02 -Review of Statistics and Demography

Ness Sandoval

Sociology

Saint Louis University

Outline

- Measurement
- Descriptive Statistics
- Simple Linear Regression
- Advanced Topics in Regression
- Review of Demography
- Example 1
- Example 2

Measurement

	Nominal	Ordinal	Interval	Ratio
Examples	Categories	Ranks	Test scores and scales	Weight, number of responses
Properties	Identity	Identity Magnitude	Identity Magnitude Equal Interval	Identity Magnitude Equal Interval True Zero
Mathematical Operations	None	Rank order	Add; subtract	Add, subtract; multiply and divide
Type of data	Nominal	Ordered	Score	Score

Descriptive Statistics

Quantitative Research Basics

1.1 *Variation*

- Quantitative research is all about explaining variation (differences)
- Sources of variation: multiple cases are differentiated characteristics (variables)

1.2 *Inference*

- ***Quantitative research*** is all about making predications about the unknown.
  - A ***population*** is any collection of objects (N) of research interests that are alike on at least one specific characteristic.
  - A ***sample*** is a subset of objects drawn from a population; sampling is based on probability (random sample, weighted sample, and stratified sample).

Measures of Central Tendency

Numbers that describe what is average or typical of the distribution.

- The **Mode** is the category or score with the largest frequency in the distribution

- The **Median** is the score that divides the distribution into tow equal parts so that half the cases are above it and half below it.

- The **Mean** is the arithmetic average obtained by adding up all the scores and dividing by the total number of score

Measure of Variability

- The ***variance*** is a measure of variation for interval-ratio variables; it is the average of the squared deviation from the mean.

- The ***standard deviation*** is a measure of variation for interval-ratio variables; it is equal to the square root of the variance.

Shape of the Distribution

Symmetrical distribution

- The frequencies at the right and left tails of the distribution are identical; each half of the distribution is the mirror image of the other.

Skewed distribution

- A distribution with a few extreme values on one side of the distribution.

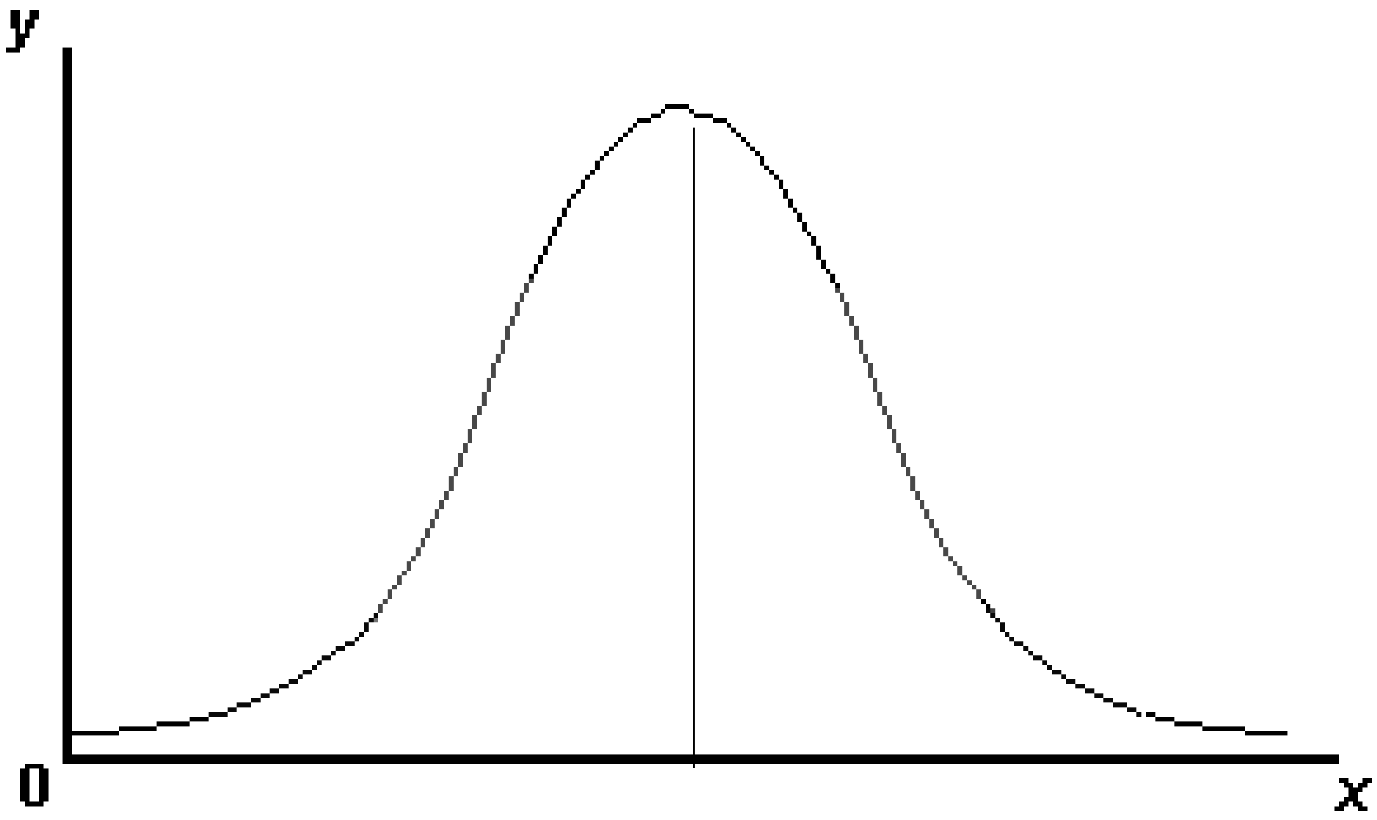
Negatively skewed distribution

- A distribution with a few extremely low values.

Positively skewed distribution

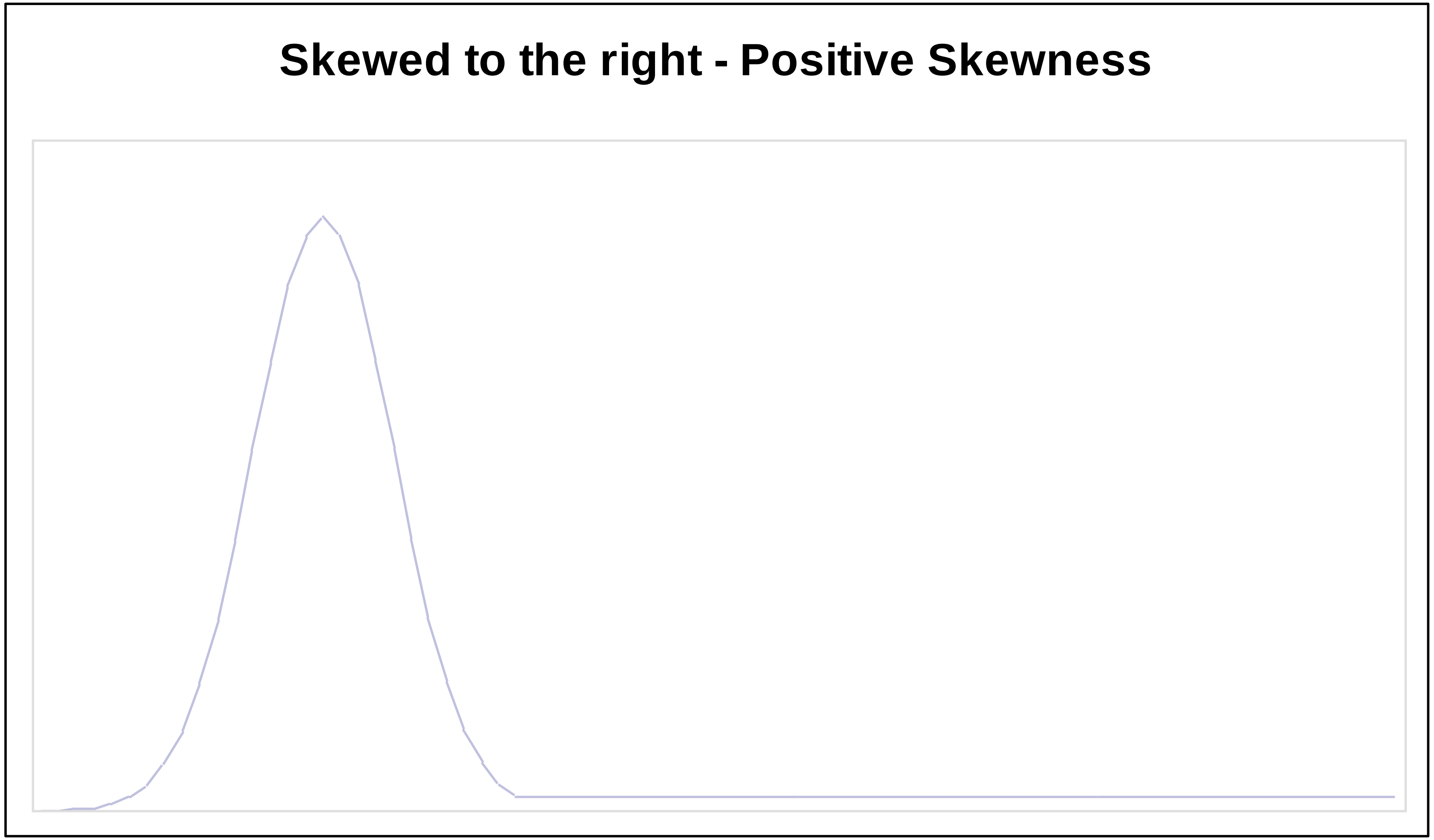
- A distribution with a few extremely high values.

The Normal Curve



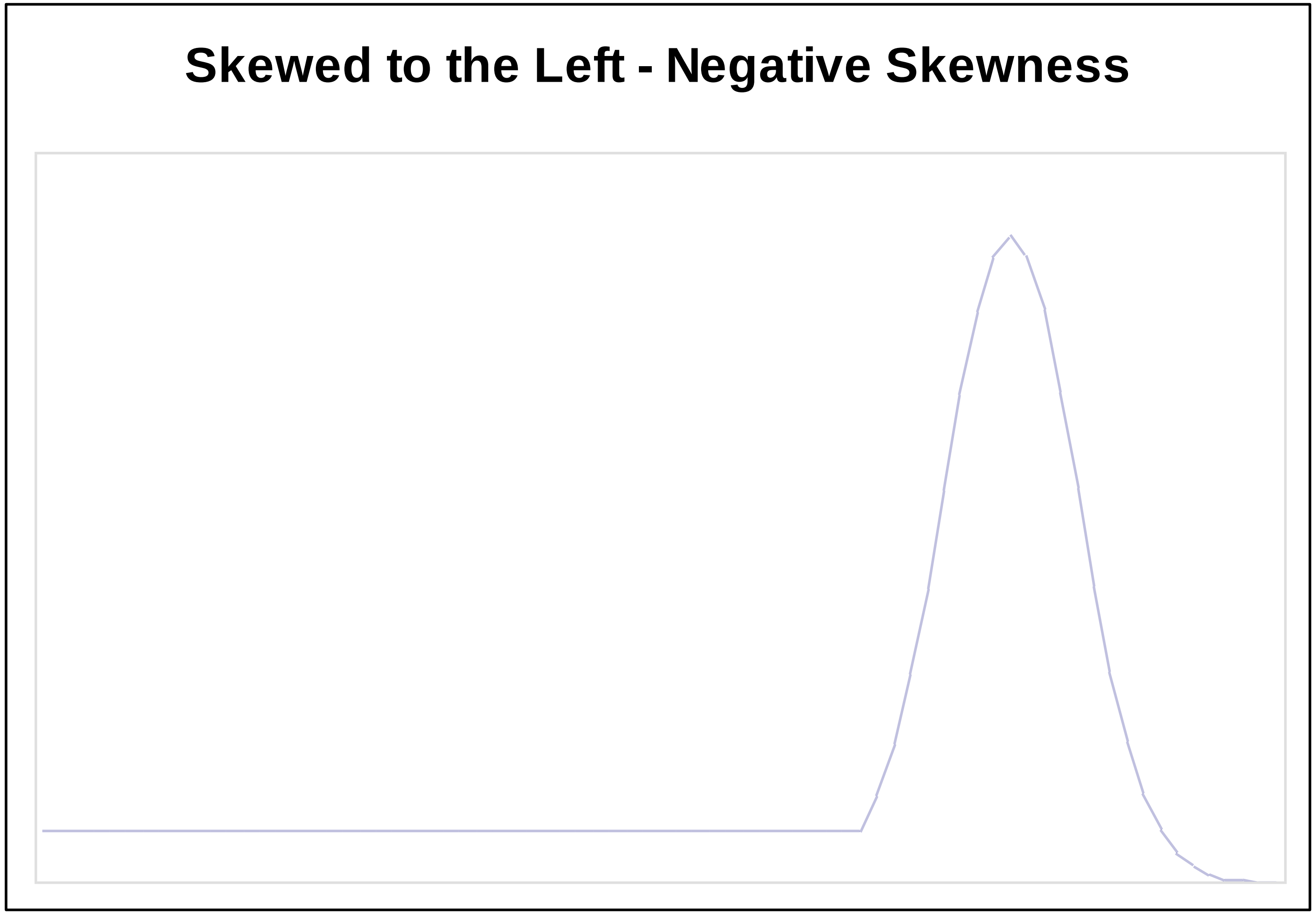
Mean=Mode=Median

Skewness in the Distribution



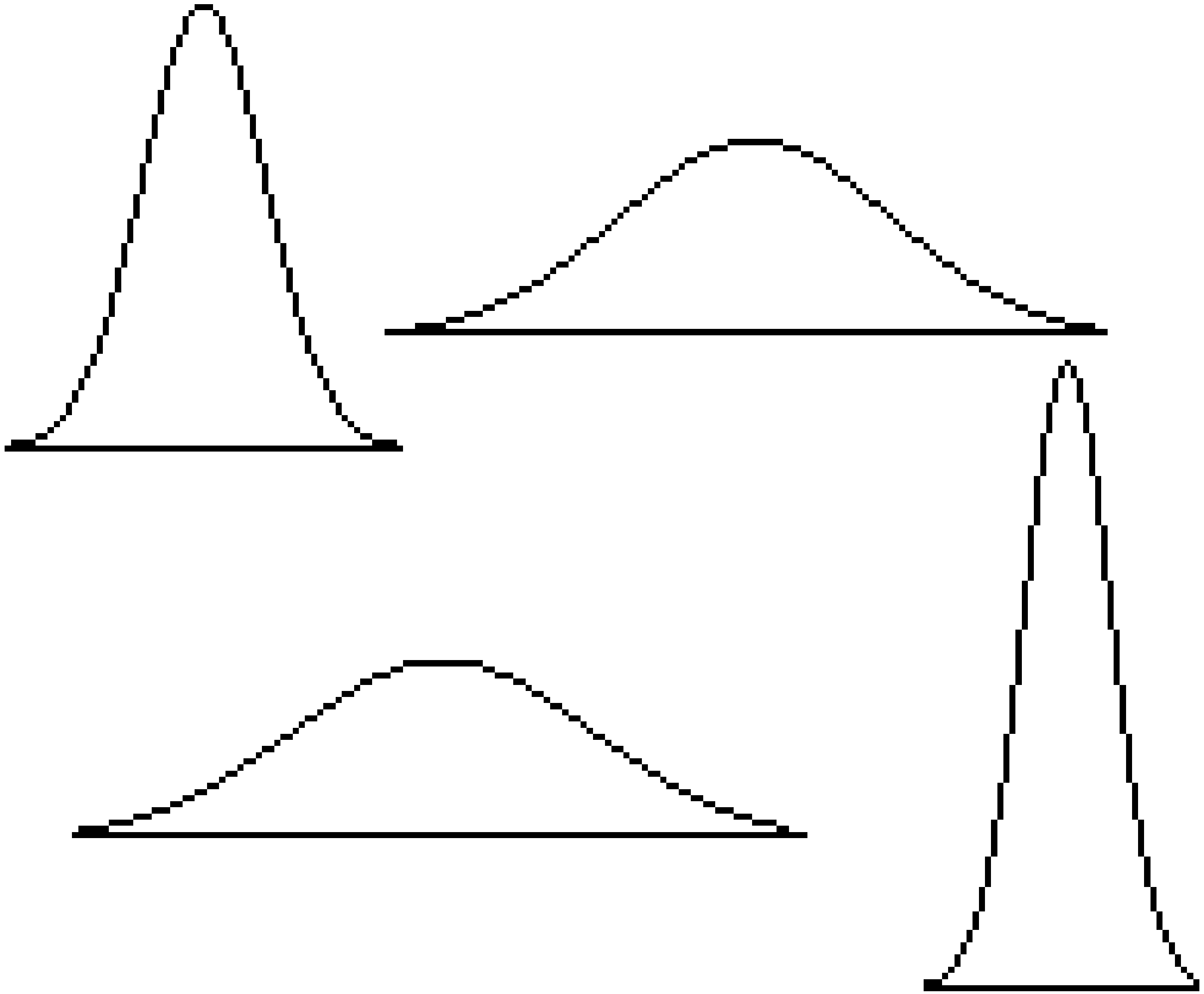
Mean>Median>Mode

Skewness in the Distribution

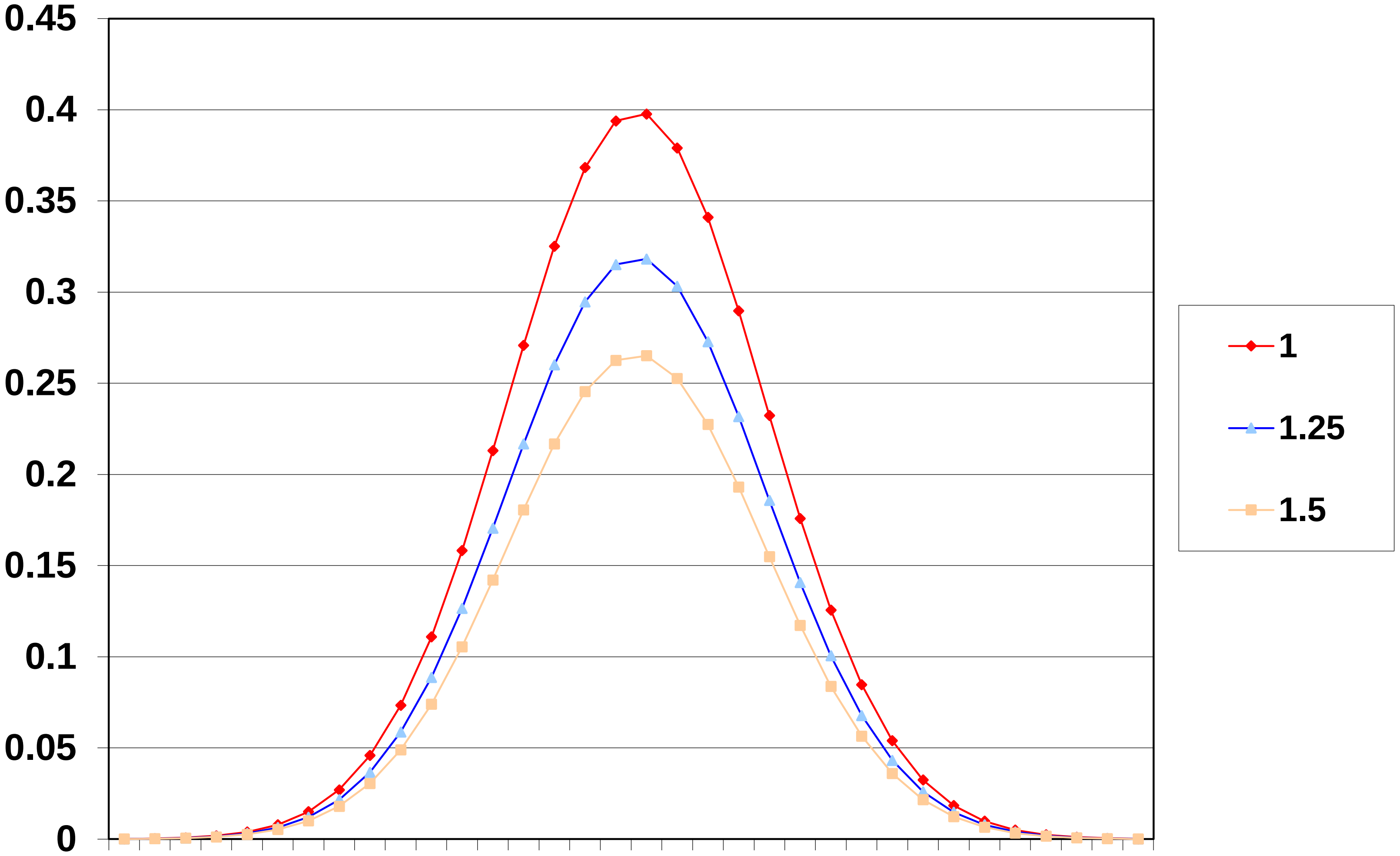


Mean<Median<Mode

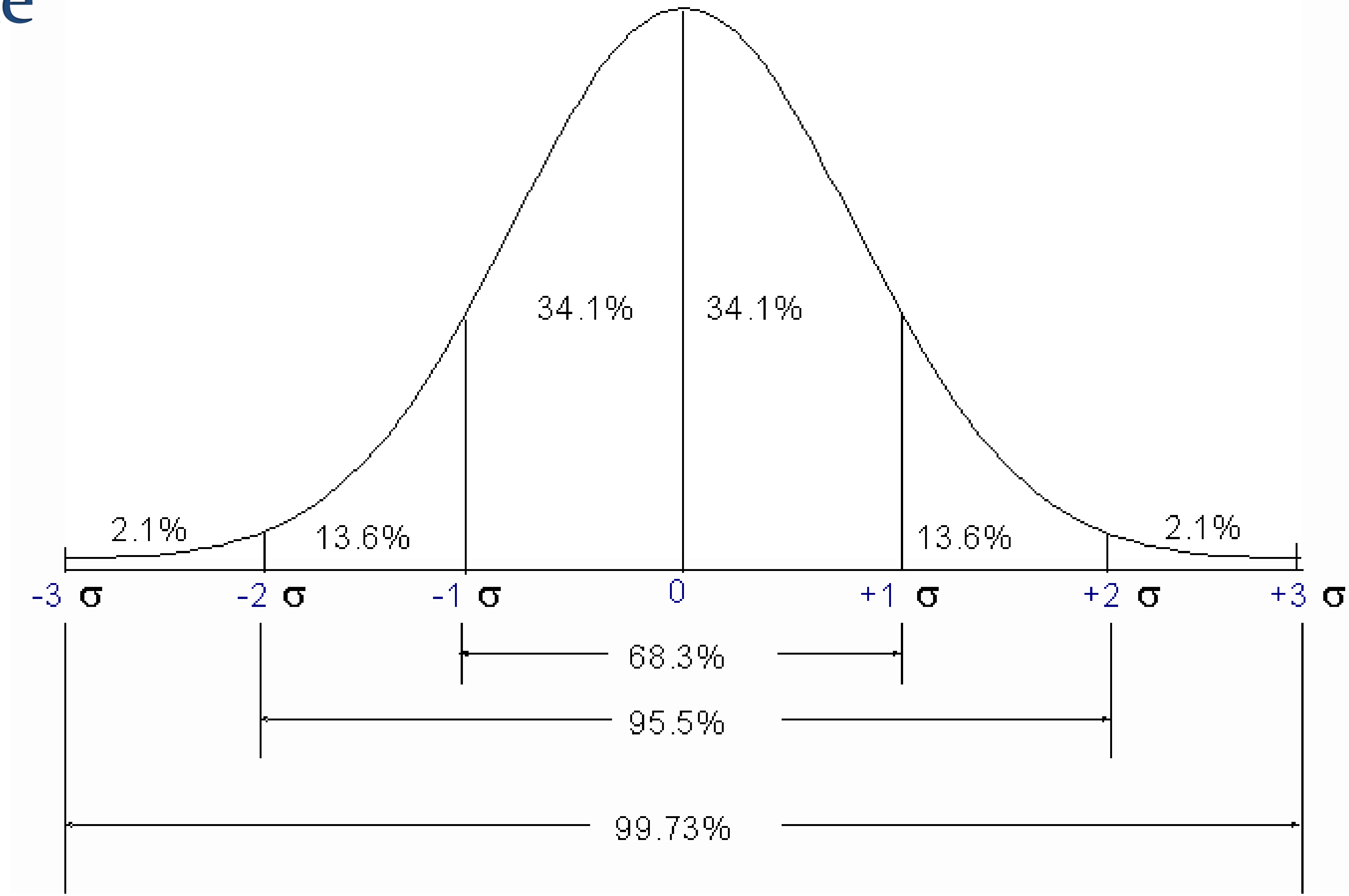
Characteristics of a Normal Curve



Normal Distribution with Equal Means but Different Standard Deviations



Percentages Under the Normal Curve or “The 68-95-99 rule”



Five Step Research Process

1. *Assumptions of Statistical Hypothesis Testing*

- Statistical hypothesis testing requires several assumptions.

  These assumptions include considerations of the level of measurement of the variable, the method of sampling, the shape of the population distribution, and the sample size.

  The specific assumptions may vary, depending on the test or the conditions of testing.

2. *State the Research and Null Hypotheses*

- *Research Hypothesis (H_1)*- A statement reflecting the substantive hypothesis. It is always expressed in terms of population parameters, but its specific form varies from test to test.

- *Null Hypothesis (H_0)* – A statement of “no difference” which contradicts the research hypothesis and is always expressed in terms of population parameters.

3. Select Sampling Distribution & Establish Critical Region

- *One-tailed test* – A type of hypothesis test that involves a directional hypothesis. It specifies that the values of one group are either larger or smaller than some specified population value.

- *Right-tailed test* – A one-tail test in which the sample outcome is hypothesized to be at the right tail of the sampling distribution.

- *Left-tailed test* – A one-tailed test in which the sample outcome is hypothesized to be at the left tail of the sampling distribution.

- *Two-tailed test* – A type of hypothesis test that involves a non-directional research hypothesis. We are equally interested in whether the values are less than or greater than one another. The sample outcome may be located at both the low and high ends of the sampling distribution.

- *Select the Test Statistic (Z or t)*

- *Select the P Value* – The probability associated with the obtained Value of Z.

- *Select Alpha* = Alpha is the level of probability at which the null hypothesis is rejected. It is customary to set alpha at the .05, .01 or .001.

4 Computing the Test Statistic

5 Making the decision and interpreting the results

Errors in Hypothesis Testing

- *Type 1 Error* – The probability associated with rejecting a null hypothesis when it is true.

- *Type II Error* – The probability associated with failing to reject null hypothesis when it is false.

	H_o is True	H_o is False
Reject H_o	Type I Error	Correct Decision
Do not reject H_o	Correct Decision	Type II Error

Simple Linear Regression

What is wrong with this analysis?

```
. corr  rincom98  sex age  race
(obs=946)

             | rincom98      sex      age     race
-------------+------------------------------------
    rincom98 |   1.0000
         sex |  -0.2383   1.0000
         age |   0.2117  -0.0420   1.0000
        race |  -0.1376   0.0594  -0.1581   1.0000
```

```
. tab race

    race of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
      white |      1,114       78.01       78.01
      black |        202       14.15       92.16
      other |        112        7.84      100.00
------------+-----------------------------------
      Total |      1,428      100.00


. gen white=race

. recode white 1=1 2=0 3=0
(white: 314 changes made)

. tab white

      white |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        314       21.99       21.99
          1 |      1,114       78.01      100.00
------------+-----------------------------------
      Total |      1,428      100.00

. gen black=race

. recode black 1=0 2=1 3=0
(black: 1428 changes made)

. tab black

      black |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,226       85.85       85.85
          1 |        202       14.15      100.00
------------+-----------------------------------
      Total |      1,428      100.00

. gen other=race

. recode other 1=0 2=0 3=1
(other: 1428 changes made)

. tab other

      other |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,316       92.16       92.16
          1 |        112        7.84      100.00
------------+-----------------------------------
      Total |      1,428      100.00

.
```

This looks better!

```
. corr  rincom98  sex age  white black other
(obs=946)

             | rincom98      sex      age    white    black    other
-------------+------------------------------------------------------
    rincom98 |   1.0000
         sex |  -0.2383   1.0000
         age |   0.2117  -0.0420   1.0000
       white |   0.1580  -0.0758   0.1337   1.0000
       black |  -0.1351   0.0766  -0.0412  -0.7562   1.0000
       other |  -0.0675   0.0175  -0.1502  -0.5538  -0.1261   1.0000
```

Causality and Notion of Ceteris Paribus

- The goal of regression models is to infer that one variable (such as education) has a *causal effect* on another variable (worker productivity).

- Ceteris Paribus
  - Means other (relevant) factors being equal

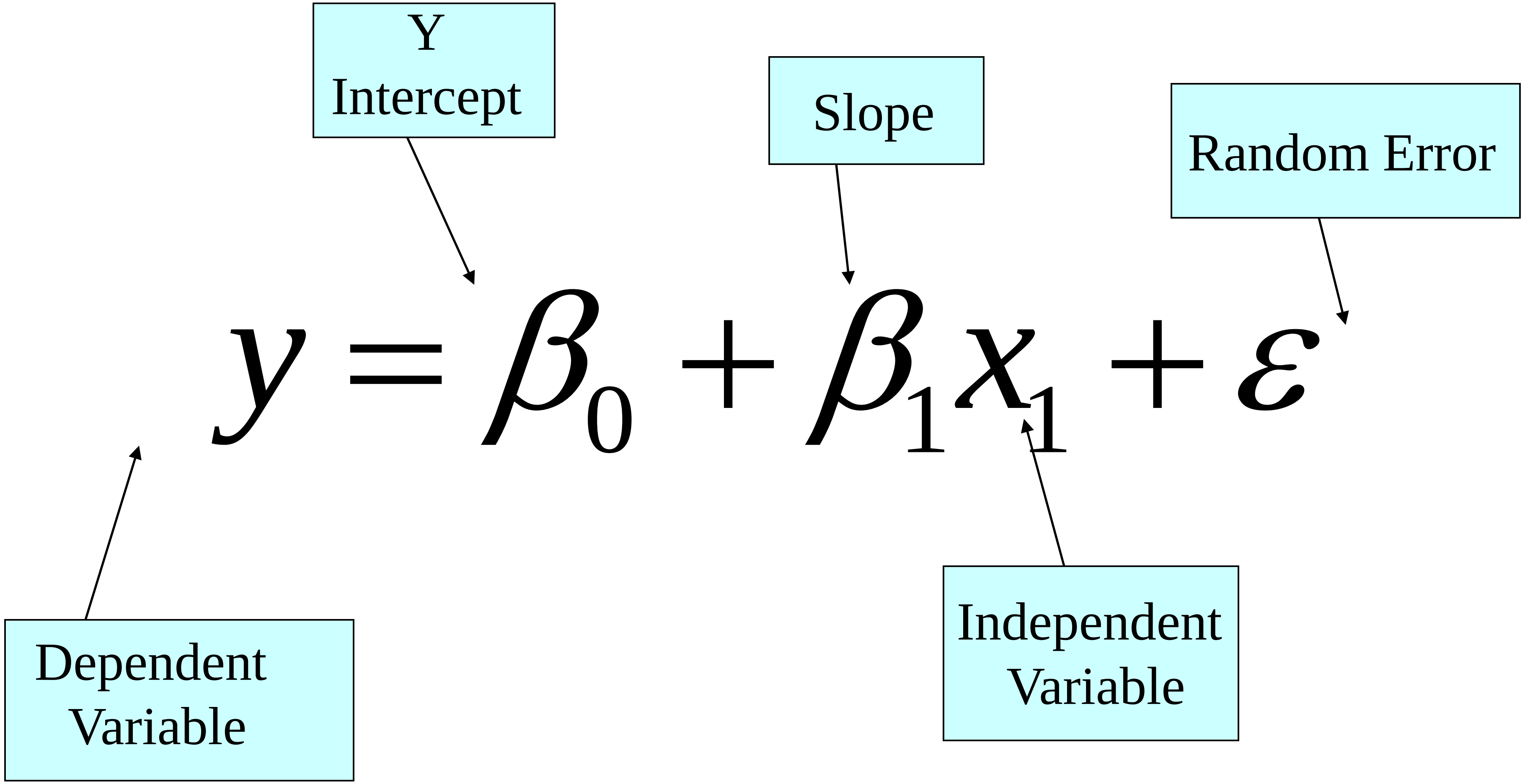
  - Plays an important part role in causal analysis

Ceteris Paribus Example

$$wage = f(educ, lanuage, training)$$

- Observed factors
  - Very easy
    - e.g., speaks English (Yes or No)
    - e.g., high school diploma
- Unobserved factors
  - Difficult
    - e.g., ability to speak English (this is what may matter in wages)
    - e.g., breadth of vocabulary

The Regression Model



Y Intercept

Slope

Random Error

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

Dependent Variable

Independent Variable

Terminology for Simple Regression

y	x
Dependent Variable	Independent Variable
Explained Variable	Explanatory Variable
Response Variable	Control Variable
Predicted Variable	Predictor Variable
Regressand	Regressor

Example

$$(1.1)\quad crime = \beta_0 + \beta_1 wage_m + \varepsilon$$

$$(1.2)\quad wage = \beta_0 + \beta_1 educ + \varepsilon$$

$$(1.3)\quad \Delta y = \beta_1 \Delta x \; if \;\; \Delta\varepsilon = 0$$

$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- ε= the deviation of the value Y from the mean value of the distribution given X.

  This error term may be conceived as representing:

  1. The effects on Y of variables not explicitly included in the equation

  2. A residual random element in the dependent variable.

ε

- For historical reasons, the most popular philosophy about ε is that it is "random noise." Some researchers also call it "disturbance" or "unobserved."

  More specifically, that it is:

  1. uncorrelated with other variables

  2. has a mean value of zero.

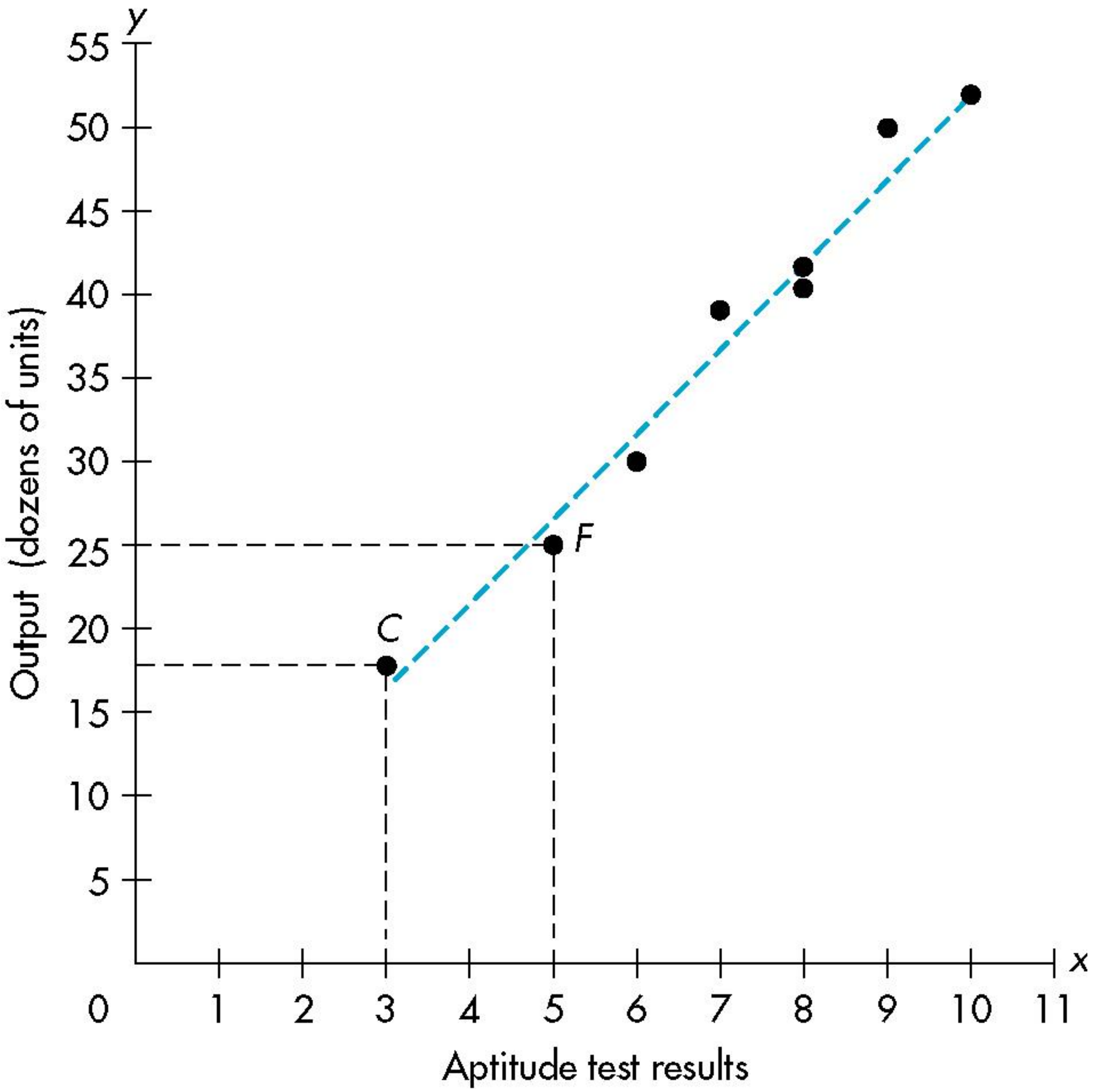
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- The y intercept.
  - Geometrically, it represents the value of E(y) where the regression surface (or plane) crosses the y axis.
  - Substantively, it is the expected value of y when the X equal 0.
  - It has more MEANINGFUL properties when we moved to the case of multiple regression.

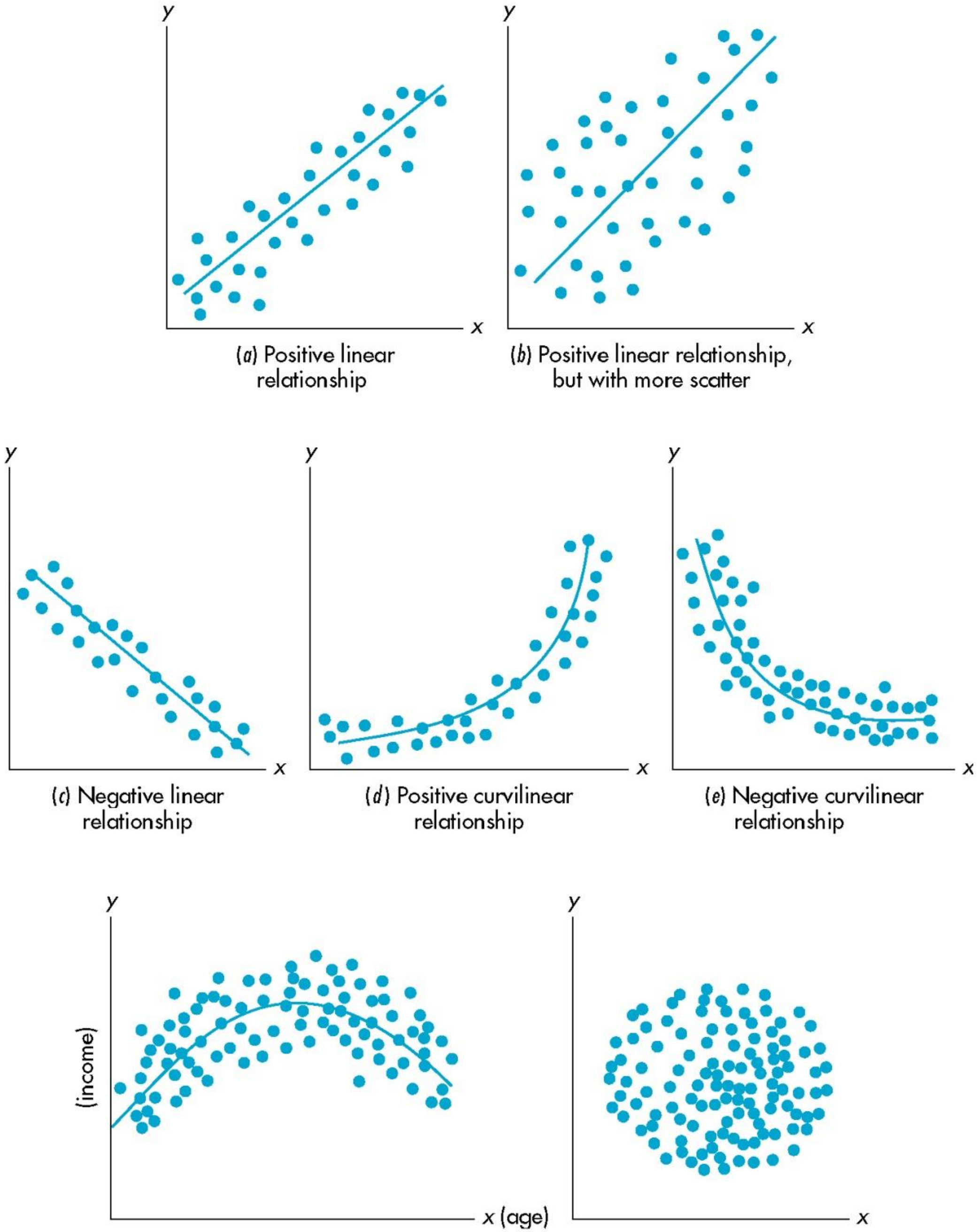
$$y = \beta_0 + \beta_1 x_1 + \varepsilon$$

- The slope coefficient (also called regression coefficient, metric coefficient, etc.)

- It represents the change in y associated with a one-unit increase in x.



y
55
50
45
40
35
30
25
20
15
10
5
0
Output (dozens of units)
F
C
x
1
2
3
4
5
6
7
8
9
10
11
Aptitude test results



(*a*) Positive linear relationship

(*b*) Positive linear relationship, but with more scatter

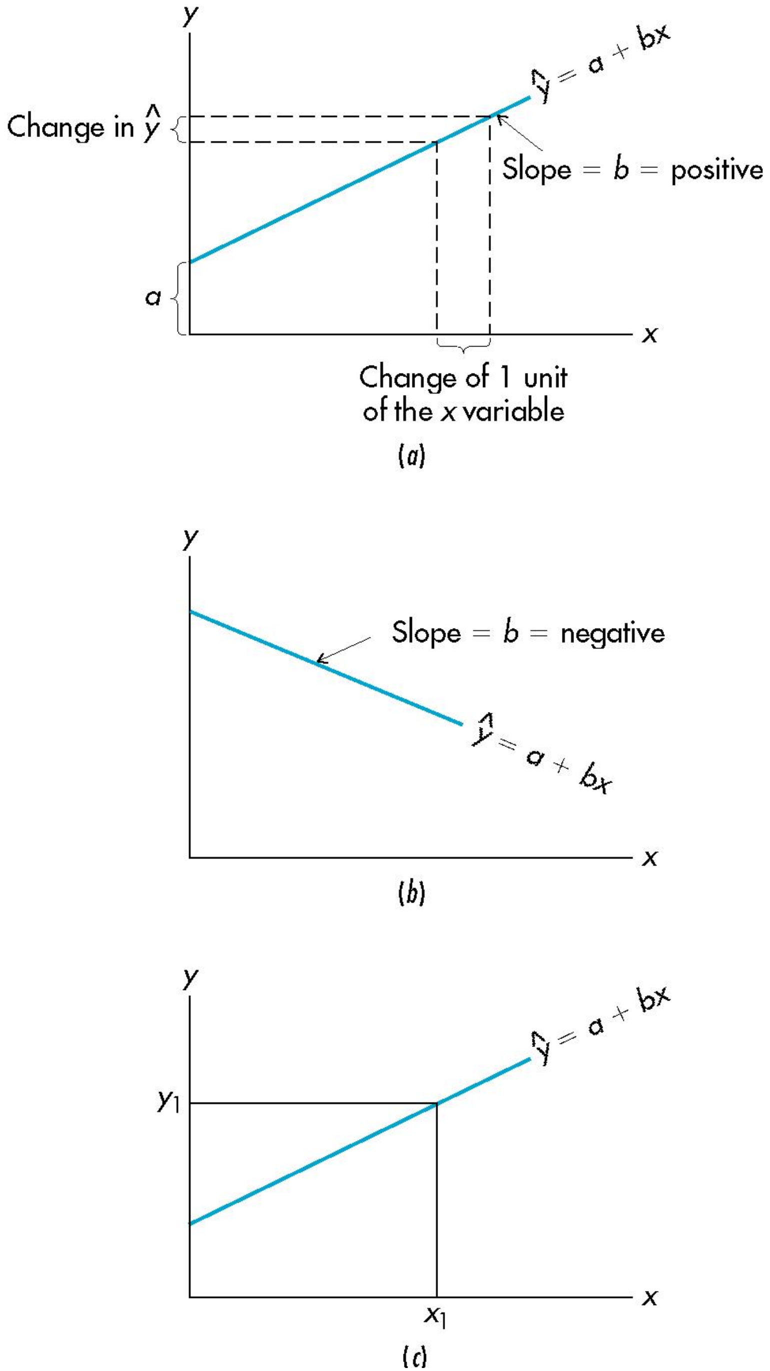
(*c*) Negative linear relationship

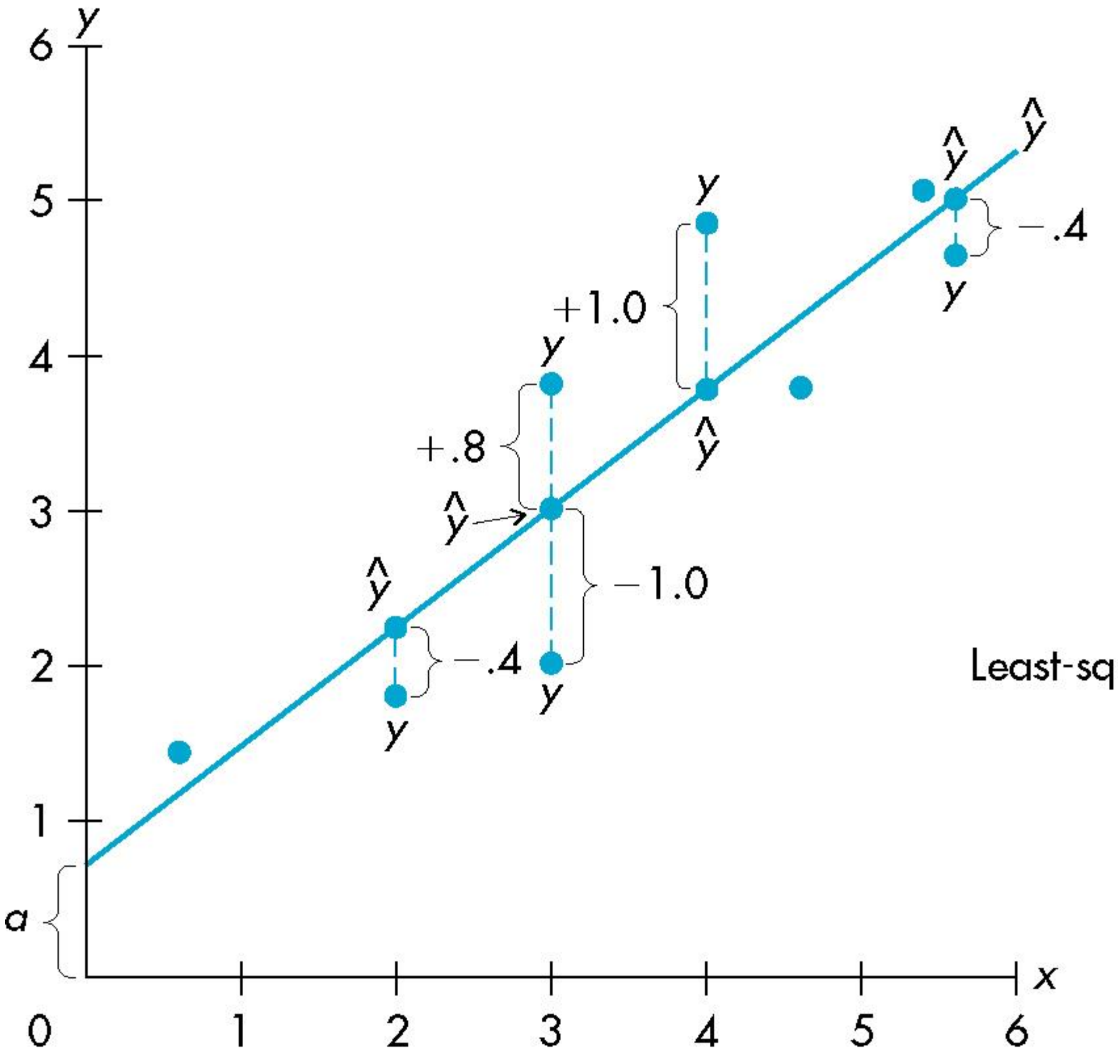
(*d*) Positive curvilinear relationship

(*e*) Negative curvilinear relationship

(*f*) Curvilinear relationship

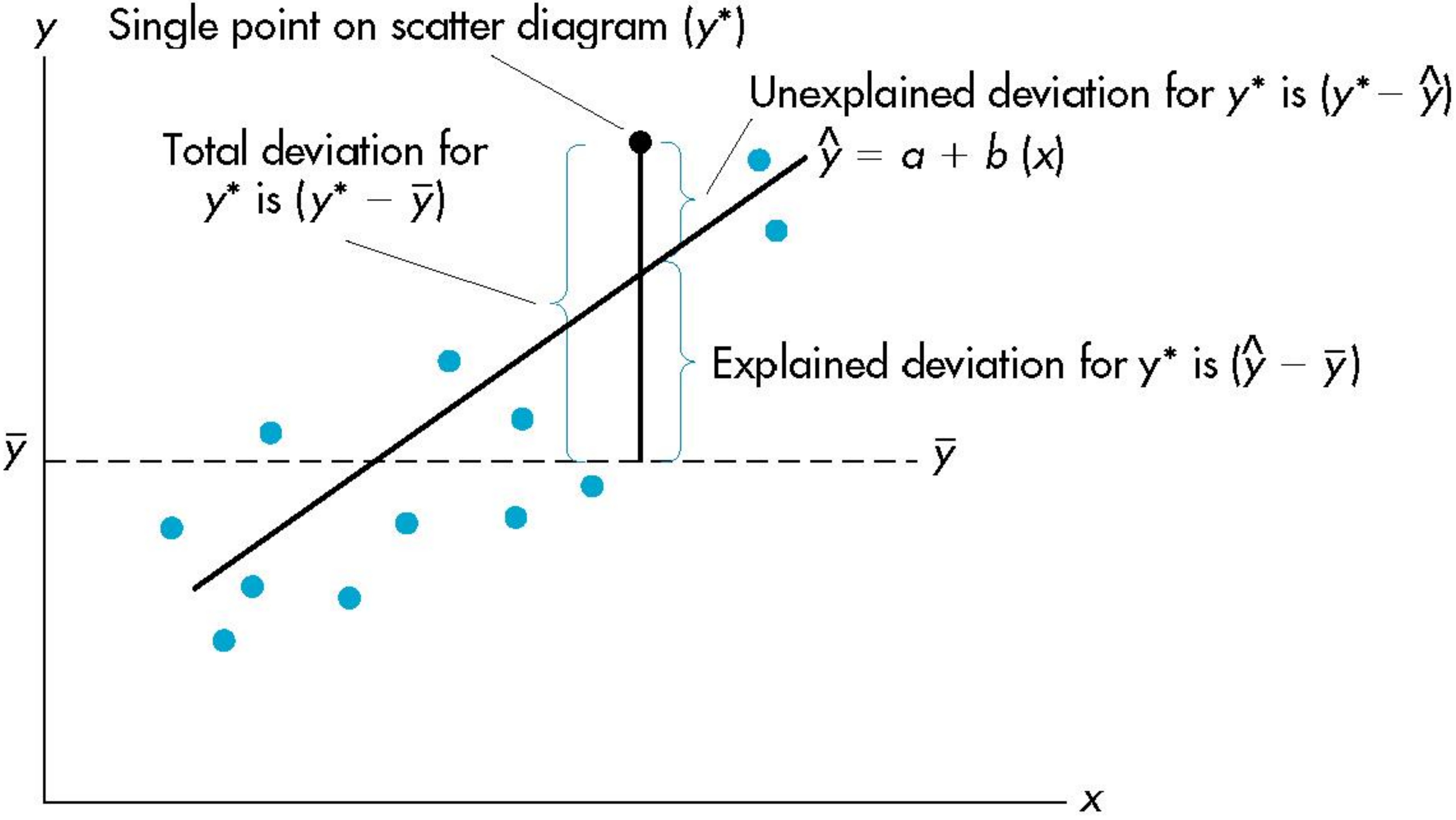
(*g*) No relationship

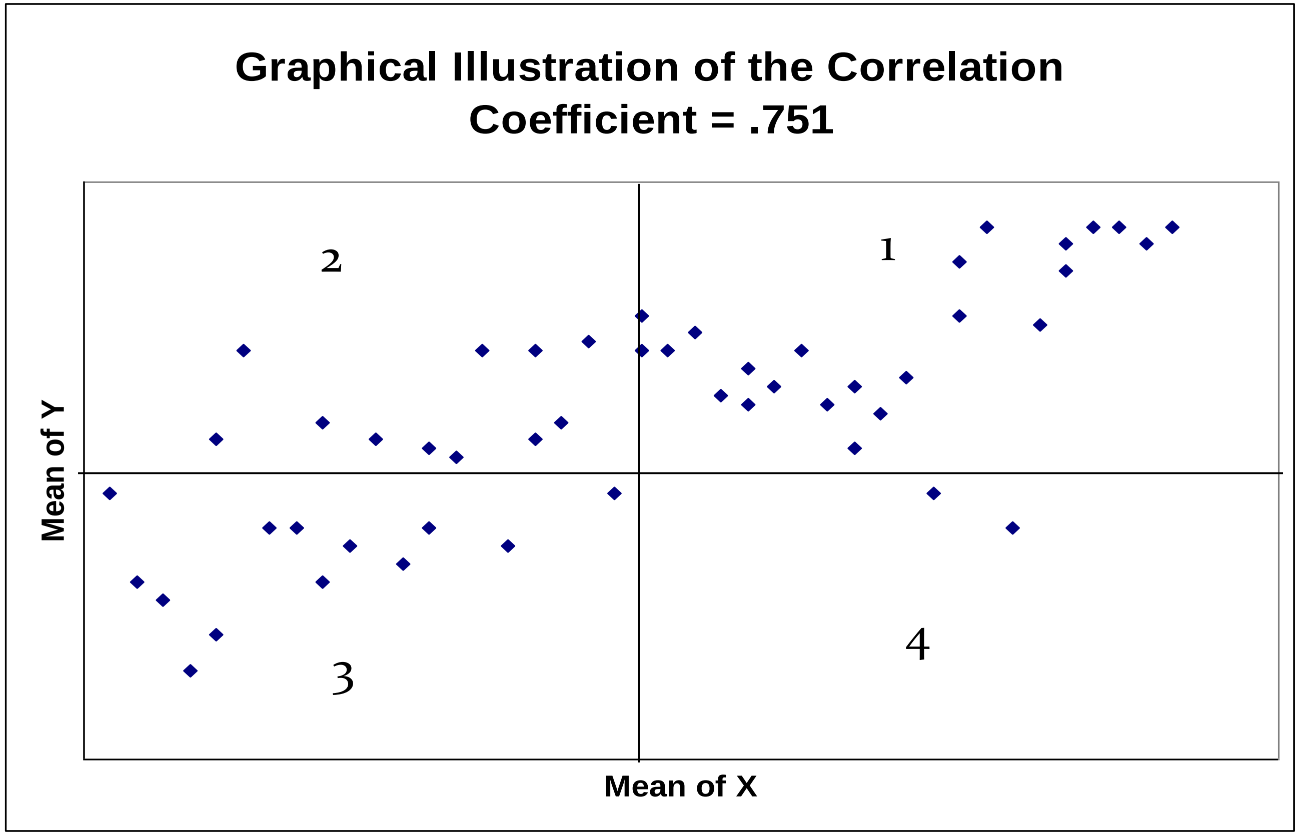

y
$\hat{y} = a + bx$
Change in $\hat{y}$
Slope = b = positive
a
x
Change of 1 unit of the x variable
(a)
y
Slope = b = negative
$\hat{y} = a + bx$
x
(b)
y
$\hat{y} = a + bx$
y_1
x
x_1
(c)



Least-squares properties: $\Sigma(y - \hat{y}) = 0$

$\Sigma(y - \hat{y})^2$ = minimum or
"least" value


y
Single point on scatter diagram (y^*)
Unexplained deviation for y^* is $(y^* - \hat{y})$
Total deviation for y^* is $(y^* - \bar{y})$
$\hat{y} = a + b\,(x)$
Explained deviation for y* is $(\hat{y} - \bar{y})$
$\bar{y}$
$\bar{y}$
x



Quadrant	$y_i - \overline{y}$	$x_i - \overline{x}$	$(y_i - \overline{y})(x_i - \overline{x})$
1	+	+	+
2	+	-	-
3	-	-	+
4	-	+	-

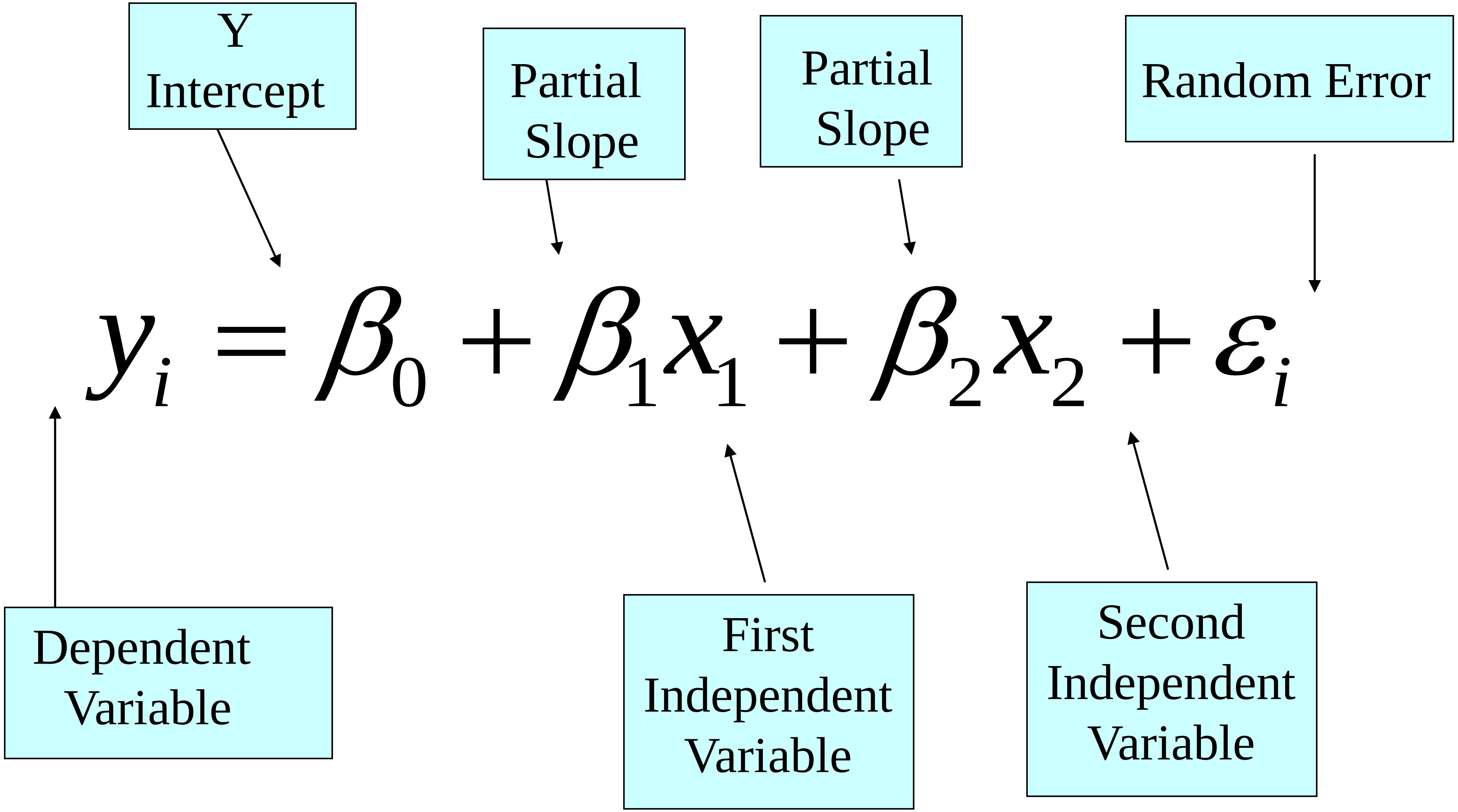
Example: The Problem

Educ (years)	Wage (000)
24	90
20	85
18	75
16	50
12	30

$$(1.2)\quad wage = \beta_0 + \beta_1 educ + \varepsilon$$

Advanced Topics In Regression

The Regression Model- Two or More Independent Variables



Y Intercept

Partial Slope

Partial Slope

Random Error

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i$$

Dependent Variable

First Independent Variable

Second Independent Variable

BLUE
Best Linear Unbiased Estimator - Part 1

R1. Linear In Parameters

R2. Random Sampling

R3. All variables are measured without error

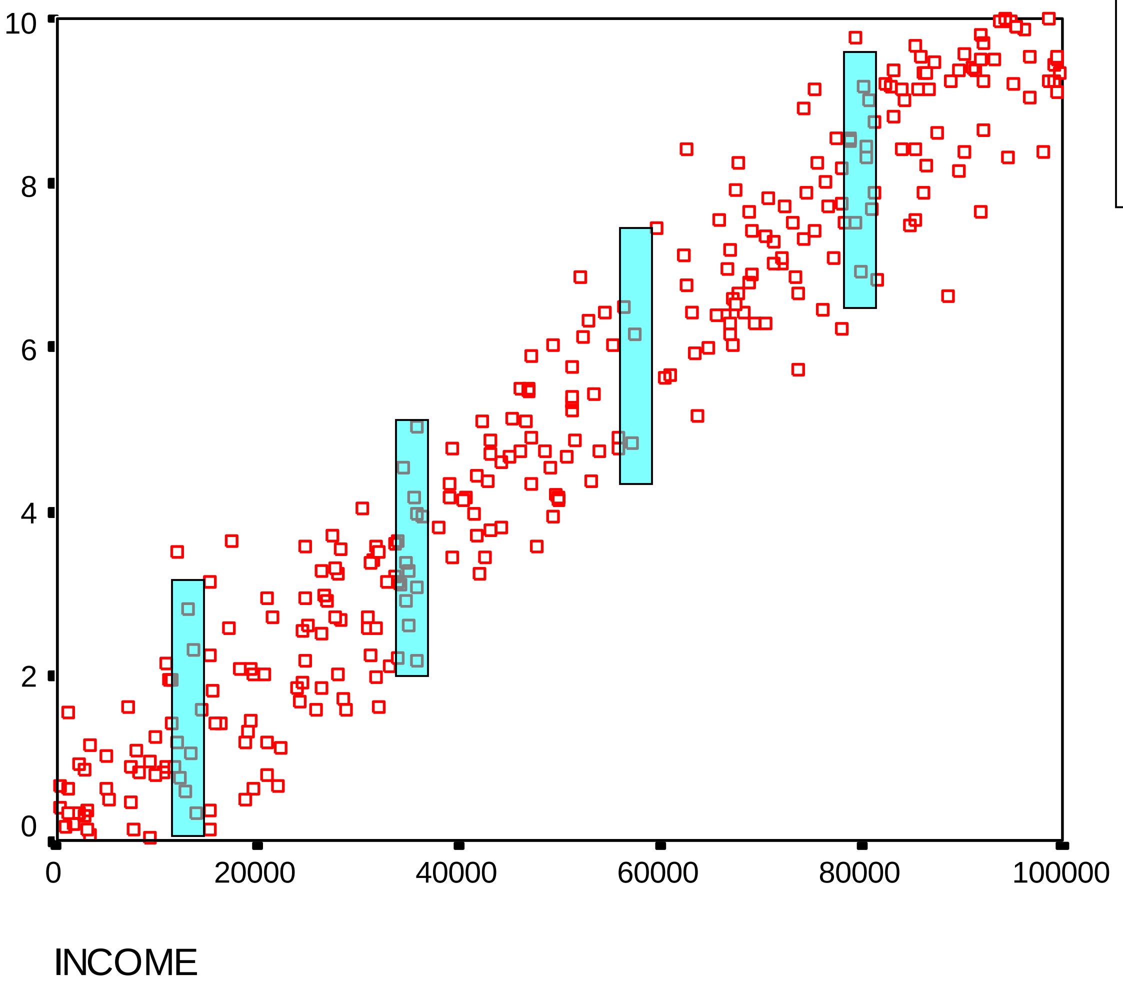
R.4 Zero Conditional Mean

R5. No Perfect Collinearity

R6. Homoskedasticity

Regression Assumptions

- Homoskedasticity: Equal Error Variance

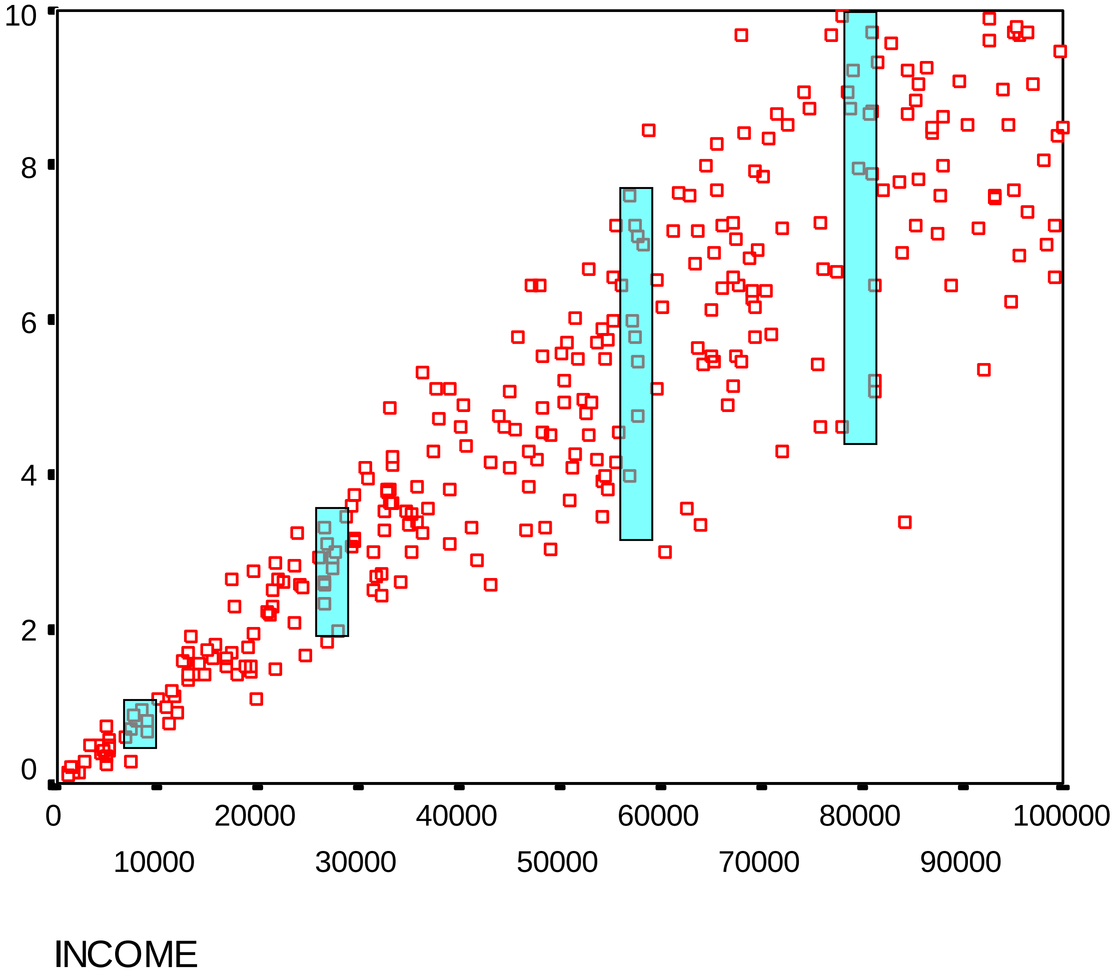


Examine error at different values of X. Is it roughly equal?

Here, things look pretty good.

Regression Assumptions

- Heteroskedasticity: Unequal Error Variance



At higher values of X, error variance increases a lot.

This looks pretty bad.

What is multicollinearity?

- When two IVs are highly correlated, they both convey essentially the same information.

- In this case, neither may contribute significantly to the model after the other one is included. But together they contribute a lot.

- If you removed both variables from the model, the fit would be much worse. So the overall model fits the data well, but neither IV makes a significant contribution when it is added to the model last.

- When this happens, the IV variables are *collinear* and the results show *multicollinearity*.

Why is multicollinearity a problem?

- If your goal is simply to predict Y from a set of X variables, then multicollinearity is not a problem.
  - The predictions will still be accurate, and the overall R^2 (or adjusted R^2) quantifies how well the model predicts the Y values.
- If your goal is to understand how the various X variables impact Y, then multicollinearity is a big problem.

	Impact						
Violation	F	R-square	β	Std Error of Estimate	Std Error of β	T	Number of Violations
Measurement Error in Dependent variable	👍	👍	👍	👍	☹↑	☹↓	2
Measurement Error in Independent variable	☹	☹	☹	☹	☹	☹	6
Irrelevant Variable	👍	👍	👍	👍	☹↑	☹↓	2
Omitted variable	☹	☹	☹	☹	☹	☹	6
Incorrect functional form	☹	☹	☹	☹	☹	☹	6
Heteroskedasticity	☹	👍	👍	☹	☹	☹	4
Collinearity	👍	👍	👍	👍	☹↑	☹↓	2
Simultaneity Bias	☹	☹	☹	☹	☹	☹	6

Legend for Table

👍= The statistic is still reliable and unbiased.

☹ = The statistic is biased, and thus cannot be relied upon.

↓ = Downward bias in estimation.

↑ = Upward bias in estimation.

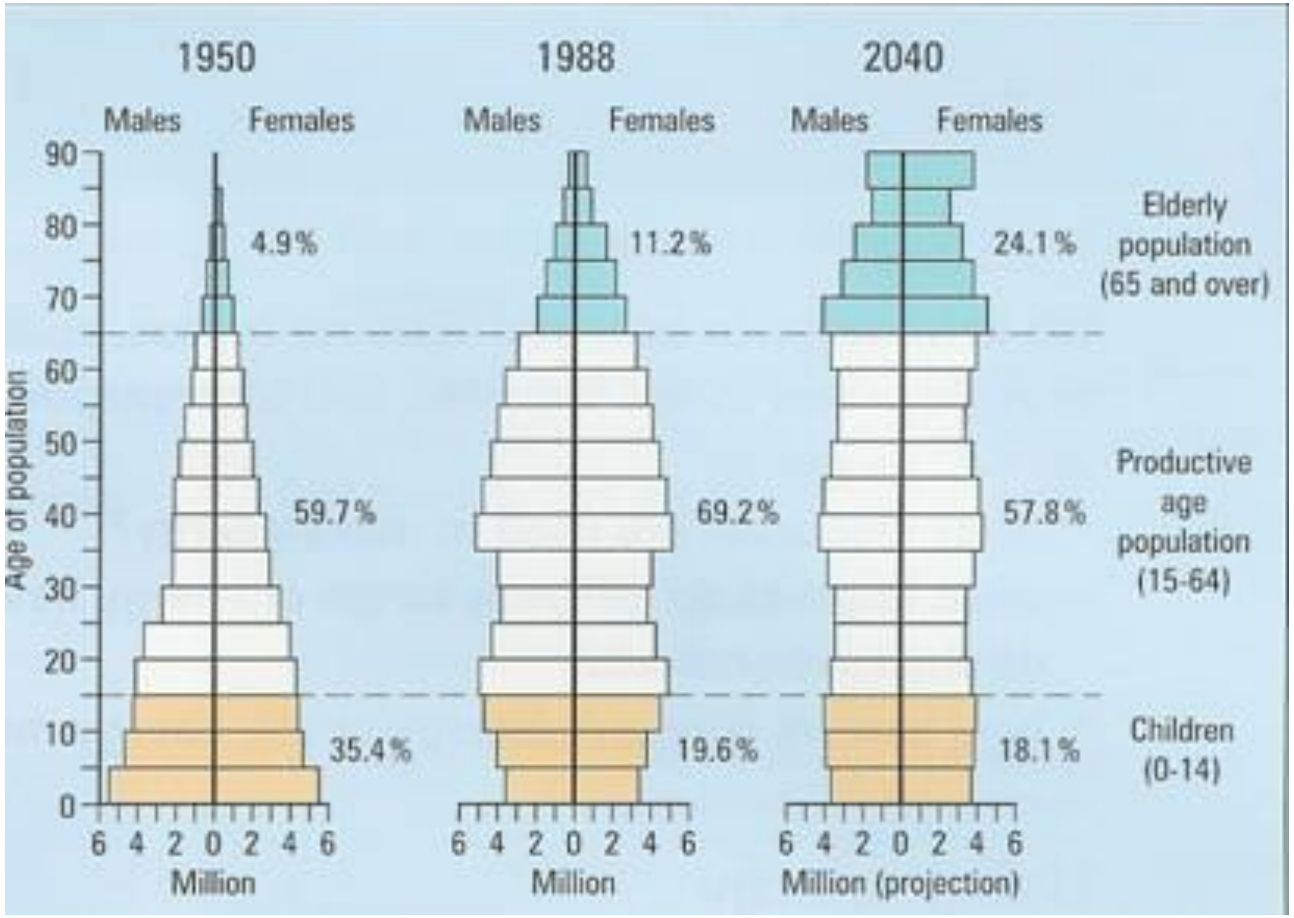
Review of Basic Concepts in Demography

What is Demography?

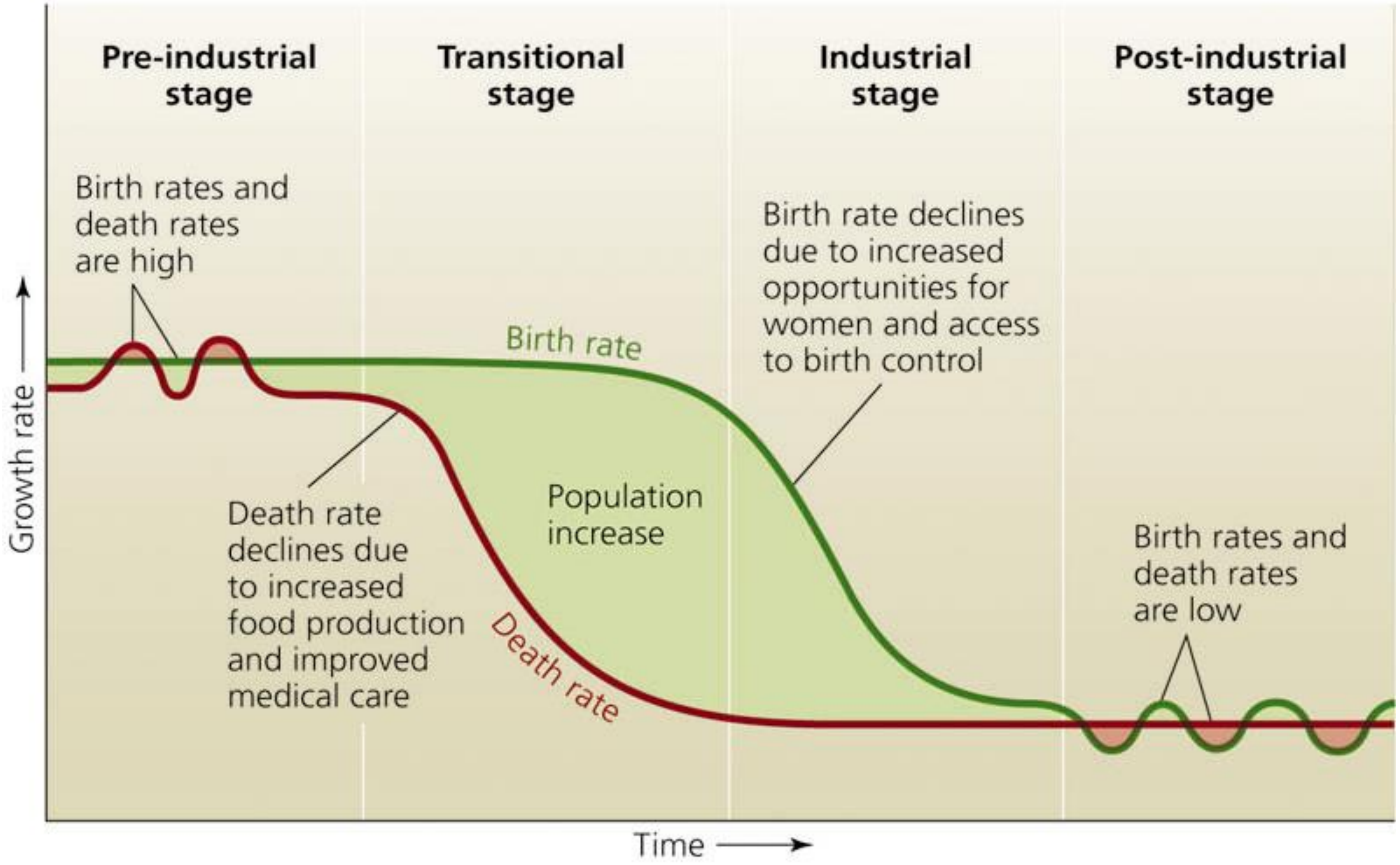
- Demography is the scientific study of human populations.
  - 1855 (Achille Guillard)
    - Demos – people
    - Graphein – write about a particular subject

What is Demography?

Mathematical knowledge of **populations**, their general movement of the **population**, and their physical, civil, intellectual and moral state.

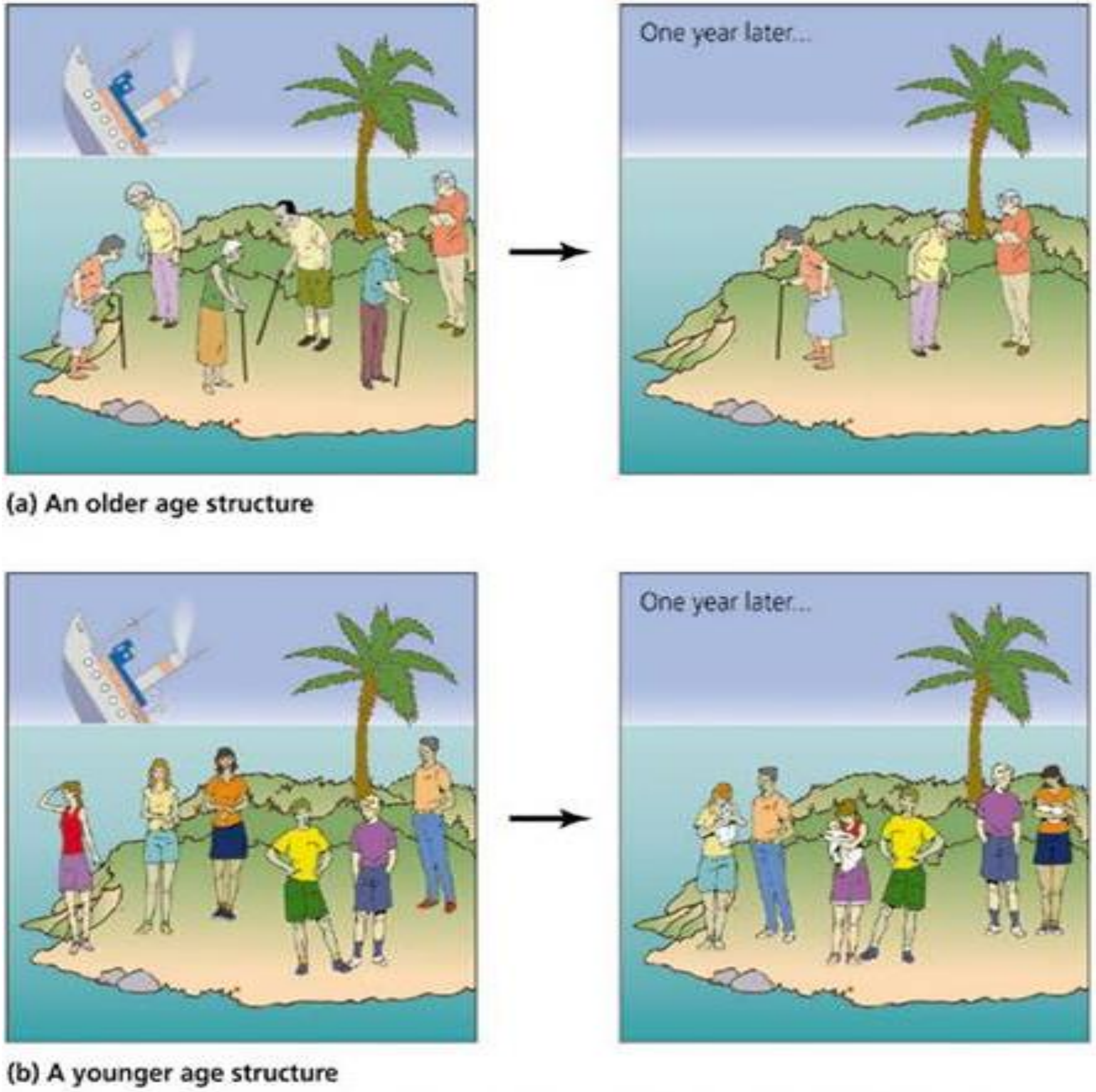


Demographic transition: Stages



Age structure

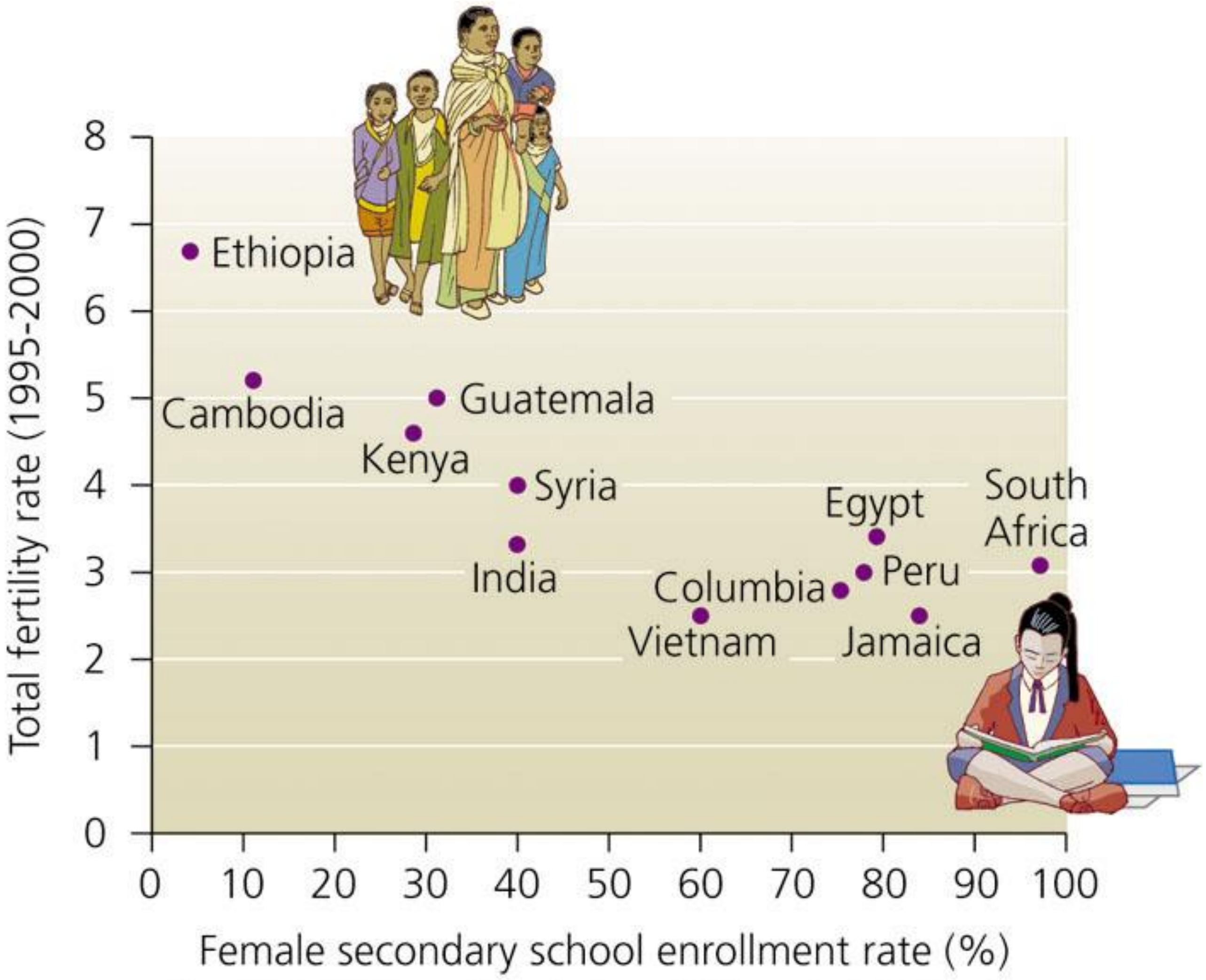
Age structure can influence population growth rates.



(a) An older age structure

(b) A younger age structure

Female education and TFR



Female literacy and school enrollment are correlated with total fertility rate:

More-educated women have fewer children.

Ecological footprints

Residents of some countries consume more resources—and thus use more land—than residents of others.

Shown are **ecological footprints** of an average citizen from various nations.

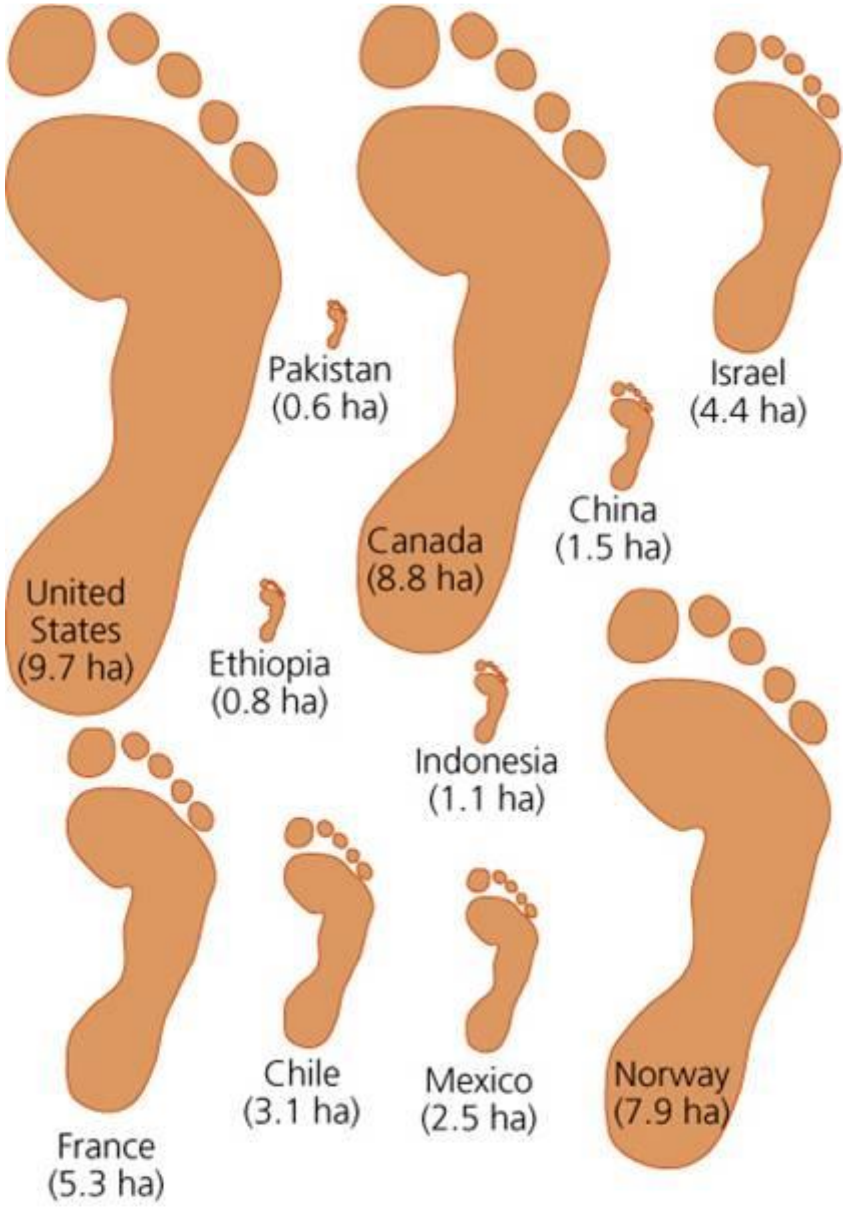
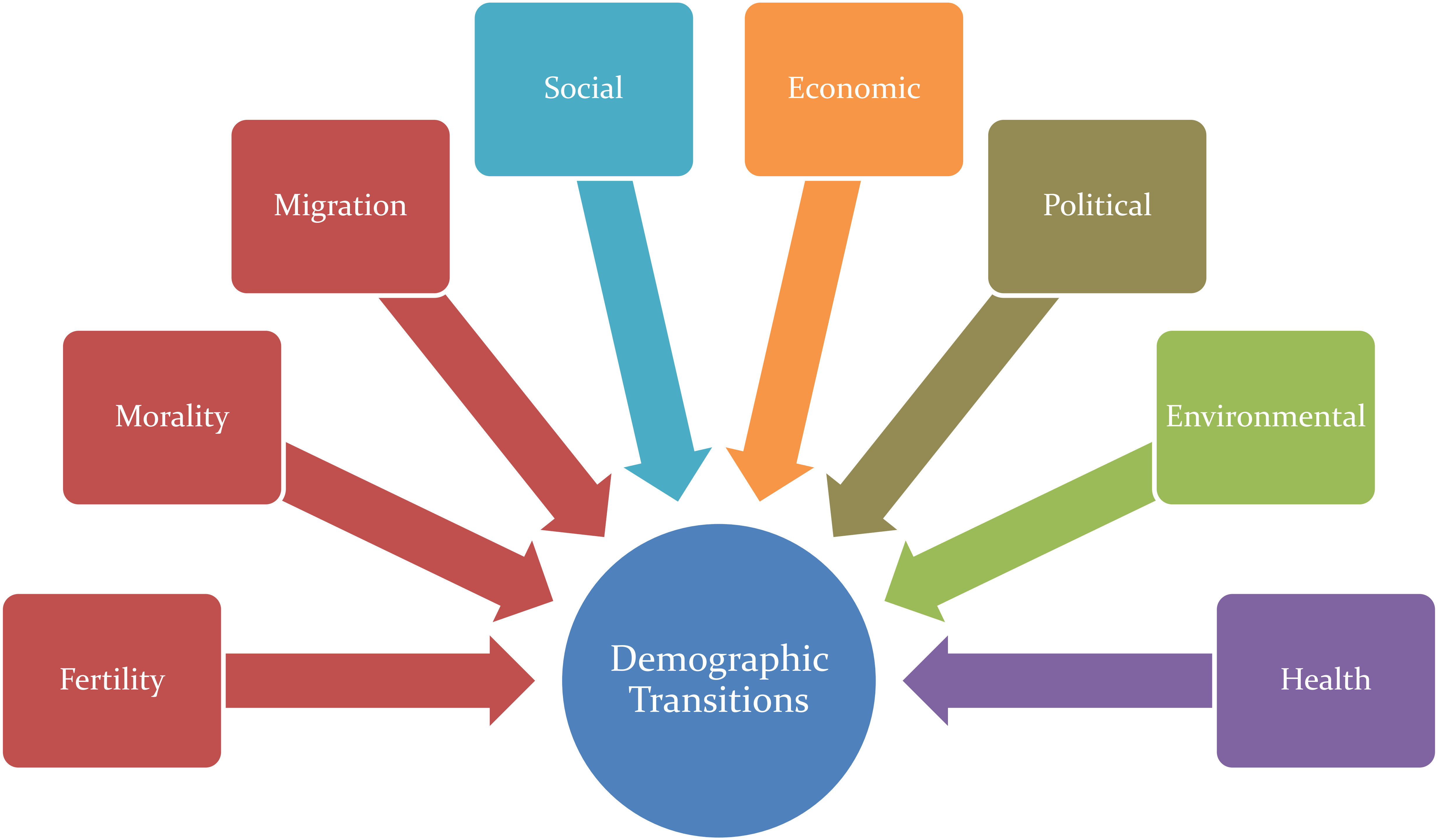


Figure 7.23

The “IPAT” model

Shows how **P**opulation, **A**ffluence, and **T**echnology interact to create **I**mpact on our environment.

$$\mathbf{I} = \mathbf{P} \times \mathbf{A} \times \mathbf{T}$$



Fertility

Morality

Migration

Social

Economic

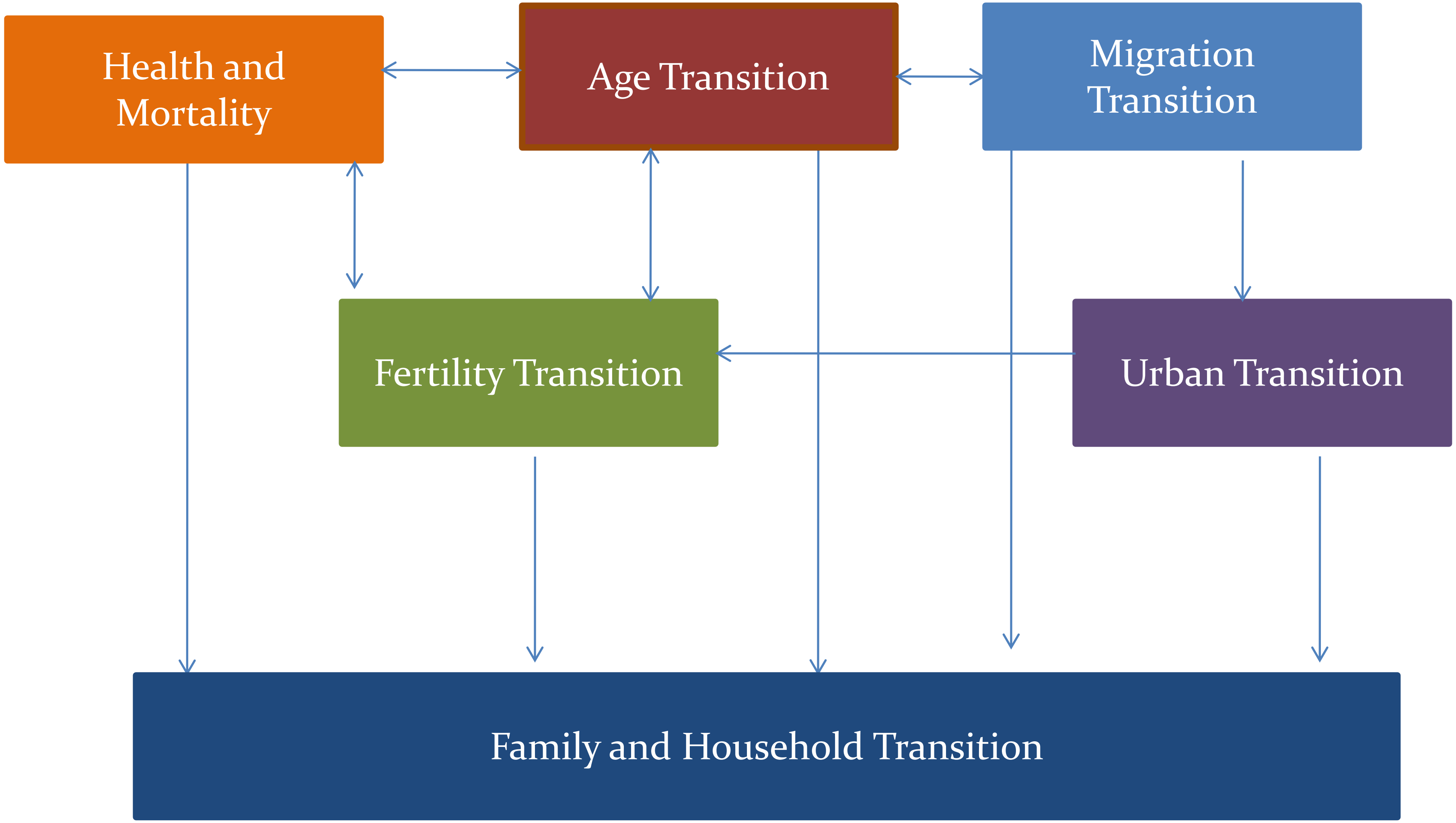
Political

Environmental

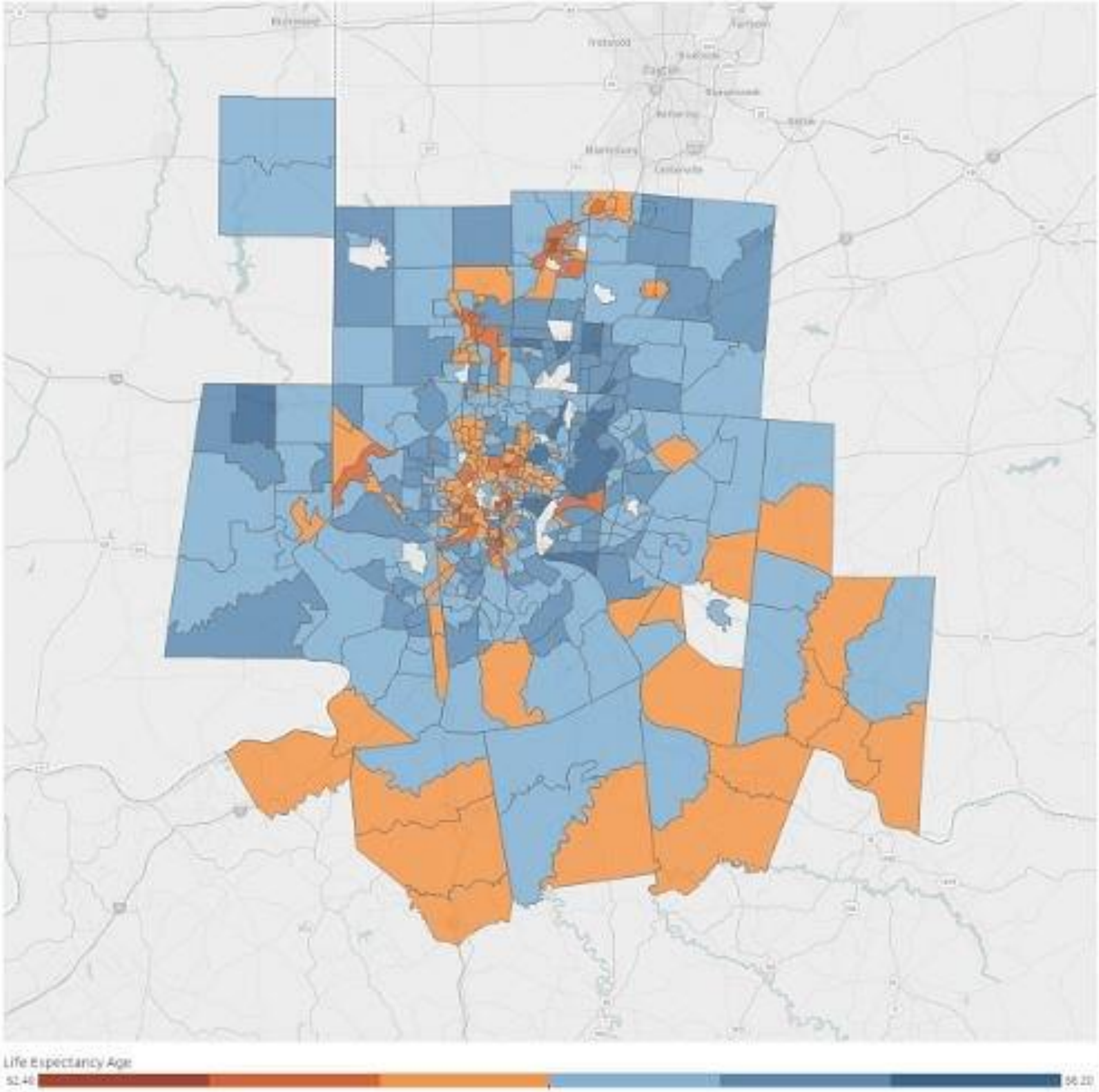
Health

Demographic Transitions

Decomposition of the Demographic Transition Model

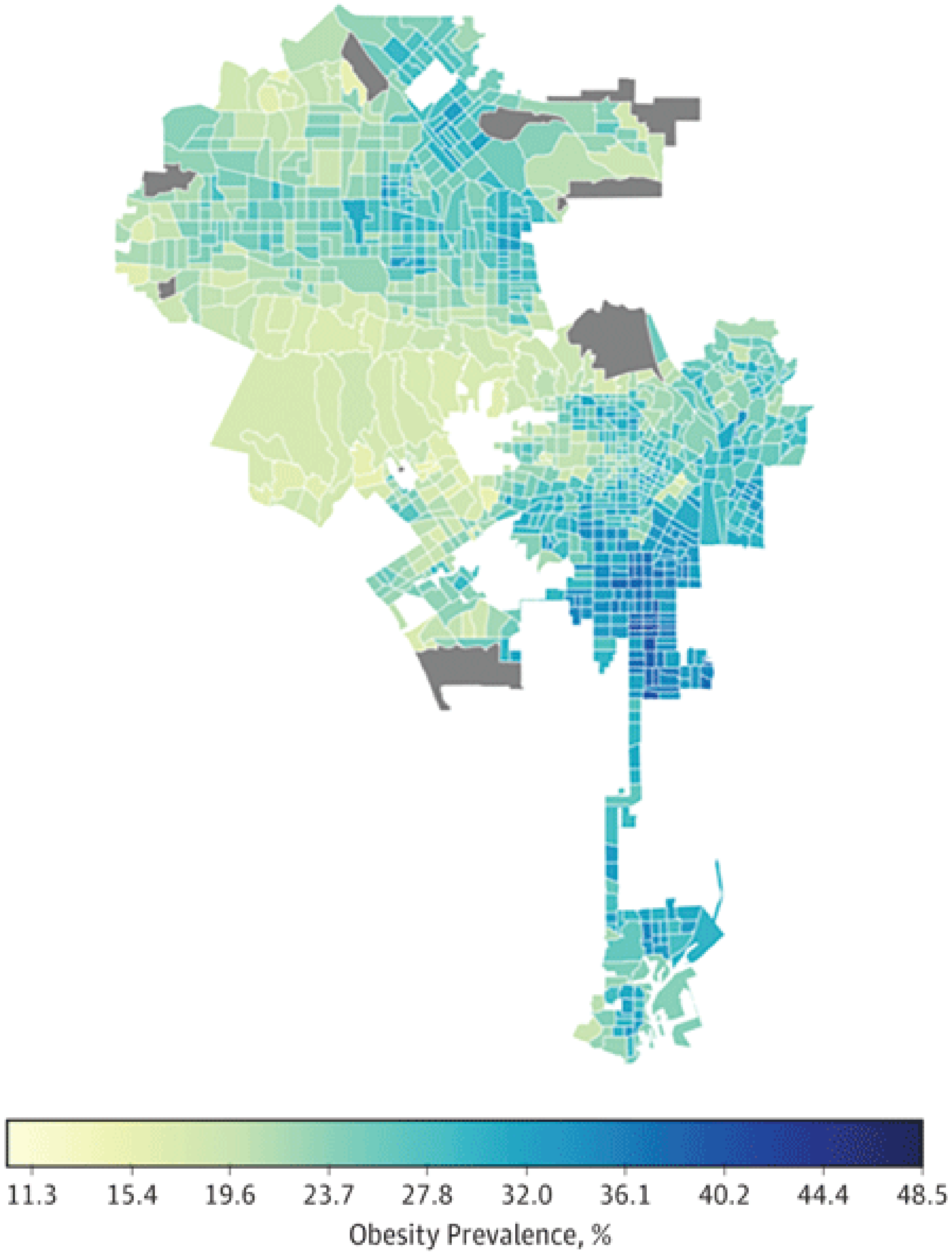


Example 1
Life Expectancy



Source: National Center for Health Statistics. U.S. Small-Area Life Expectancy Estimates Project (USALEEP): Life Expectancy Estimates File for Ohio, Kentucky, and Indiana, 2010-2015. National Center for Health Statistics. 2018. Available from: https://www.cdc.gov/nchs/nvss/usaleep/usaleep.html

Example 2
Obesity Prevalence


11.3
15.4
19.6
23.7
27.8
32.0
36.1
40.2
44.4
48.5
Obesity Prevalence, %