Lecture (chapter 13): Association between variables measured at the interval-ratio level

Ernesto F. L. Amaral

November 12–14, 2018
Advanced Methods of Social Research (SOCI 420)

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

TEXAS A&M UNIVERSITY

Chapter learning objectives

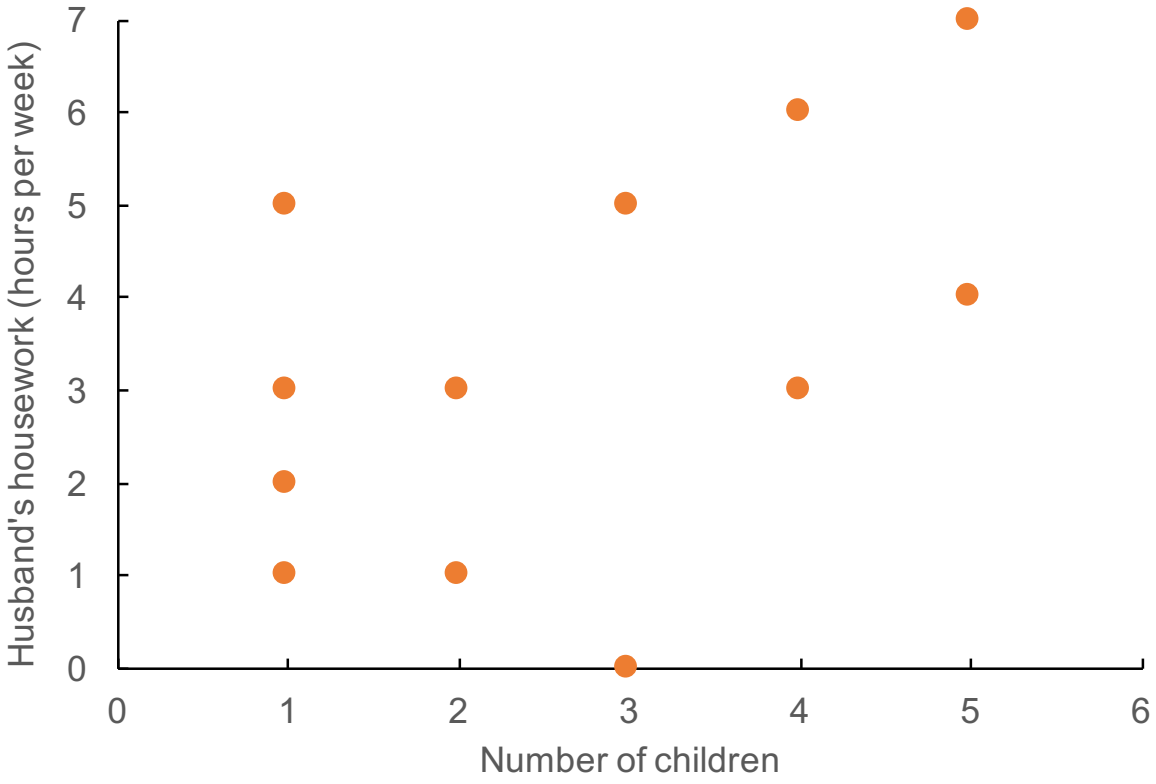
- Interpret a scatterplot
- Calculate and interpret slope (*b*), *Y* intercept (*a*), Pearson's *r*, and r^2
- Find and explain the least-squares regression line and use it to predict values of *Y*
- Explain the concepts of total, explained, and unexplained variance
- Use regression and correlation techniques to analyze and describe a bivariate association
- Test Pearson's *r* for significance: five-step model

Scatterplots

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

Example of a scatterplot

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

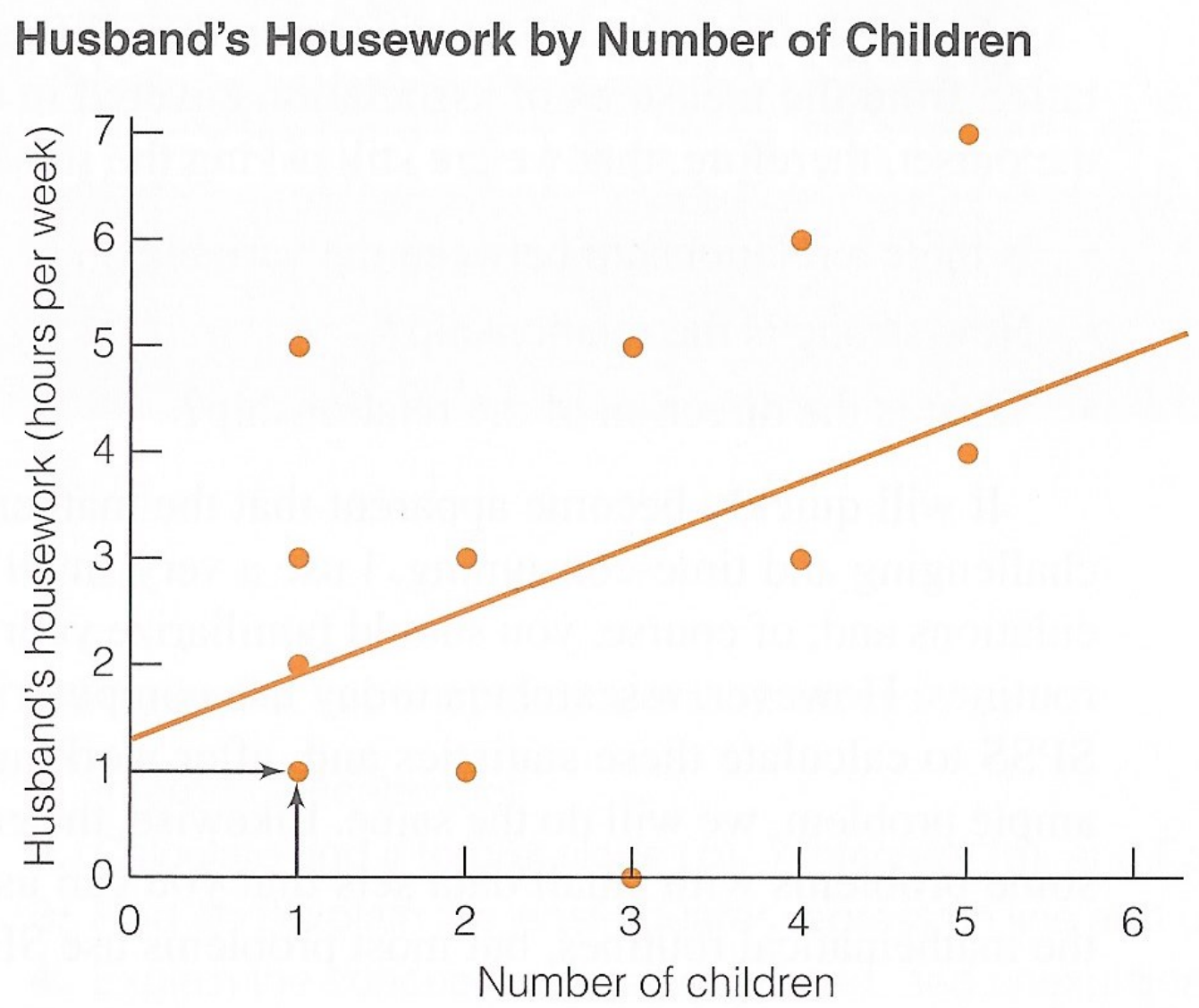


Source: Healey 2015, p.343.

Regression line

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

Scatterplot with regression line



Source: Healey 2015, p.344.

Use of scatterplots

- Scatterplots can be used to answer these questions

1. Is there an association?

2. How strong is the association?

3. What is the pattern of the association?

1. Is there an association?

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

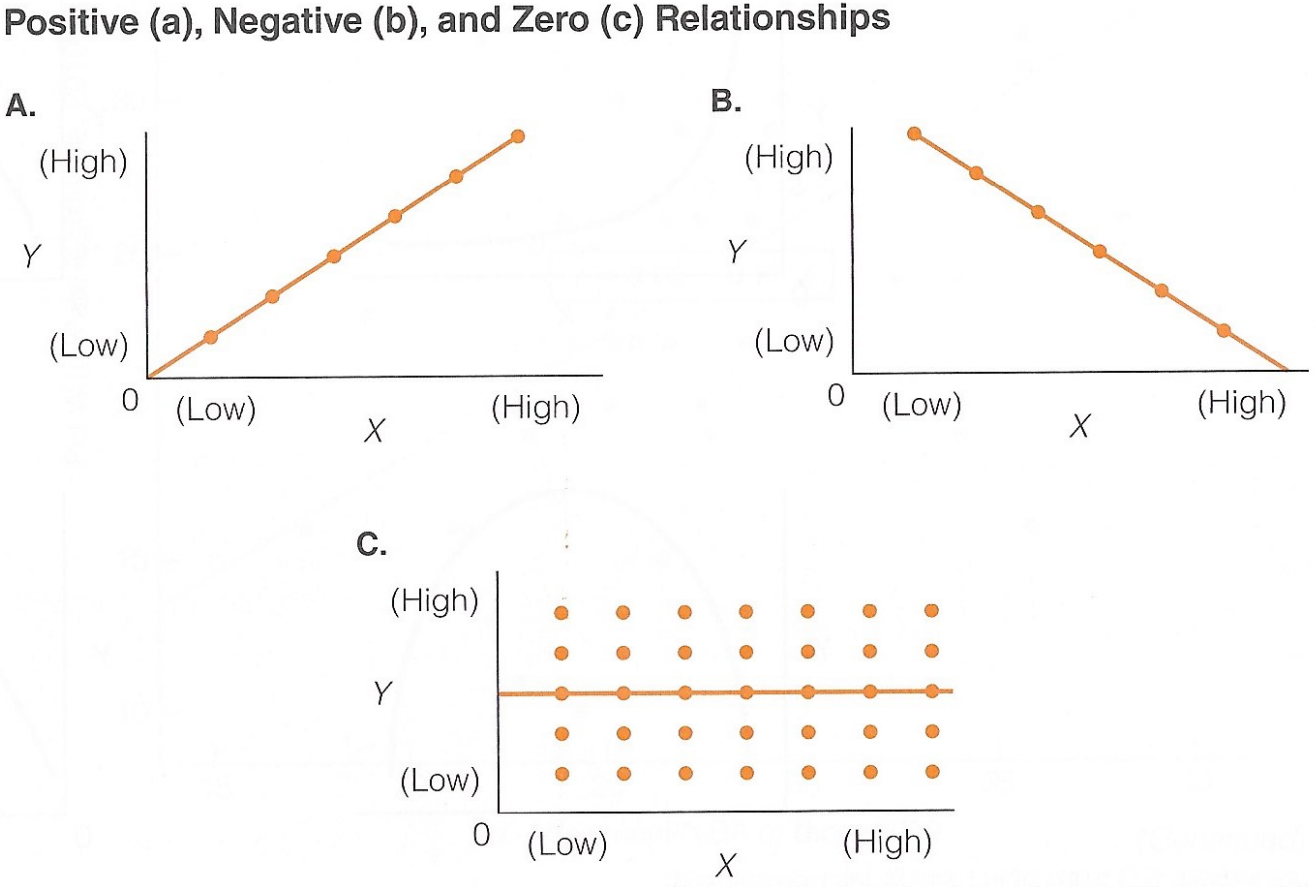
- If the regression line has an angle to the X axis
  - We can conclude that an association exists between the two variables
  - The line is not parallel to the X axis

2. How strong is the association?

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

3. Pattern of the association

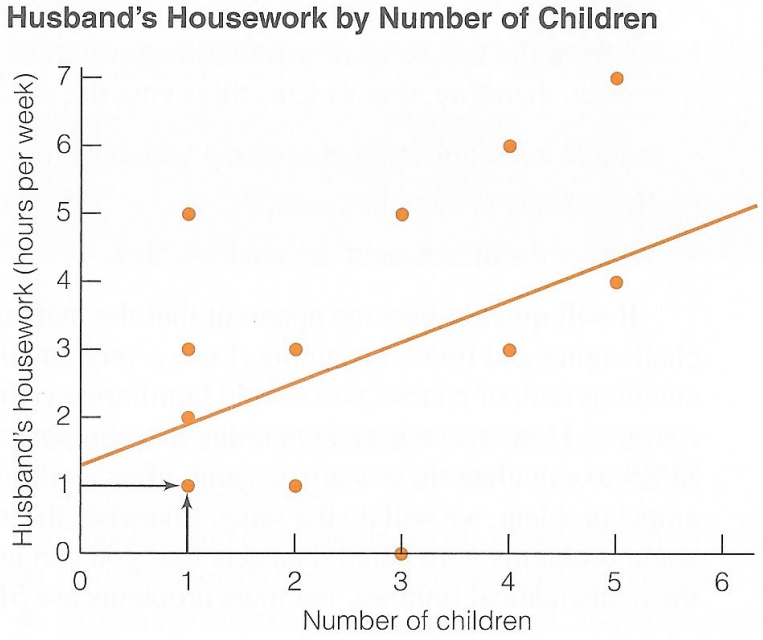
- The pattern or direction of association is determined by the angle of the regression line



Source: Healey 2015, p.345.

Check for linearity

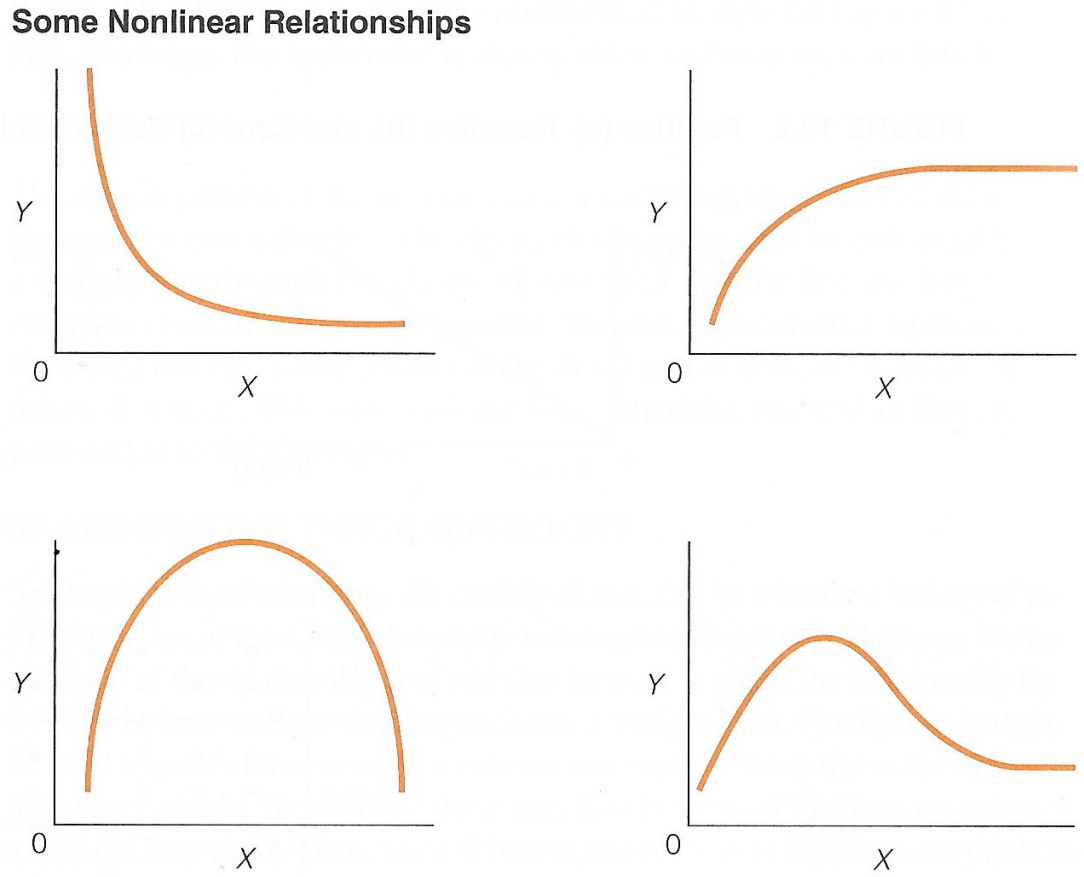
- Scatterplots can be used to check for linearity
  - An assumption of scatterplots and linear regression analysis is that X and Y have a linear correlation
  - In a linear association, the dots of a scatterplot form a straight line pattern



Source: Healey 2015, p.344.

Nonlinear associations

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

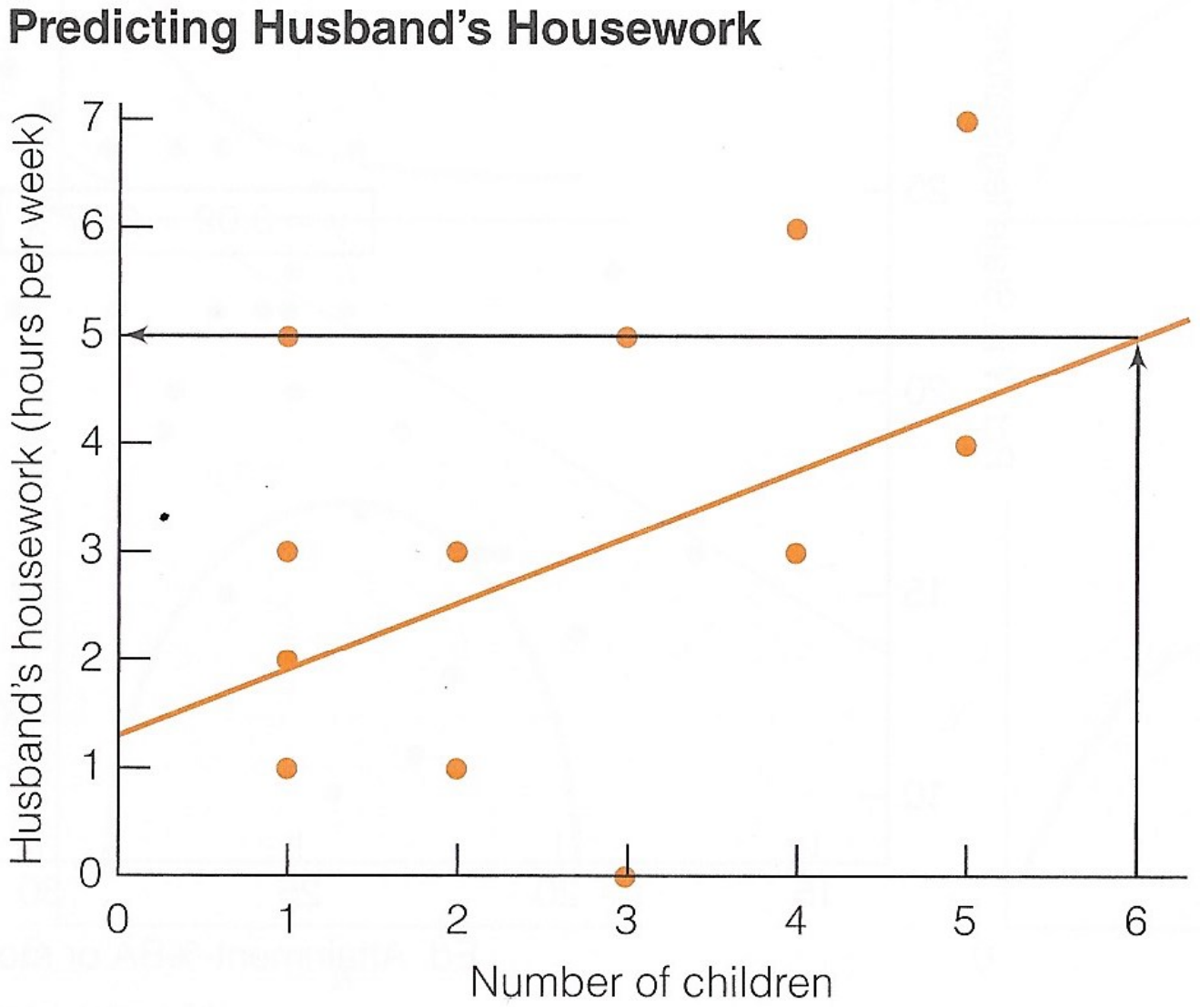


Source: Healey 2015, p.346.

Prediction

- Scatterplots can be used to predict values of Y (Y' or $\hat{Y}$) based on values of X
- Locate a particular X value on the horizontal axis
- Draw a vertical line up to the regression line
- Then draw a horizontal line over to the vertical axis

Example of prediction



Source: Healey 2015, p.348.

Estimating the regression line

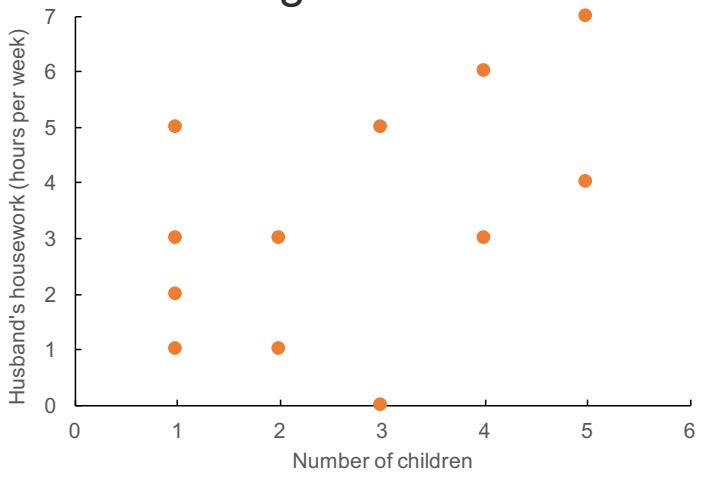
- The regression line touches each conditional mean of Y
  - Or the line comes as close as possible to all scores

- The dots above each value of X can be thought of as conditional distributions of Y
  - In previous chapters, column percentages were the conditional distributions of Y for each value of X

Conditional means of Y

- Conditional means of Y are found by summing all Y values for each value of X and dividing by the number of cases

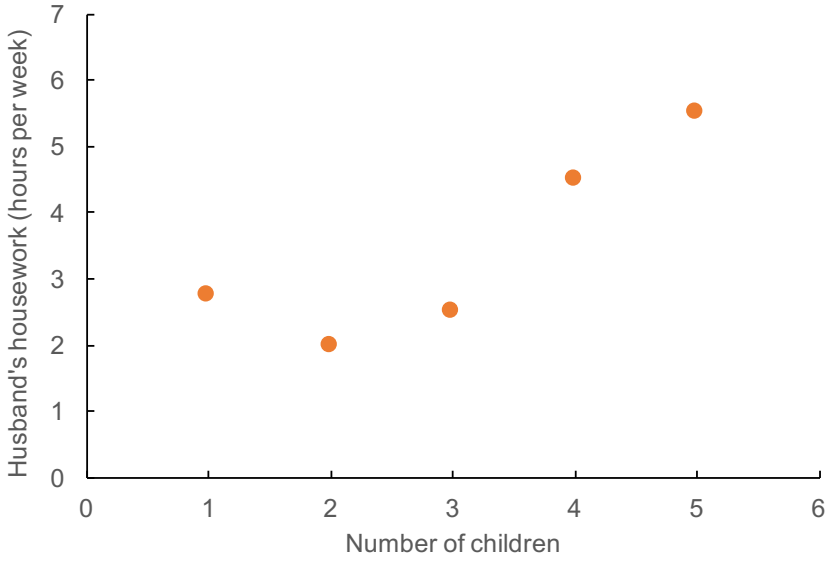
Original data



Conditional means of Y

Number of Children (*X*)	Husband's Housework (*Y*)	Conditional Mean of *Y*
1	1, 2, 3, 5	2.75
2	3, 1	2.00
3	5, 0	2.50
4	6, 3	4.50
5	7, 4	5.50

Conditional means of Y



Source: Healey 2015, p.349.

Regression coefficients

- Ordinary least squares (OLS) simple regression
  - OLS: linear regression
  - Simple: only one independent variable

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

- Where
  - Y = score on the dependent variable
  - X = score on the independent variable
  - $a = \beta_0$ = the Y intercept or the point where the regression line crosses the Y axis
  - $b = \beta_1$ = slope of the regression line or the amount of change produced in Y by a unit change in X

Computing the slope (*b*)

- Before using the formula for the regression line, we need to estimate *a* and *b*
- First, estimate *b*

$$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2}$$

- The numerator of the formula is the "covariation" of X and Y
  - How much X and Y vary together
  - Its value reflects the direction and strength of the association between X and Y

Computing the Y intercept (*a*)

- The intercept (*a*) is the point where the regression line crosses the Y axis

- Estimate *a* using the mean for X, the mean for Y, and *b*

$$a = \bar{Y} - b\bar{X}$$

Example

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

Number of Children and Husband's Contribution to Housework (fictitious data)

Family	Number of Children	Hours per Week Husband Spends on Housework
A	1	1
B	1	2
C	1	3
D	1	5
E	2	3
F	2	1
G	3	5
H	3	0
I	4	6
J	4	3
K	5	7
L	5	4

Source: Healey 2015, p.343.

Example: calculation table

- Calculation of *b* is simplified if you set up a computation table

Computation of the Slope (*b*)

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

$$\overline{X} = \frac{32}{12} = 2.67$$

$$\overline{Y} = \frac{40}{12} = 3.33$$

Source: Healey 2015, p.351.

Example: slope and intercept

- Based on previous table, estimate the slope (*b*)

$$b = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sum(X - \bar{X})^2} = \frac{18.32}{26.68} = 0.69$$

- Estimate the intercept (*a*)

$$a = \bar{Y} - b\bar{X} = 3.33 - (0.69)(2.67) = 1.49$$

Example: interpretations

- Regression equation with *a*=1.49 and *b*=0.69

$$Y' = 1.49 + (0.69)X$$

  - *b* = 0.69
    - For every additional child in the dual-career household, husbands perform on average an additional 0.69 hours (around 36 minutes) of housework per week
  - *a* = 1.49
    - The regression line crosses the Y axis at 1.49
    - When there are zero children in a dual-career household, husbands perform on average 1.49 hours of housework per week

Example: predictions

- What is the predicted value of Y (Y') when X equals 6?

  $$Y' = 1.49 + (0.69)X = 1.49 + (0.69)(6) = 5.63$$

  - In dual-career families with 6 children, the husband is predicted to perform on average 5.63 hours of housework a week

- What about when X equals 7?

  $$Y' = 1.49 + (0.69)X = 1.49 + (0.69)(7) = 6.32$$

  - In dual-career families with 7 children, the husband is predicted to perform on average 6.32 hours of housework a week
  - Notice how the difference in these two predicted values equals *b* (6.32 – 5.63 = 0.69)

GSS example

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

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

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

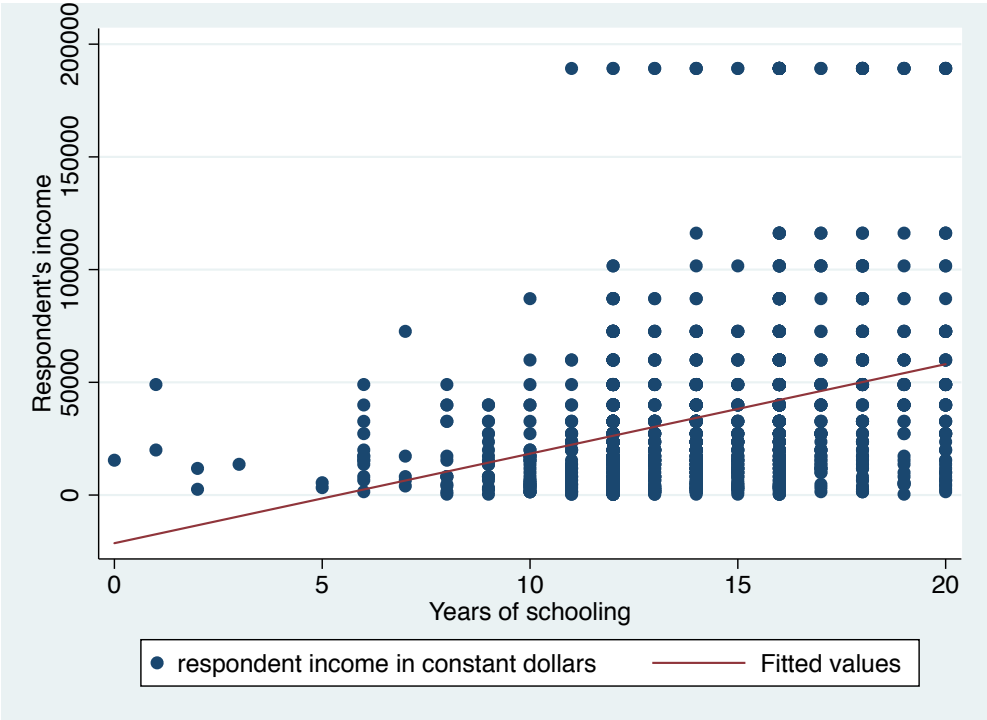
```
. svy: reg conrinc educ
(running regress on estimation sample)

Survey: Linear regression

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

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

Source: 2016 General Social Survey.

Edited figure

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



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

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

Pearson's *r*

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

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

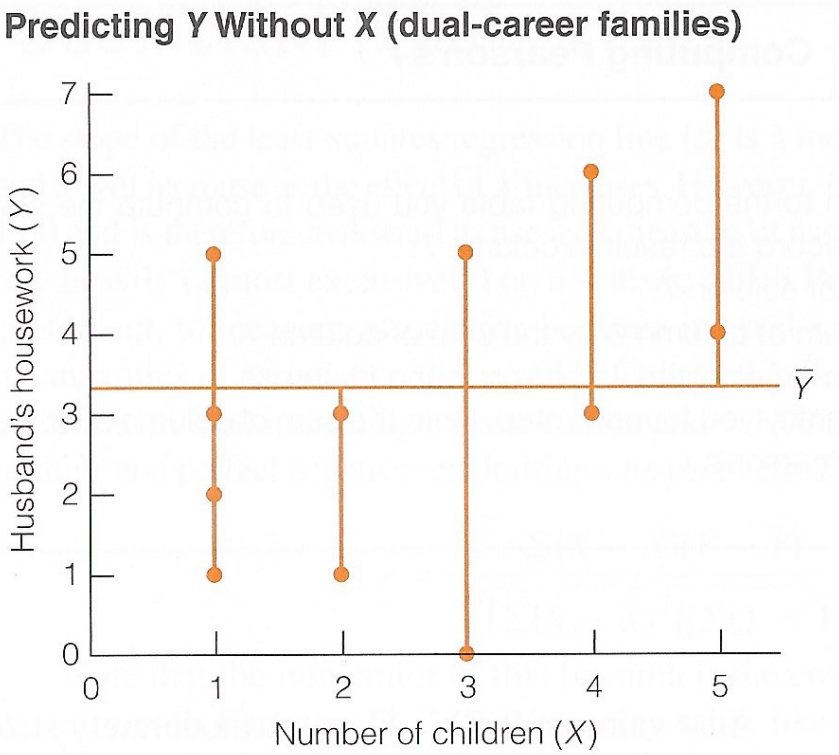
- Pearson's *r* indicate the direction of association
  - –1.00 indicates perfect negative association
  - 0.00 indicates no association
  - +1.00 indicates perfect positive association
- It doesn't have a direct interpretation of strength

Coefficient of determination (r^2)

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

Predicting Y without X

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

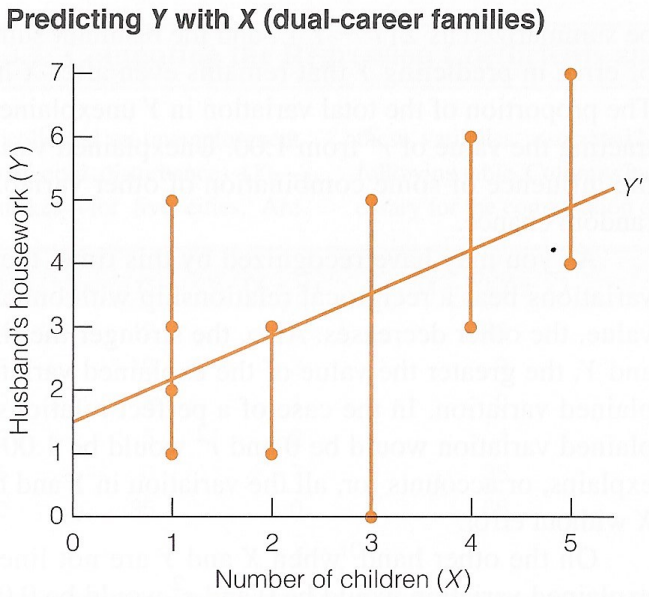


Source: Healey 2015, p.356.

Predicting Y with X

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

$Y' = a + bX$



Source: Healey 2015, p.357.

Estimating r^2

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

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

Source: Healey 2015, p.357.

Unexplained variation

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

Example: Pearson's *r*

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

Computation of Pearson's *r*

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

Source: Healey 2015, p.354.

Example: calculate *r*

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

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

$$r = 0.50$$

Example: interpretation

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

Test Pearson's *r* for significance

- Use the five-step model

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

Step 1: Assumptions,requirements

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

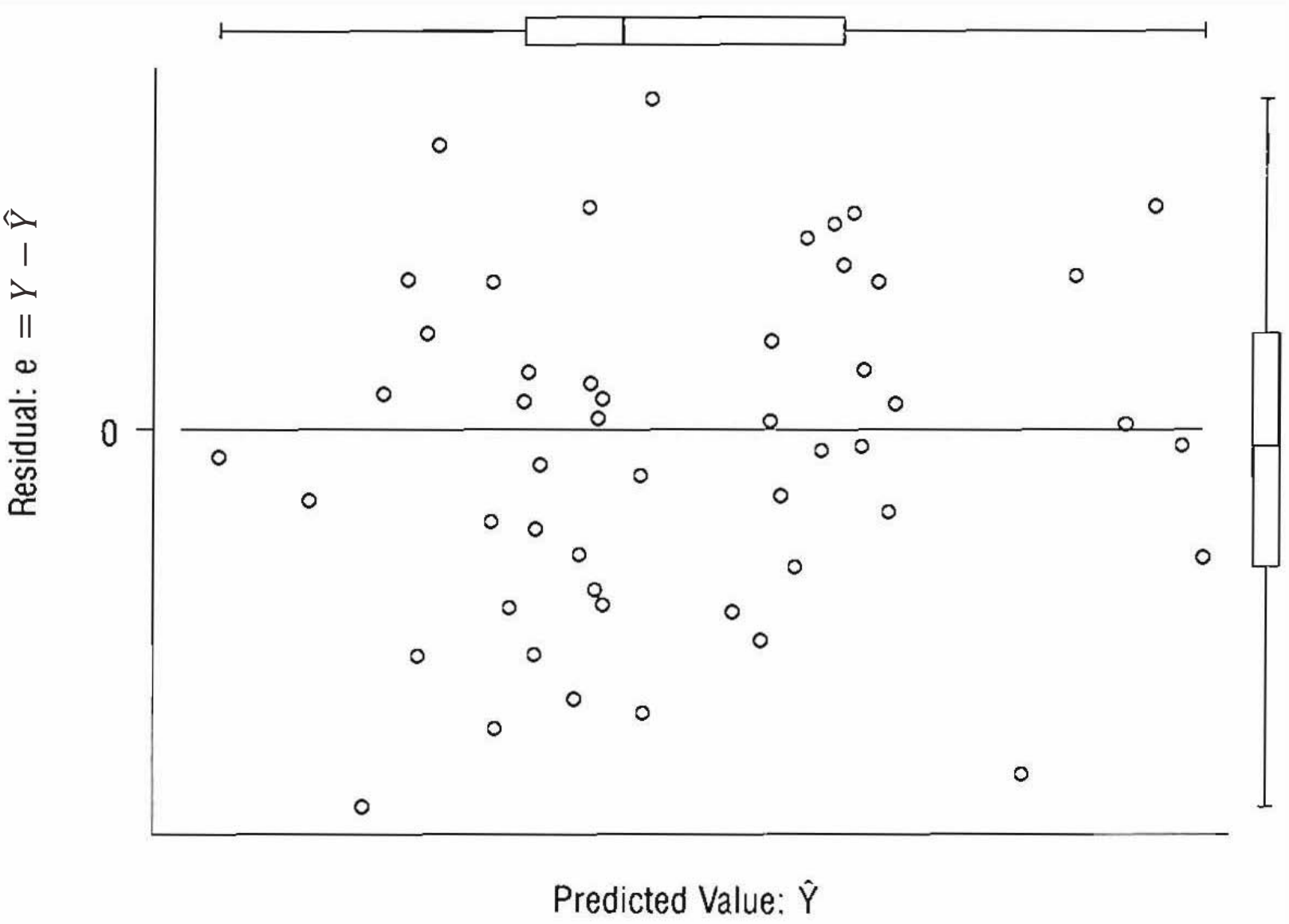


Figure 2.10 "All clear" *e*-versus-$\hat{Y}$ plot (artificial data).

Fonte: Hamilton, 1992: 52.

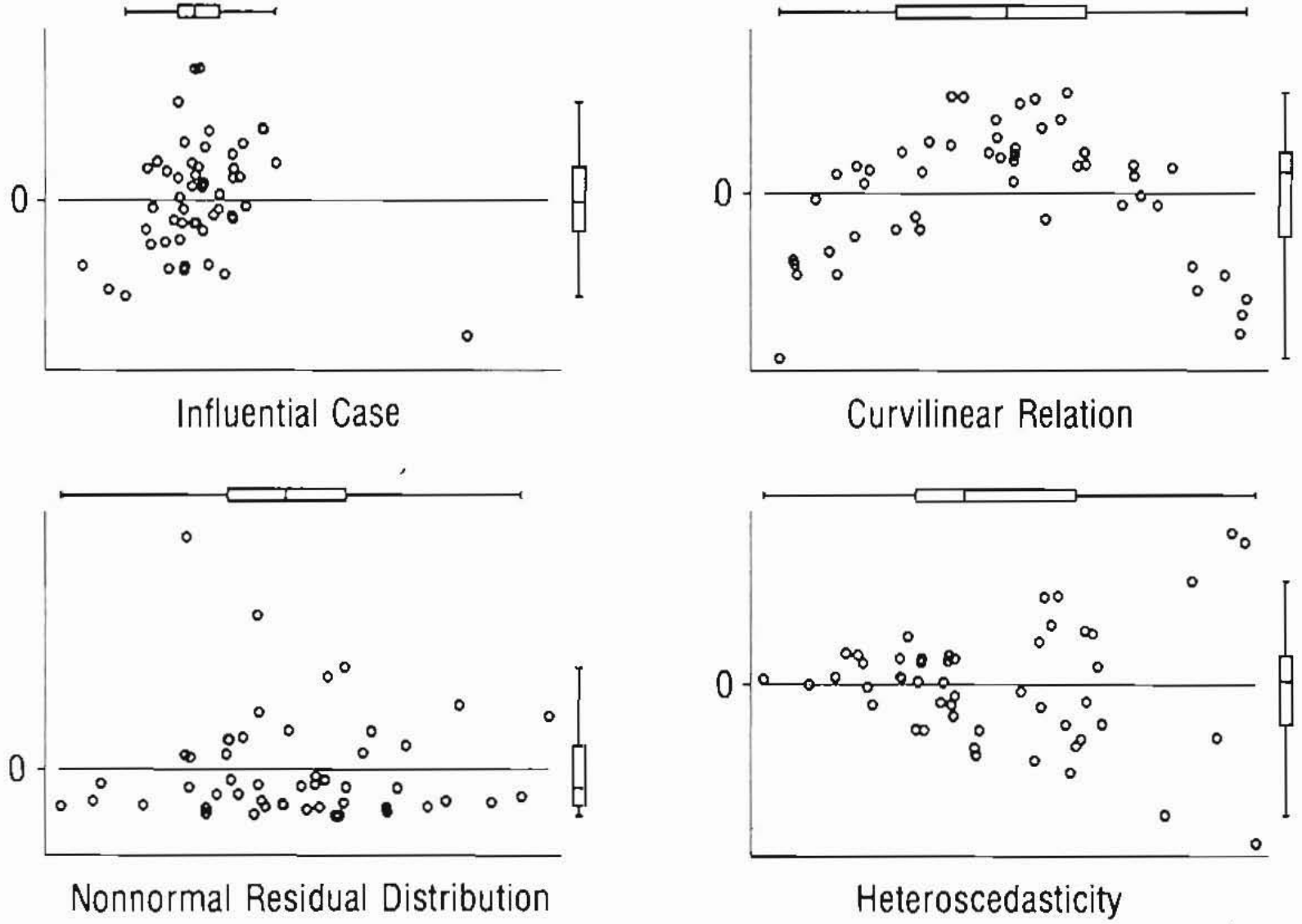


Figure 2.11 Examples of trouble seen in e-versus-$\hat{Y}$ plots (artificial data).

Fonte: Hamilton, 1992: 53.

Step 2: Null hypothesis

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

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

Step 3: Distribution, critical region

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

Step 4: Test statistic

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

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

$$t(obtained) = 1.83$$

Step 5: Decision, interpret

- *t(obtained)* = 1.83
  - This is not beyond the *t(critical)* = ±2.228
  - The *t*(obtained) does not fall in the critical region, so we ***fail to reject*** the H_0

- The two variables are not correlated in the population
  - The correlation between number of children (X) and hours per week husband spends on housework (Y) is not statistically significant

Correlation matrix

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

A Correlation Matrix Showing the Relationships Among Four Variables

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

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

Source: Healey 2015, p.360.

GSS example

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

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

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

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

Source: 2016 General Social Survey.

Edited table

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

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

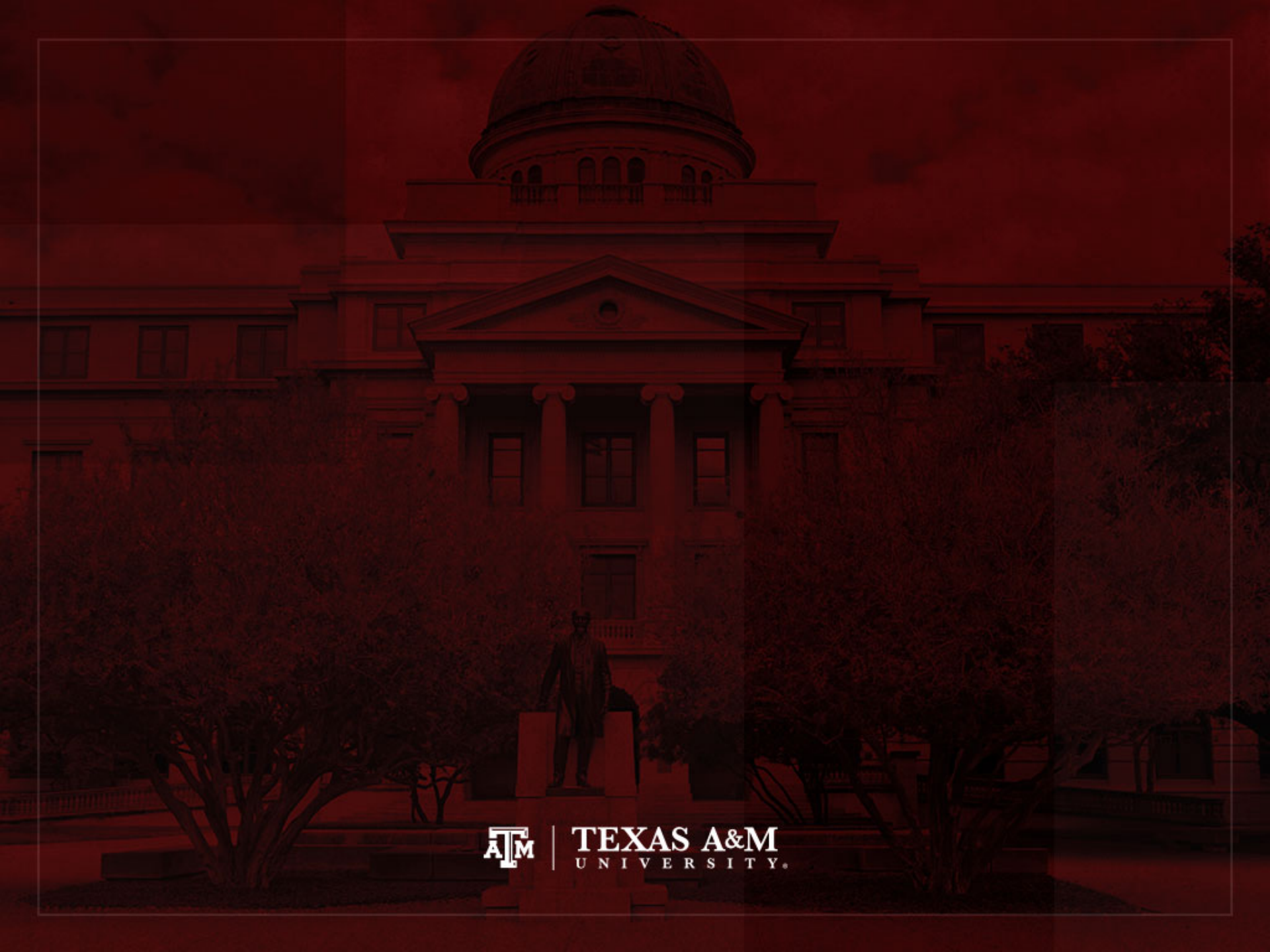
Note: Pearson's *r* and coefficient of determination (r^2) were generated taking into account the survey weight of the General Social Survey. *Significant at p<0.10; **Significant at p<0.05; ***Significant at p<0.01.
Source: 2016 General Social Survey.

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


TEXAS A&M
UNIVERSITY