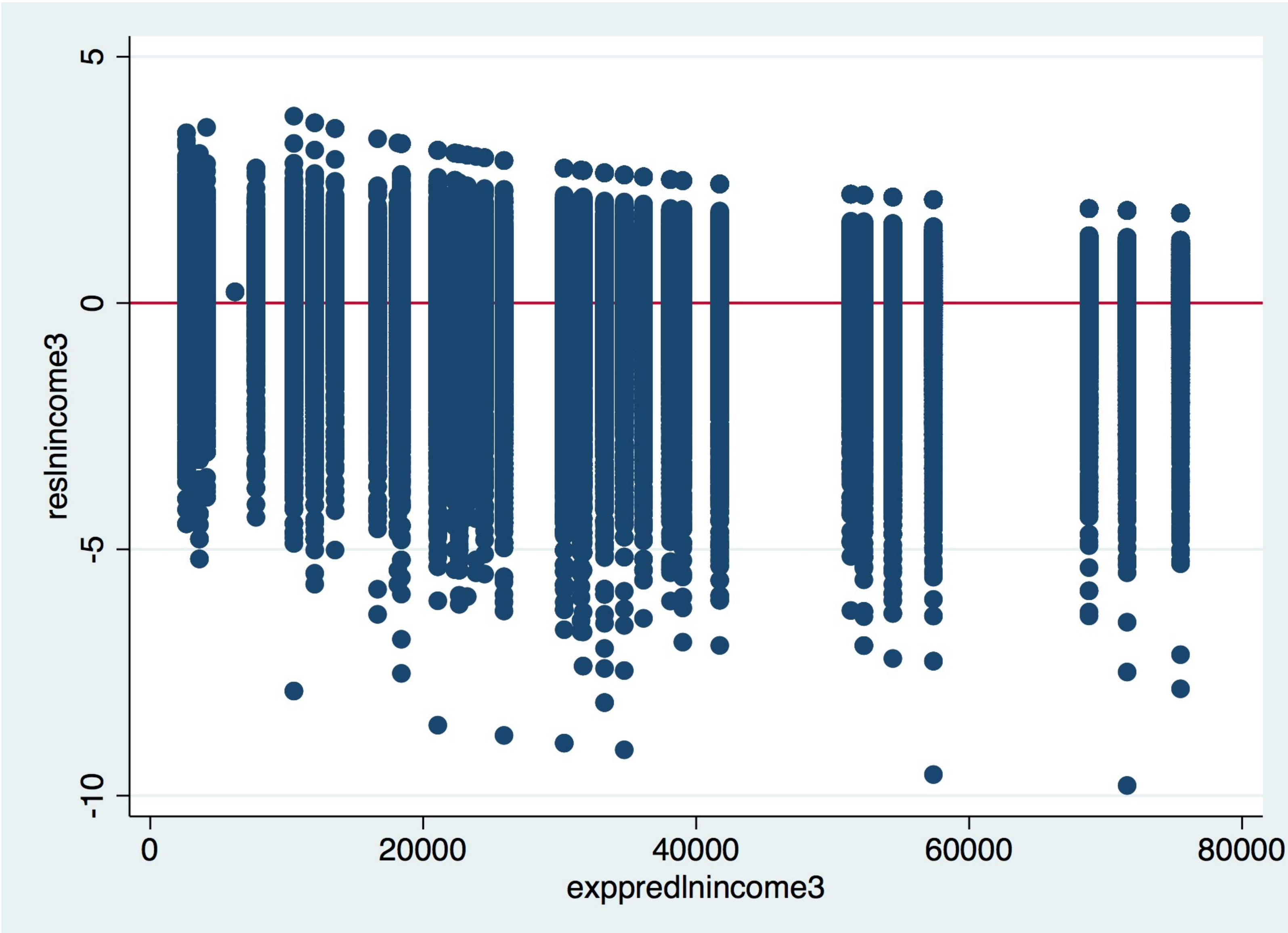
Source: 2018 American Community Survey.

Residuals: ln(income)=F(age group, educ. group)

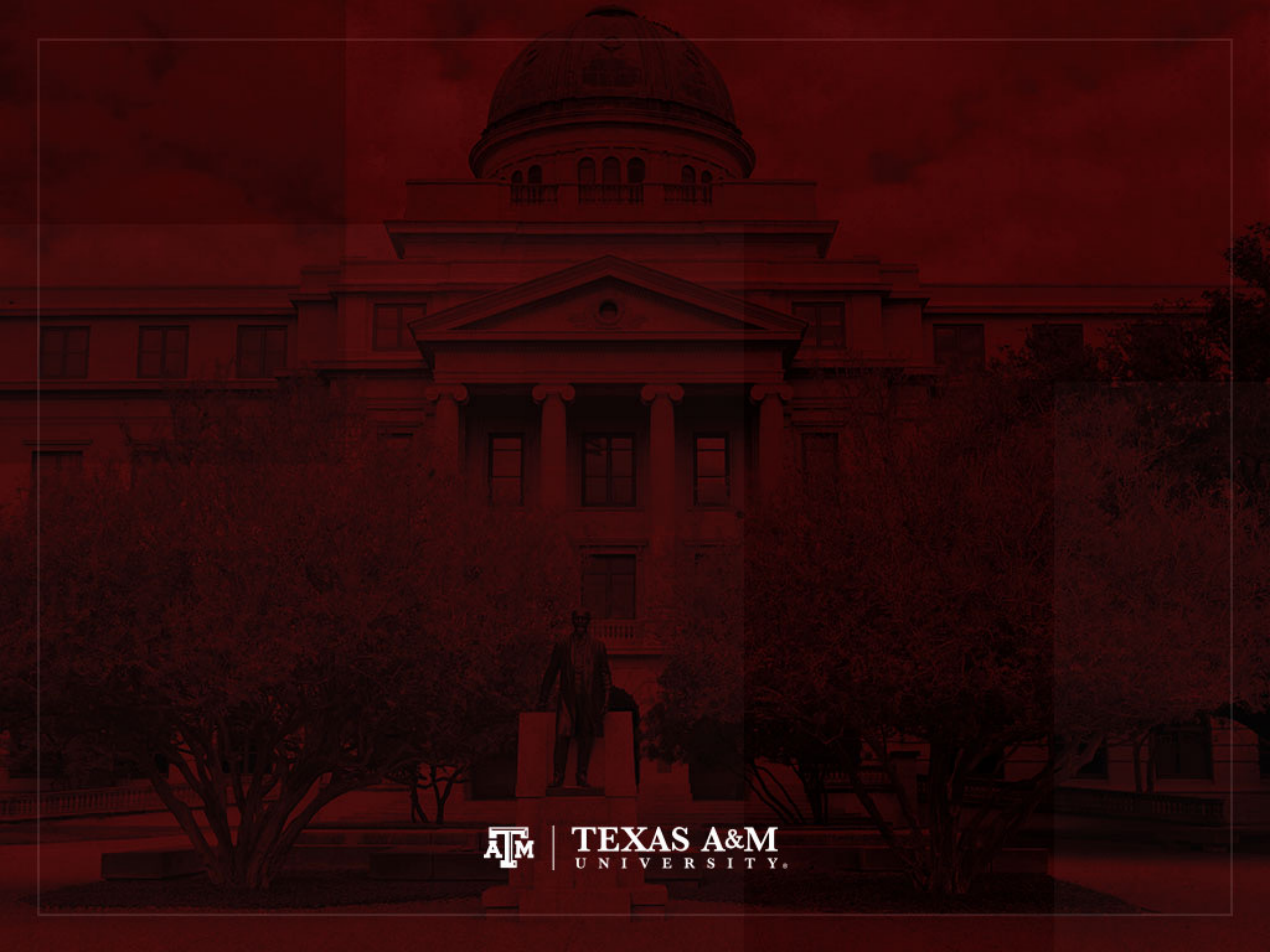


Source: 2018 American Community Survey.

Residuals: Exp. ln(income)=F(age group, educ. group)



Source: 2018 American Community Survey.


TEXAS A&M
UNIVERSITY

Full OLS model

- Dependent variable
  - Natural logarithm of income
- Independent variables
  - **Sex:** female; male (reference)
  - **Age group:** 16–19; 20–24; 25–34; 35–44; 45–54 (reference); 55–64; 65+
  - **Education group:** less than high school, high school (reference), some college, college, graduate school
  - **Race/ethnicity:** White (reference); African American; Hispanic; Asian; Native American; Other races
  - **Marital status:** married (reference); separated, divorced, widowed; never married
  - **Migration status:** non-migrant (reference); internal migrant; international migrant

Source: 2018 American Community Survey.

Command in Stata

```
. svy: reg lnincome i.female ib45.agegr ib2.educgr i.raceth i.marital i.migrant
(running regress on estimation sample)

Survey: Linear regression

Number of strata   =        212                 Number of obs     =     127,785
Number of PSUs     =     79,499                 Population size   =  13,849,398
                                                Design df         =      79,287
                                                F(  20,   79268)  =     1818.83
                                                Prob > F          =      0.0000
                                                R-squared         =      0.3577
```

Source: 2018 American Community Survey.

Coefficients from OLS regression for natural logarithm of income

lnincome	Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf.	Interval]
female						
Female	-.4374635	.0070675	-61.90	0.000	-.4513158	-.4236111
agegr						
16-19	-1.995369	.0241877	-82.50	0.000	-2.042777	-1.947961
20-24	-.9592868	.0168846	-56.81	0.000	-.9923806	-.926193
25-34	-.2920554	.0106538	-27.41	0.000	-.3129368	-.271174
35-44	-.0705981	.0100164	-7.05	0.000	-.0902301	-.0509661
55-64	-.0751899	.0107209	-7.01	0.000	-.0962027	-.0541771
65-100	-.6377643	.0183047	-34.84	0.000	-.6736413	-.6018873
educgr						
Less than high school	-.3148089	.01281	-24.58	0.000	-.3399165	-.2897013
Some college	.1565395	.0096239	16.27	0.000	.1376767	.1754023
College	.5426535	.0101186	53.63	0.000	.5228211	.562486
Graduate school	.8081078	.0122256	66.10	0.000	.7841457	.8320698
raceth						
African American	-.172703	.012575	-13.73	0.000	-.19735	-.148056
Hispanic	-.1285316	.0085376	-15.05	0.000	-.1452652	-.111798
Asian	-.1583612	.0172829	-9.16	0.000	-.1922356	-.1244867
Native American	-.071535	.0555021	-1.29	0.197	-.1803187	.0372488
Ohter races	-.1193284	.0302909	-3.94	0.000	-.1786982	-.0599585
marital						
Separated, divorced, wid..	-.1364001	.0101838	-13.39	0.000	-.1563603	-.11644
Never married	-.2696217	.009485	-28.43	0.000	-.2882122	-.2510312
migrant						
Internal migrant	-.1211724	.0160131	-7.57	0.000	-.1525579	-.0897869
International migrant	-.4936644	.0683904	-7.22	0.000	-.6277092	-.3596197
_cons	10.76426	.0105691	1018.47	0.000	10.74355	10.78498

Source: 2018 American Community Survey.

Exponential of coefficients from OLS regression for natural logarithm of income

lnincome	Exp. Coef.	Linearized Std. Err.	t	P>|t|	[95% Conf. Interval]	
female						
Female	.6456721	.0045633	-61.90	0.000	.6367897	.6546784
agegr						
16-19	.1359635	.0032886	-82.50	0.000	.1296682	.1425645
20-24	.3831661	.0064696	-56.81	0.000	.3706932	.3960586
25-34	.7467272	.0079555	-27.41	0.000	.7312961	.7624838
35-44	.9318363	.0093336	-7.05	0.000	.9137209	.9503108
55-64	.9275673	.0099443	-7.01	0.000	.9082799	.9472643
65-100	.5284726	.0096735	-34.84	0.000	.5098487	.5477769
educgr						
Less than high school	.7299283	.0093504	-24.58	0.000	.7118298	.7484871
Some college	1.169457	.0112548	16.27	0.000	1.147604	1.191726
College	1.720566	.0174098	53.63	0.000	1.686779	1.75503
Graduate school	2.243658	.02743	66.10	0.000	2.190535	2.29807
raceth						
African American	.8413875	.0105805	-13.73	0.000	.8209033	.8623828
Hispanic	.8793858	.0075078	-15.05	0.000	.8647929	.8942249
Asian	.8535415	.0147517	-9.16	0.000	.8251124	.88295
Native American	.9309637	.0516704	-1.29	0.197	.835004	1.037951
Ohter races	.8875163	.0268836	-3.94	0.000	.8363582	.9418036
marital						
Separated, divorced, widowed	.8724934	.0088853	-13.39	0.000	.855251	.8900835
Never married	.7636683	.0072434	-28.43	0.000	.7496025	.7779981
migrant						
Internal migrant	.8858812	.0141857	-7.57	0.000	.8585092	.9141259
International migrant	.6103856	.0417445	-7.22	0.000	.5338133	.6979417
_cons	47299.9	499.9155	1018.47	0.000	46330.15	48289.95

Source: 2018 American Community Survey.

Edited table

Table 1. Coefficients and standard errors estimated with ordinary least squares models for the logarithm of wage and salary income as the dependent variable, United States, 2018

Independent variables	Model 1	Model 2	Model 3	Model 4	Model 4 Standardized coefficients
Constant	10.61***	10.70***	10.76***	10.76***	
	(0.00961)	(0.0106)	(0.0106)	(0.0106)	
Sex					
Male	ref.	ref.	ref.	ref.	ref.
Female	-0.449***	-0.444***	-0.436***	-0.437***	-0.177
	(0.00700)	(0.00700)	(0.00707)	(0.00707)	
Age groups					
16-19	-2.195***	-2.204***	-2.007***	-1.995***	-0.348
	(0.0226)	(0.0228)	(0.0241)	(0.0242)	
20-24	-1.154***	-1.142***	-0.973***	-0.959***	-0.242
	(0.0155)	(0.0155)	(0.0168)	(0.0169)	
25-34	-0.396***	-0.385***	-0.302***	-0.292***	-0.101
	(0.0103)	(0.0102)	(0.0106)	(0.0107)	
35-44	-0.100***	-0.0921***	-0.0734***	-0.0706***	-0.0236
	(0.0101)	(0.0101)	(0.0100)	(0.0100)	
45-54	ref.	ref.	ref.	ref.	ref.
55-64	-0.0545***	-0.0698***	-0.0737***	-0.0752***	-0.0214
	(0.0108)	(0.0108)	(0.0107)	(0.0107)	
65+	-0.604***	-0.631***	-0.634***	-0.638***	-0.115
	(0.0183)	(0.0183)	(0.0183)	(0.0183)	

Note: Coefficients and standard errors were generated with the complex survey design of the American Community Survey. The standardized coefficients were generated with sample weights. Standard errors are reported in parentheses. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Source: 2018 American Community Survey.

Continue...

Edited table

Table 1. Coefficients and standard errors estimated with ordinary least squares models for the logarithm of wage and salary income as the dependent variable, United States, 2018

Independent variables	Model 1	Model 2	Model 3	Model 4	Model 4 Standardized coefficients
Education groups					
Less than high school	-0.336***	-0.311***	-0.314***	-0.315***	-0.0809
	(0.0125)	(0.0129)	(0.0128)	(0.0128)	
High school	ref.	ref.	ref.	ref.	ref.
Some college	0.165***	0.156***	0.157***	0.157***	0.0549
	(0.00965)	(0.00971)	(0.00963)	(0.00962)	
College	0.579***	0.551***	0.539***	0.543***	0.178
	(0.0100)	(0.0102)	(0.0101)	(0.0101)	
Graduate school	0.848***	0.826***	0.803***	0.808***	0.204
	(0.0119)	(0.0123)	(0.0122)	(0.0122)	

Note: Coefficients and standard errors were generated with the complex survey design of the American Community Survey. The standardized coefficients were generated with sample weights. Standard errors are reported in parentheses. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Source: 2018 American Community Survey.

Continue...

Edited table

Table 1. Coefficients and standard errors estimated with ordinary least squares models for the logarithm of wage and salary income as the dependent variable, United States, 2018

Independent variables	Model 1	Model 2	Model 3	Model 4	Model 4 Standardized coefficients
Race/ethnicity					
White		ref.	ref.	ref.	ref.
African American		-0.211***	-0.172***	-0.173***	-0.0461
		(0.0126)	(0.0126)	(0.0126)	
Hispanic		-0.132***	-0.125***	-0.129***	-0.0503
		(0.00860)	(0.00853)	(0.00854)	
Asian		-0.153***	-0.166***	-0.158***	-0.0288
		(0.0176)	(0.0175)	(0.0173)	
Native American		-0.0988*	-0.0758	-0.0715	-0.00272
		(0.0540)	(0.0549)	(0.0555)	
Other races		-0.140***	-0.124***	-0.119***	-0.0123
		(0.0302)	(0.0301)	(0.0303)	

Note: Coefficients and standard errors were generated with the complex survey design of the American Community Survey. The standardized coefficients were generated with sample weights. Standard errors are reported in parentheses. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Source: 2018 American Community Survey.

Continue...

Edited table

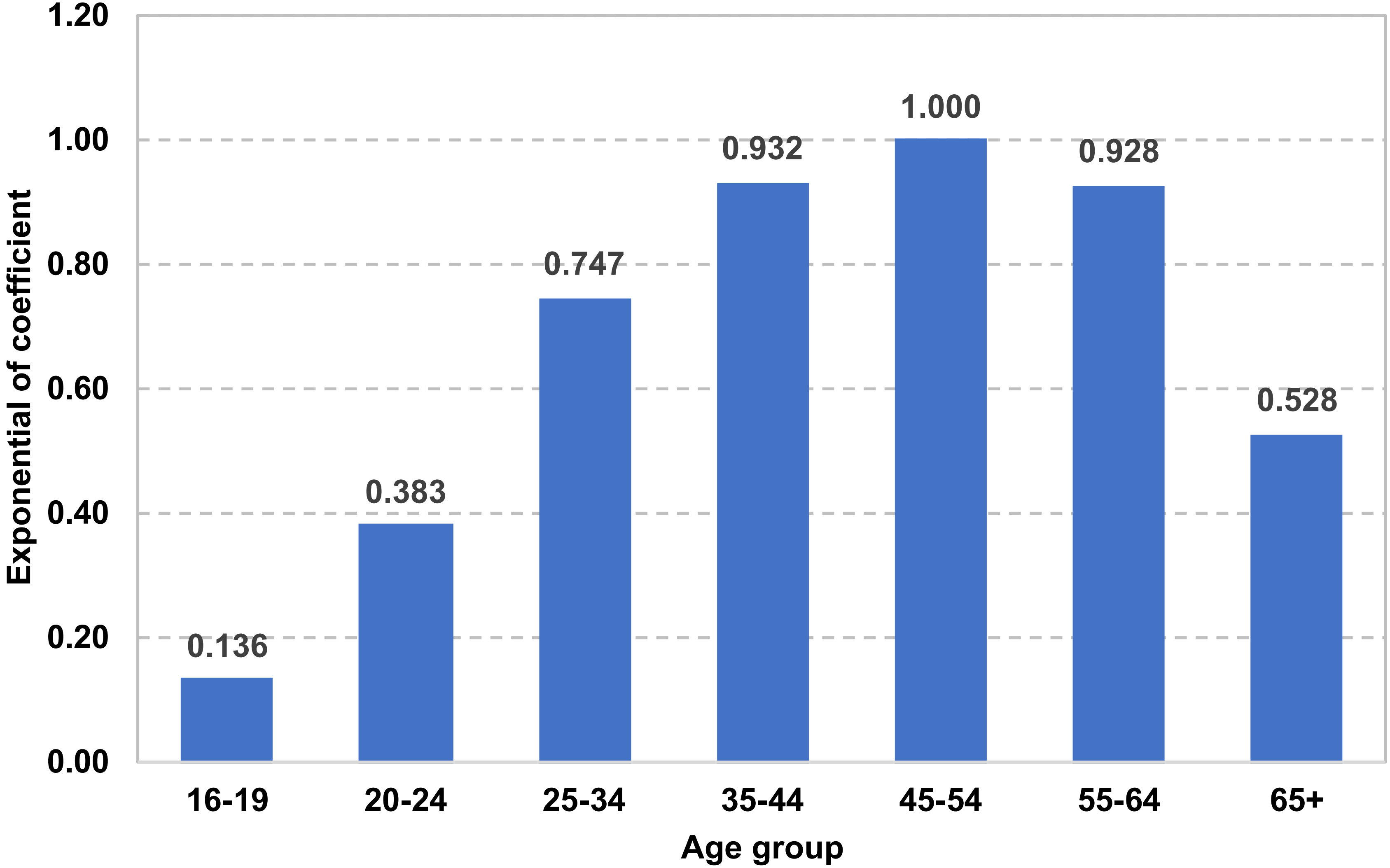
Table 1. Coefficients and standard errors estimated with ordinary least squares models for the logarithm of wage and salary income as the dependent variable, United States, 2018

Independent variables	Model 1	Model 2	Model 3	Model 4	Model 4 Standardized coefficients
Marital status					
Married			ref.	ref.	ref.
Separated, divorced, widowed			-0.139*** (0.0102)	-0.136*** (0.0102)	-0.0398
Never married			-0.270*** (0.00950)	-0.270*** (0.00948)	-0.104
Migration status					
Non-migrant				ref.	ref.
Internal migrant				-0.121*** (0.0160)	-0.0242
International migrant				-0.494*** (0.0684)	-0.0287
R^2	0.346	0.349	0.356	0.358	0.358
Observations	127,785	127,785	127,785	127,785	127,785

Note: Coefficients and standard errors were generated with the complex survey design of the American Community Survey. The standardized coefficients were generated with sample weights. Standard errors are reported in parentheses. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Source: 2018 American Community Survey.

Exponential of age group coefficients

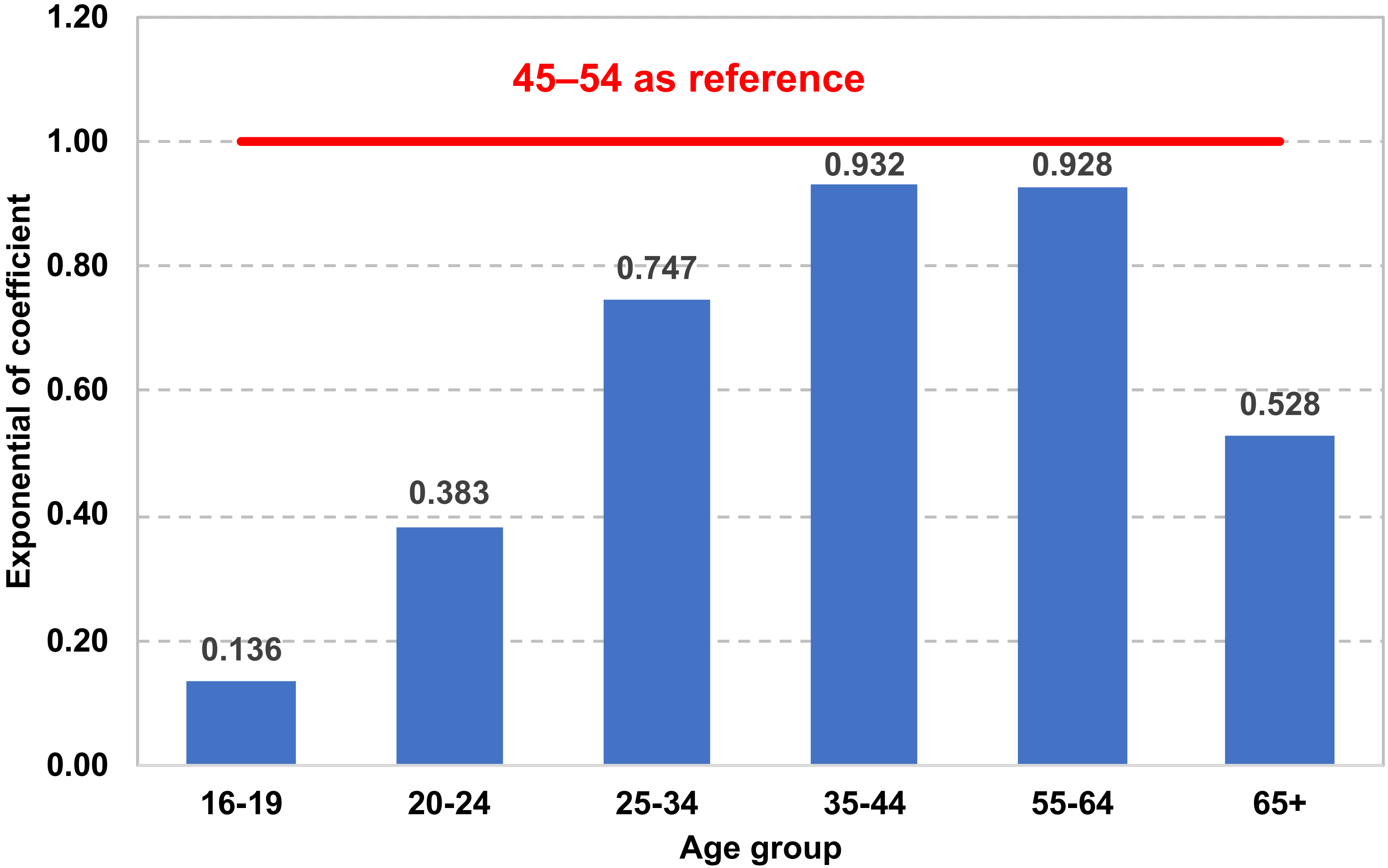
(Example of how to show regression results in conferences. Edited in Excel)



Source: 2018 American Community Survey.

Exponential of age group coefficients

(Example of how to show regression results in conferences. Edited in Excel.)

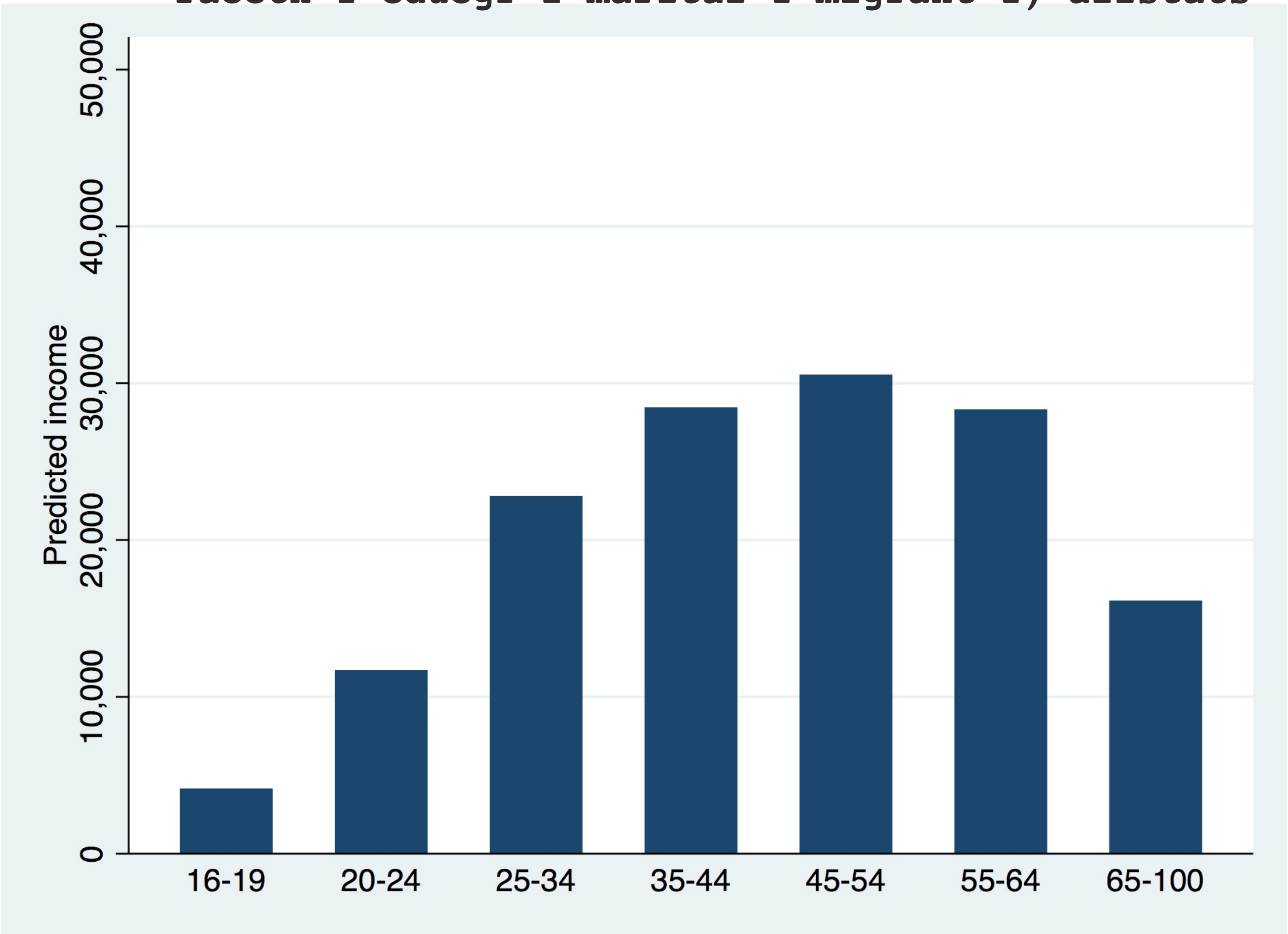


Source: 2018 American Community Survey.

Predicted female income by age

(Using “mgen” command within SPost13 package by Long and Freese, 2014)

```
mgen, stub(F) at(agegr=(16 20 25 35 45 55 65) female=1 ///
        raceth=1 educgr=2 marital=1 migrant=1) allstats
```

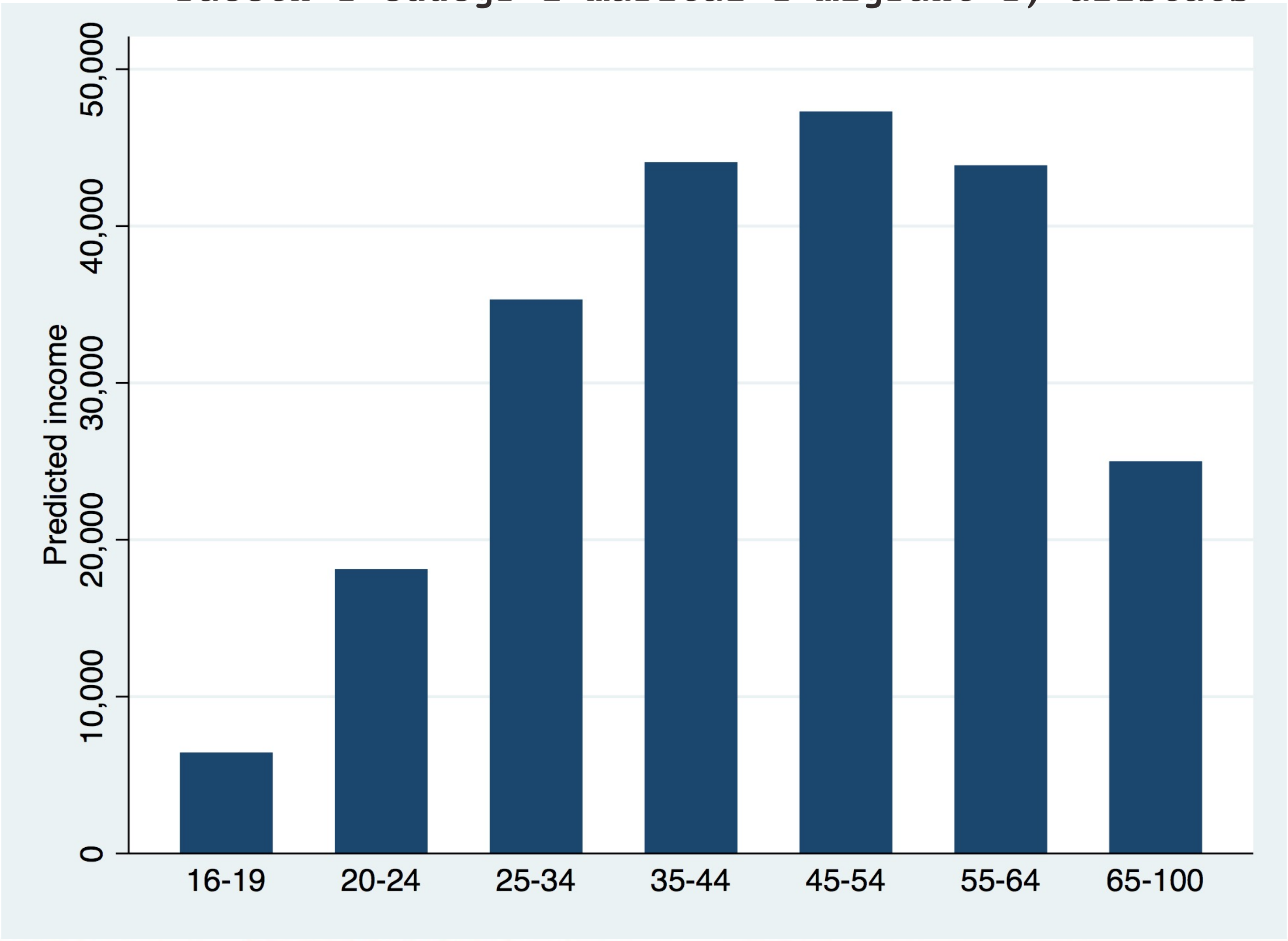


Source: 2018 American Community Survey.

Predicted male income by age

(Using "mgen" command within SPost13 package by Long and Freese, 2014)

```
mgen, stub(M) at(agegr=(16 20 25 35 45 55 65) female=0 ///
        raceth=1 educgr=2 marital=1 migrant=1) allstats
```

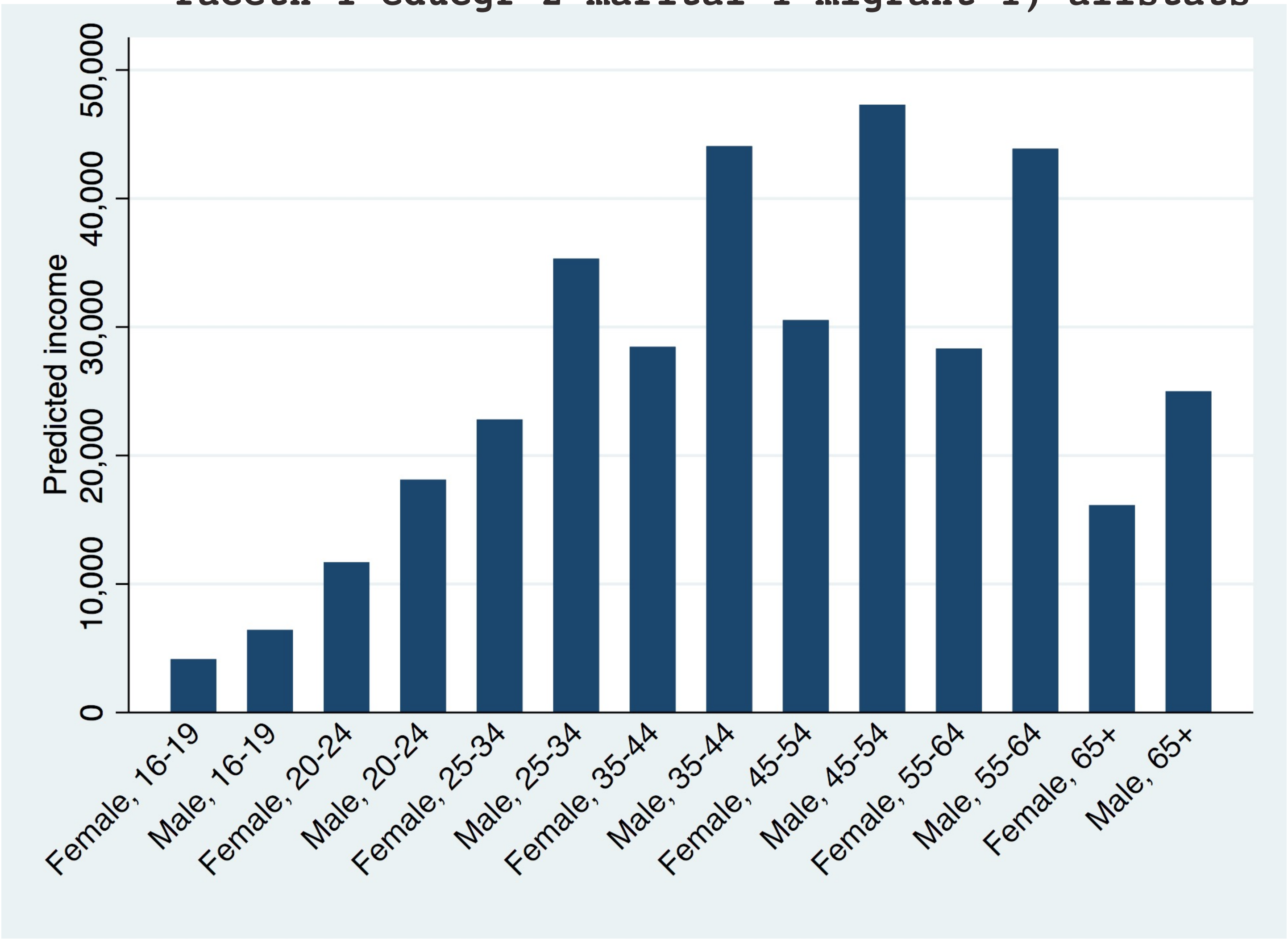


Source: 2018 American Community Survey.

Predicted income by age and sex

(Using “mgen” command within SPost13 package by Long and Freese, 2014)

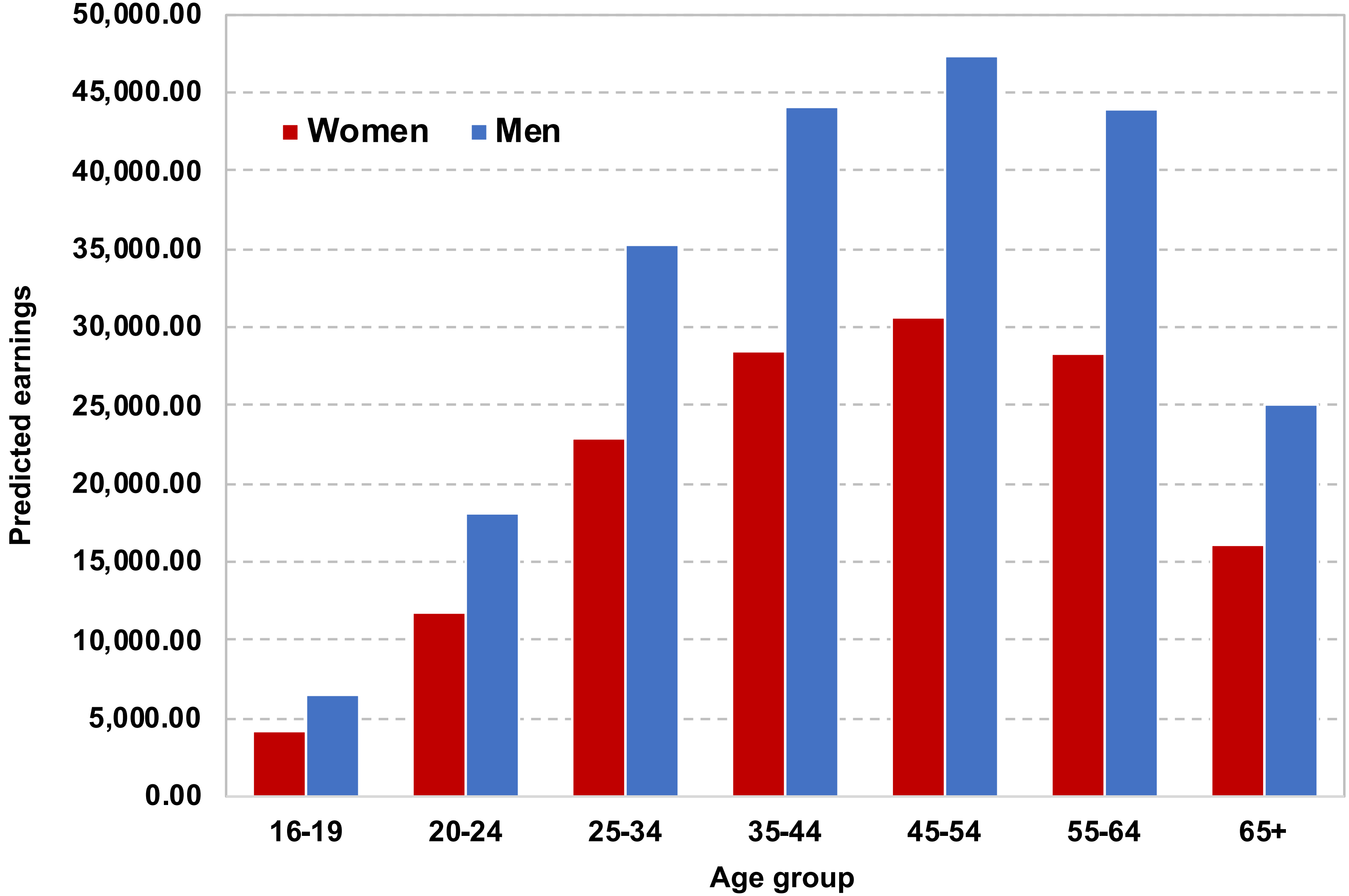
```
mgen, stub(A) at(agegr=(16 20 25 35 45 55 65) female=(0 1) ///
       raceth=1 educgr=2 marital=1 migrant=1) allstats
```



Source: 2018 American Community Survey.

Predicted income by age and sex

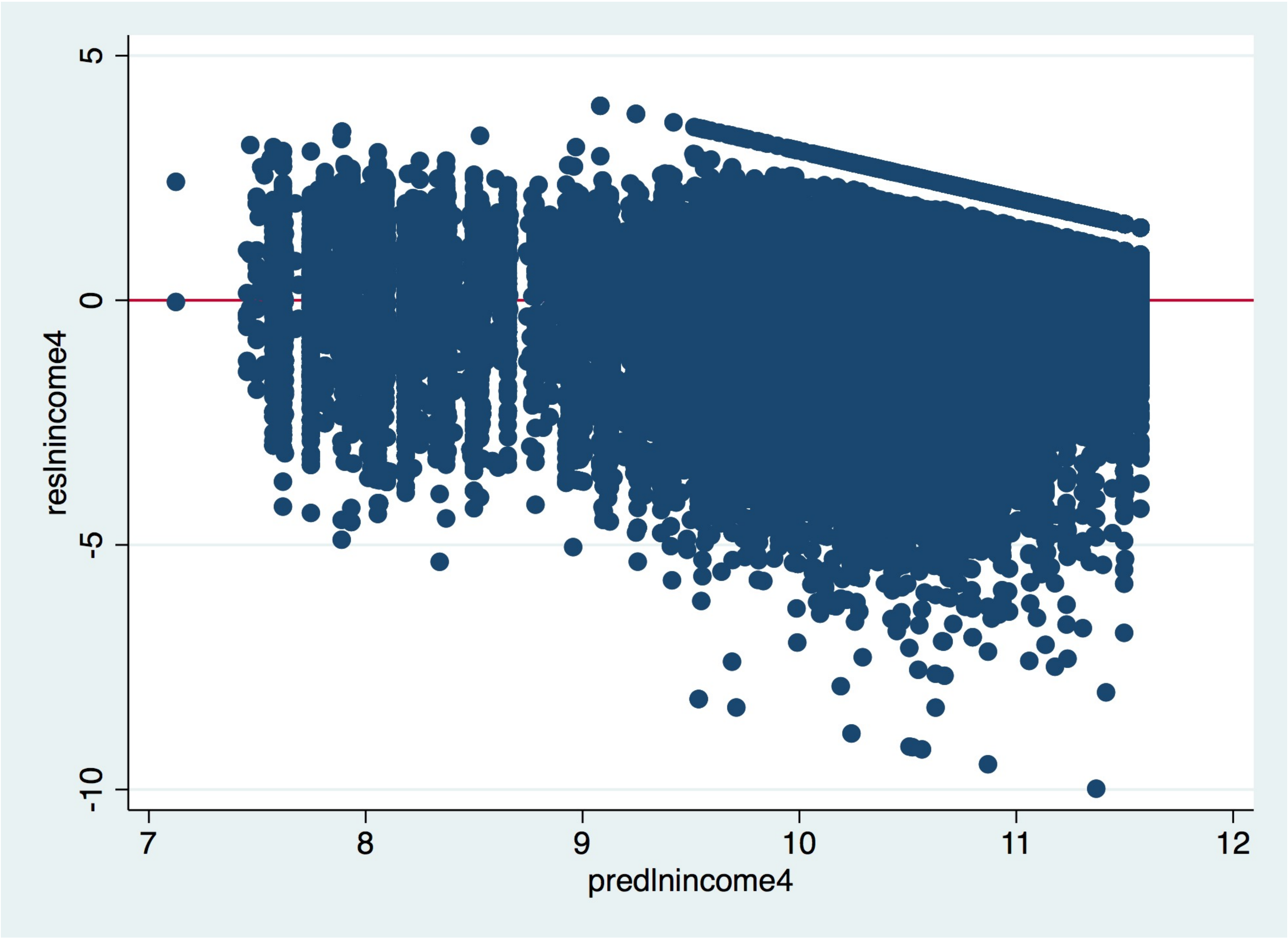
(Using "mgen" command within SPost13 package by Long and Freese, 2014. Edited in Excel.)



Source: 2018 American Community Survey.

Residuals:

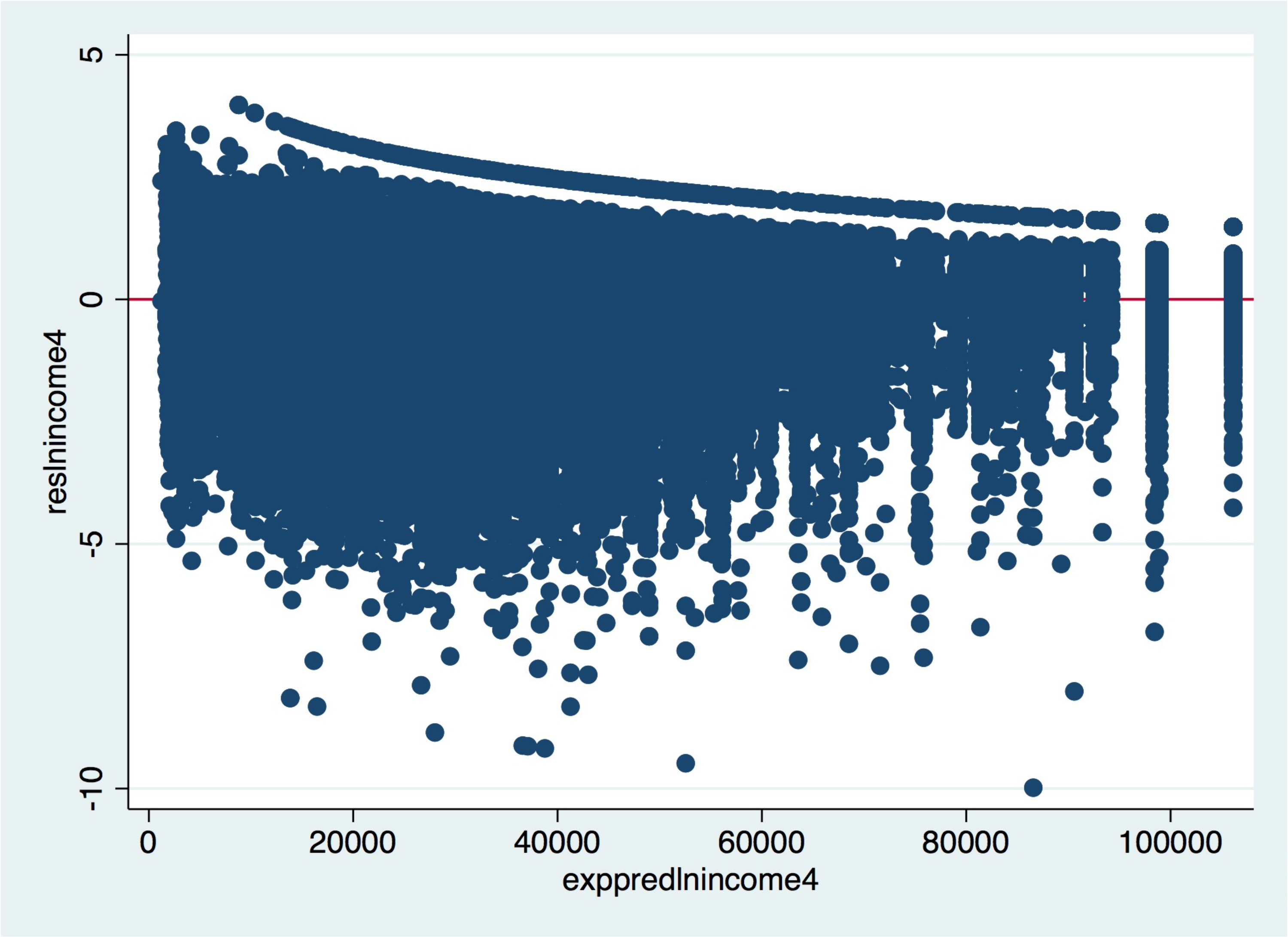
ln(income)=F(sex,age,educ,race/ethnicity,marital,migrant)



Source: 2018 American Community Survey.

Residuals:

Exp.ln(income)=F(sex,age,educ,race/ethnicity,marital,migrant)



Source: 2018 American Community Survey.

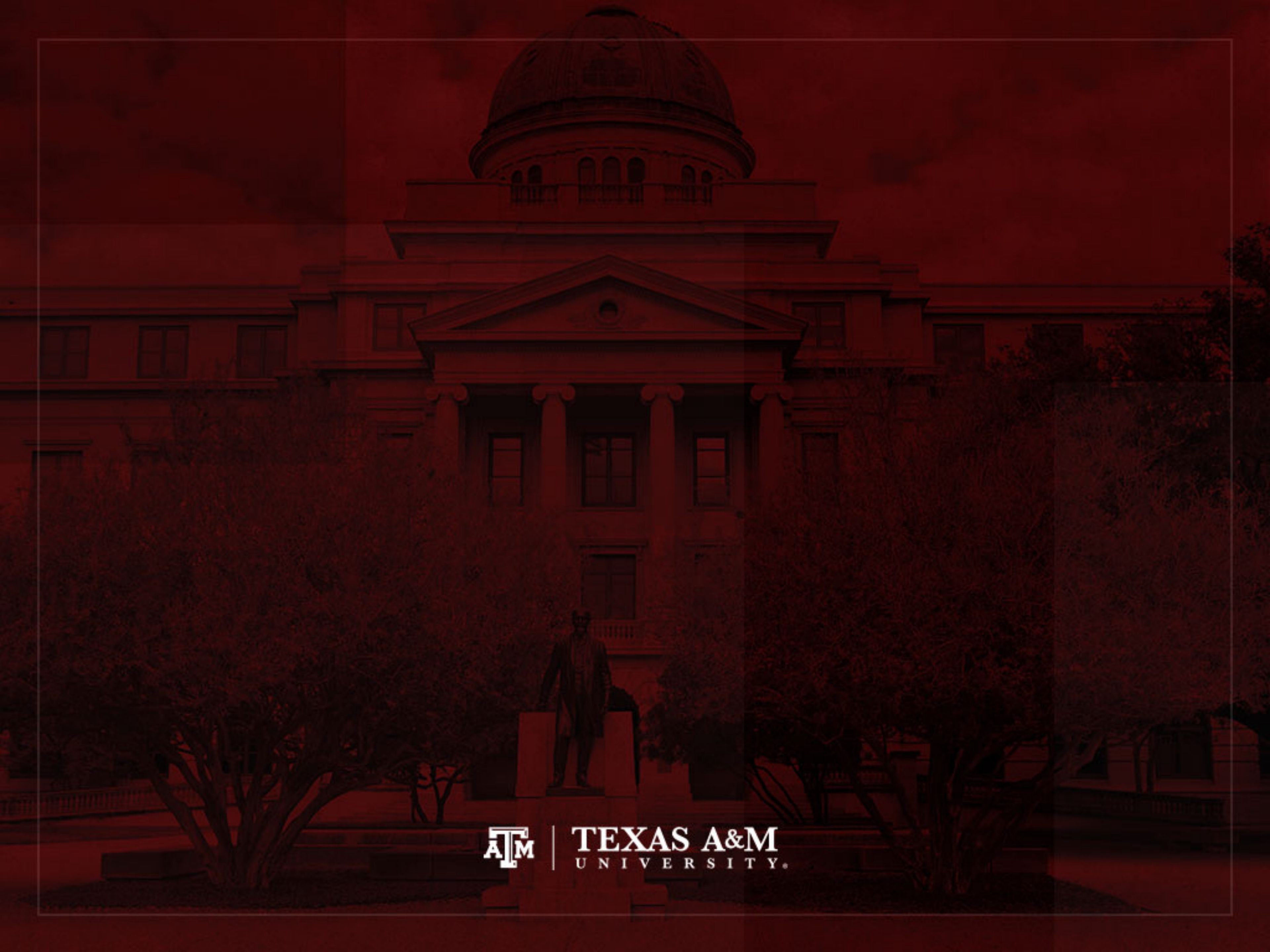
Stata practice time

- Additional material on introduction to social statistics using Stata

  http://www.ernestoamaral.com/stata2020a.html

- You can run all Stata commands that were used in this lecture using this DO-file

http://www.ernestoamaral.com/docs/Stata2020a/Stata05.txt


TEXAS A&M
UNIVERSITY