Mathematics 231

Lecture 7

Liam O'Brien

Announcements

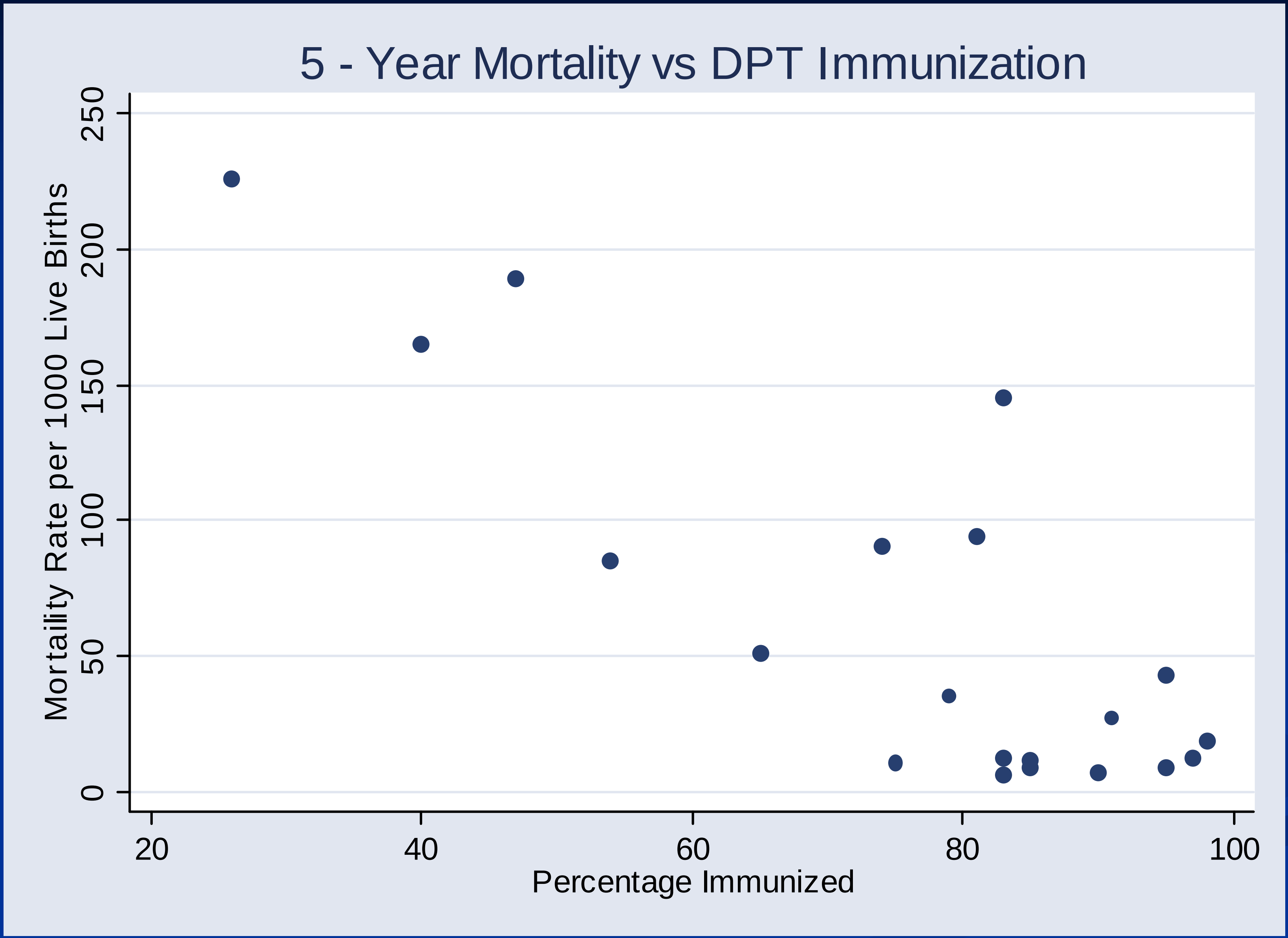
- Reading
  - Today M&M 2.3 108-121
  - Next class M&M 2.3 117-119
    M&M 2.4 125-132

Linear Relationships & Regression

- Linear relationship between two variables.
- Response and explanatory variables.
- Regression line.
- Least squares criterion.

Response and Explanatory Variables

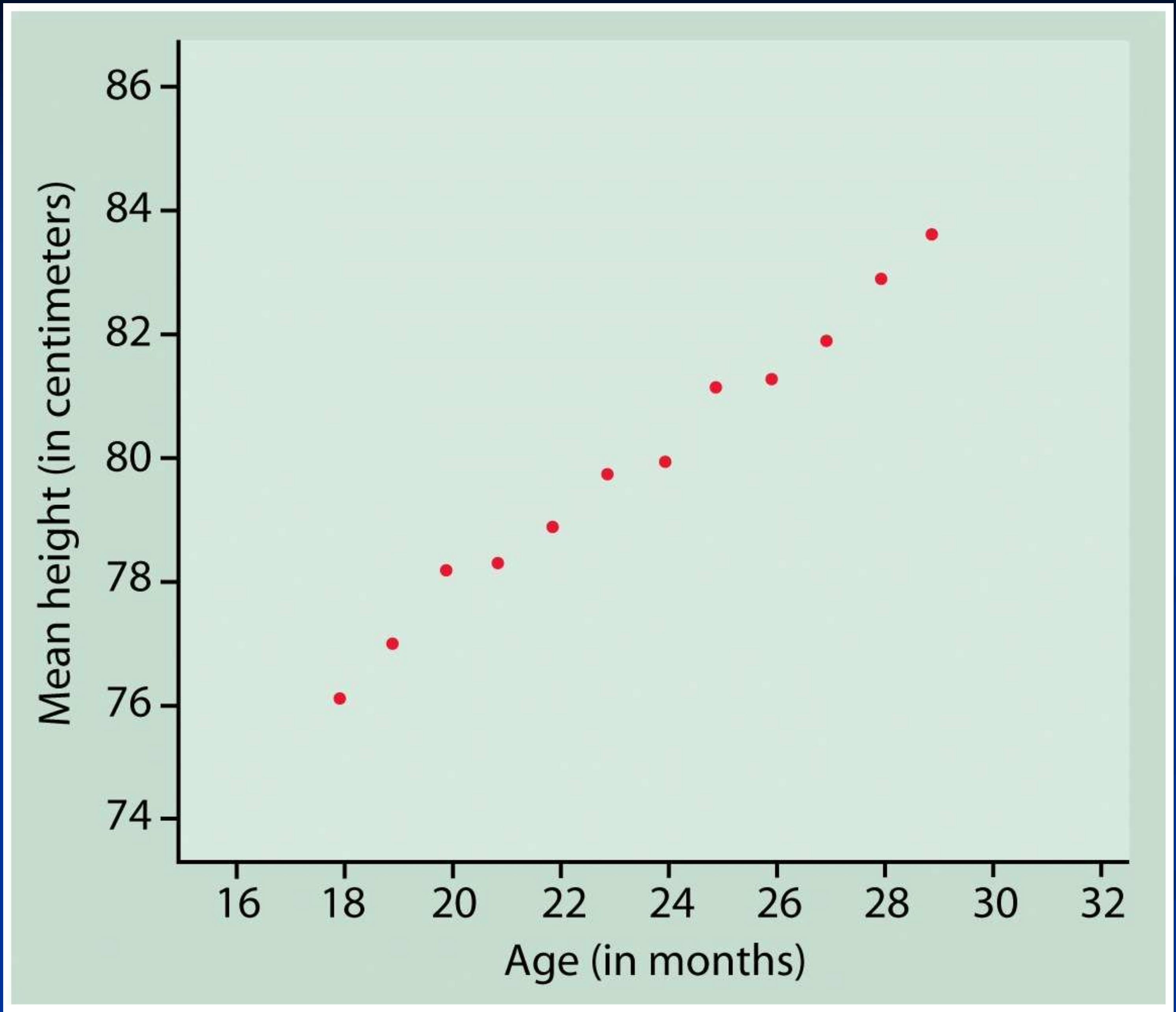
- A **response variable**, denoted as Y, measures the outcome of an experiment, survey, or study. Y is the variable we want to explain or predict.
- An **explanatory variable**, denoted as X, is a variable that may affect, explain or predict (but not necessarily cause) the response variable.


5 - Year Mortality vs DPT Immunization
Mortaility Rate per 1000 Live Births
0
50
100
150
200
250
20
40
60
80
100
Percentage Immunized

Example: Height and Age

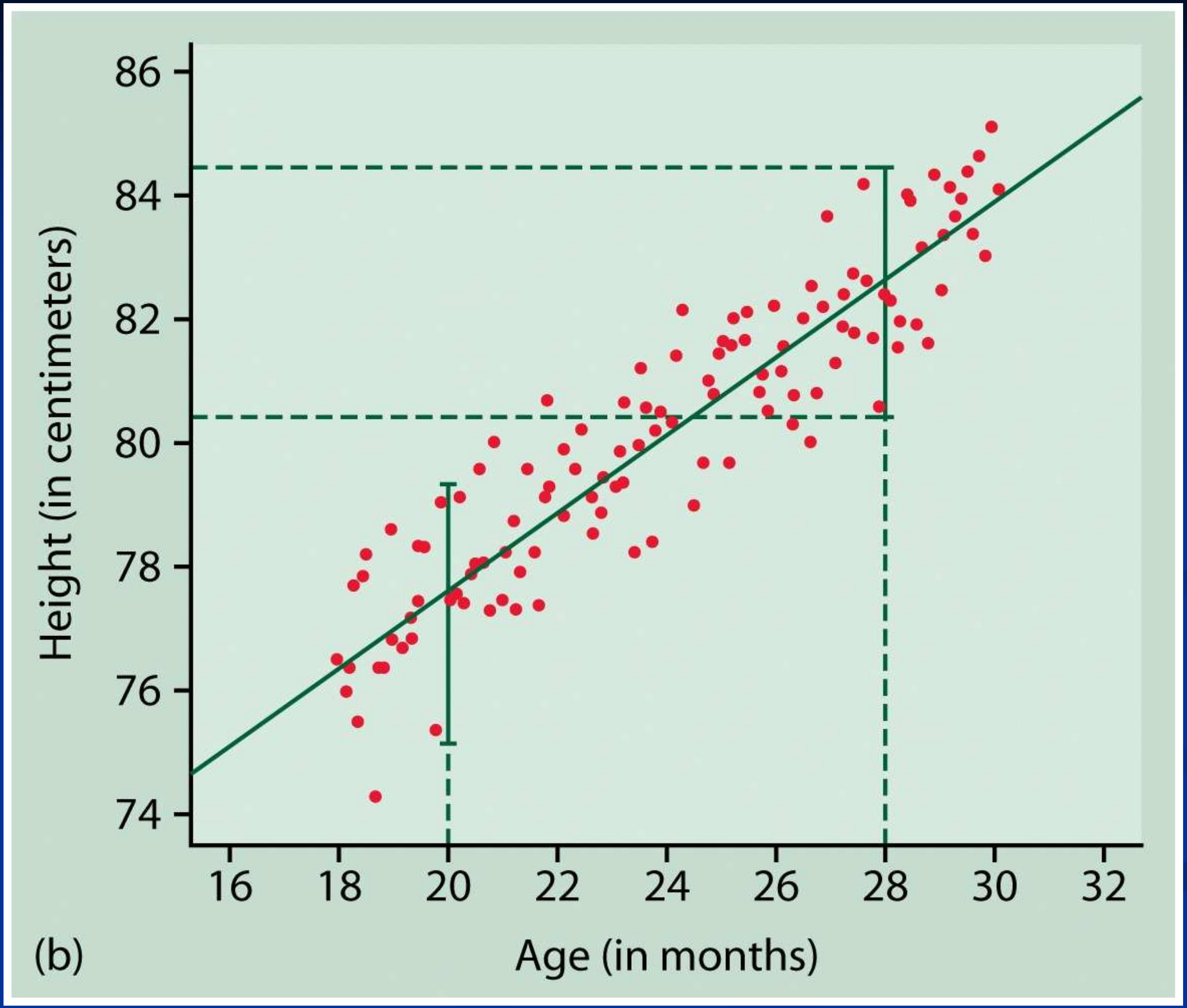
TABLE 2.7 Mean height of Kalama children

Age x in months	Height y in centimeters
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5


Mean height (in centimeters)
74
76
78
80
82
84
86
16
18
20
22
24
26
28
30
32
Age (in months)

Conditional Distributions

- In general, we can consider the distribution of Y variables (e.g., height) for observations that satisfy some condition $X = x$ (e.g., age equals 28 months).
- This is called the **conditional distribution of Y given $X = x$.**
- In a scatter plot, the conditional distribution of Y given $X = x$ is the distribution of points in the vertical strip above a given value of x.


86
84
82
80
78
76
74
Height (in centimeters)
16
18
20
22
24
26
28
30
32
Age (in months)
(b)

Conditional Mean

- Conditional distributions have center, spread, and shape properties like all distributions.

- The mean value of Y in the vertical strip above a given value x is called the **conditional mean of *Y* given *X* = *x***.

Linear Regression

- Linear regression is used to explain or predict *Y* using *X*.
- It quantifies the relationship between the two variables in terms of a straight line.
- Suppose we have *n* pairs of *Y* and *X*,

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n y_n)$$

- How can we find the straight line that best "fits" or describes these data?

Linear Regression

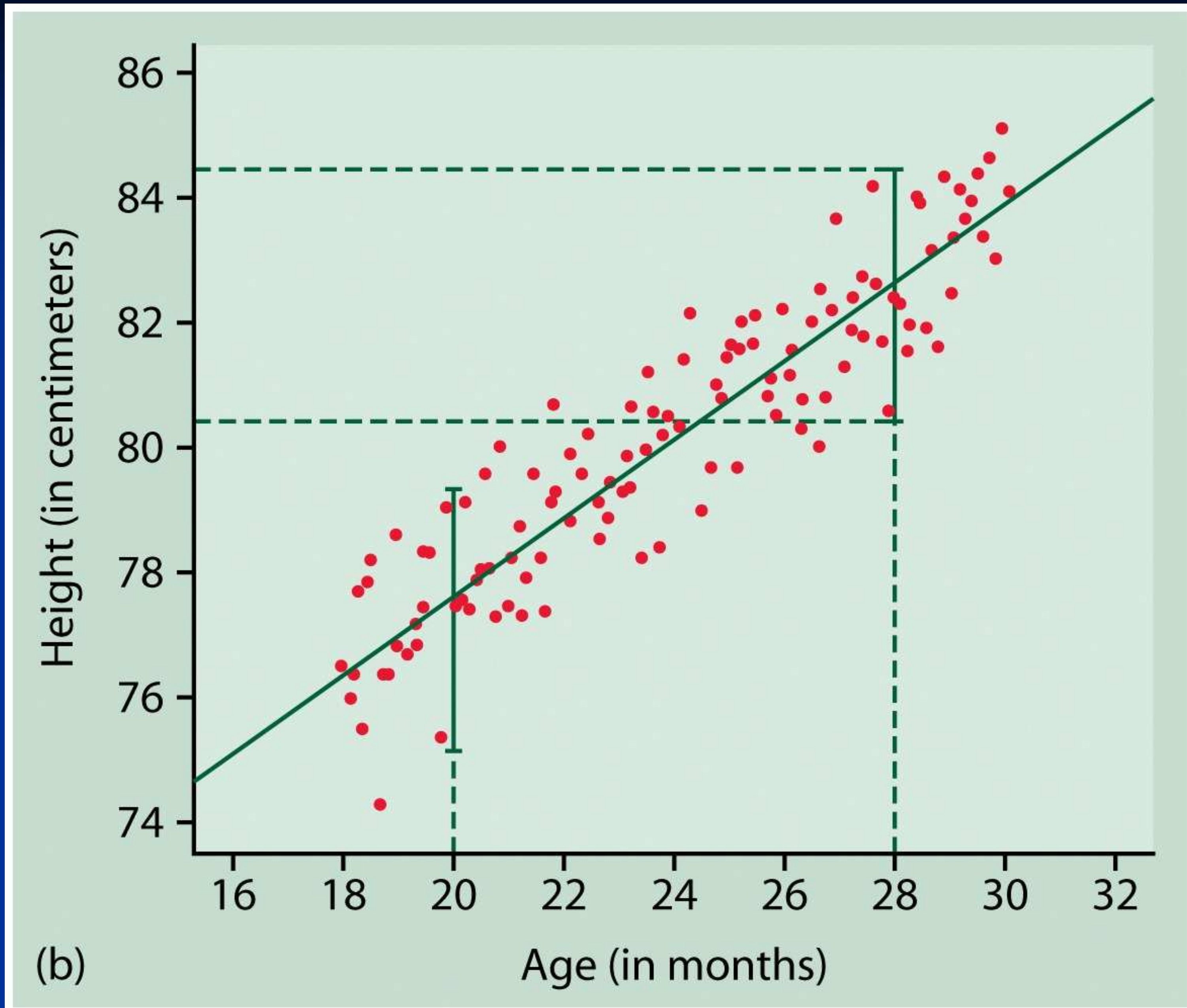
- This line has an equation of the form:

$$\hat{y}_i = a + bx_i$$

where $\hat{y}_i$ (y-hat) is the predicted value of Y,

a is the y-intercept (the value of Y when $X = 0$),

and b is the slope of the line.

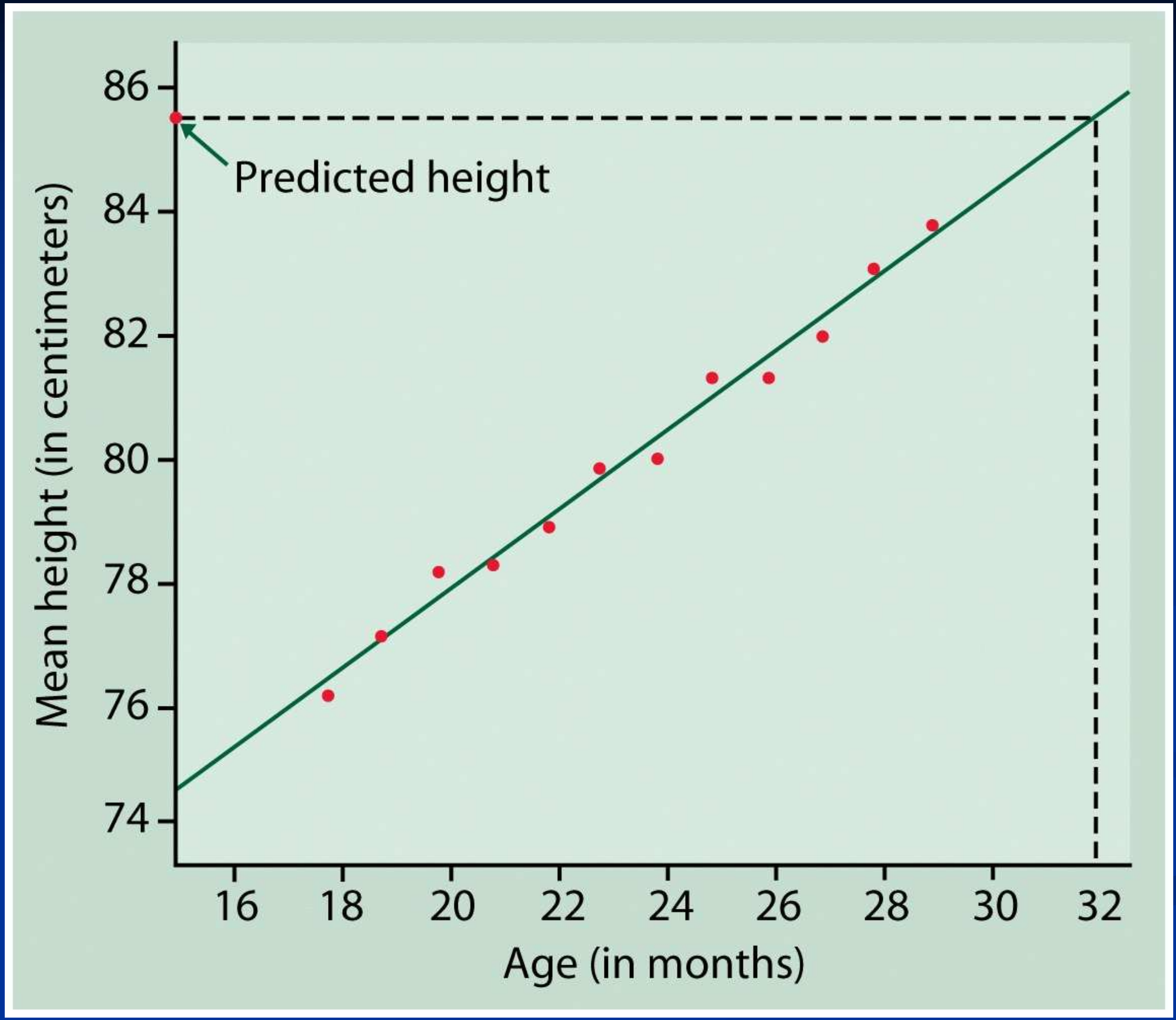

86
84
82
80
78
76
74
Height (in centimeters)
16
18
20
22
24
26
28
30
32
Age (in months)
(b)

Definition 1

- For any particular value, x_i, the **predicted (or fitted)** value is:

$$\hat{y}_i = a + bx_i$$

and is the y-value of the line at x_i.



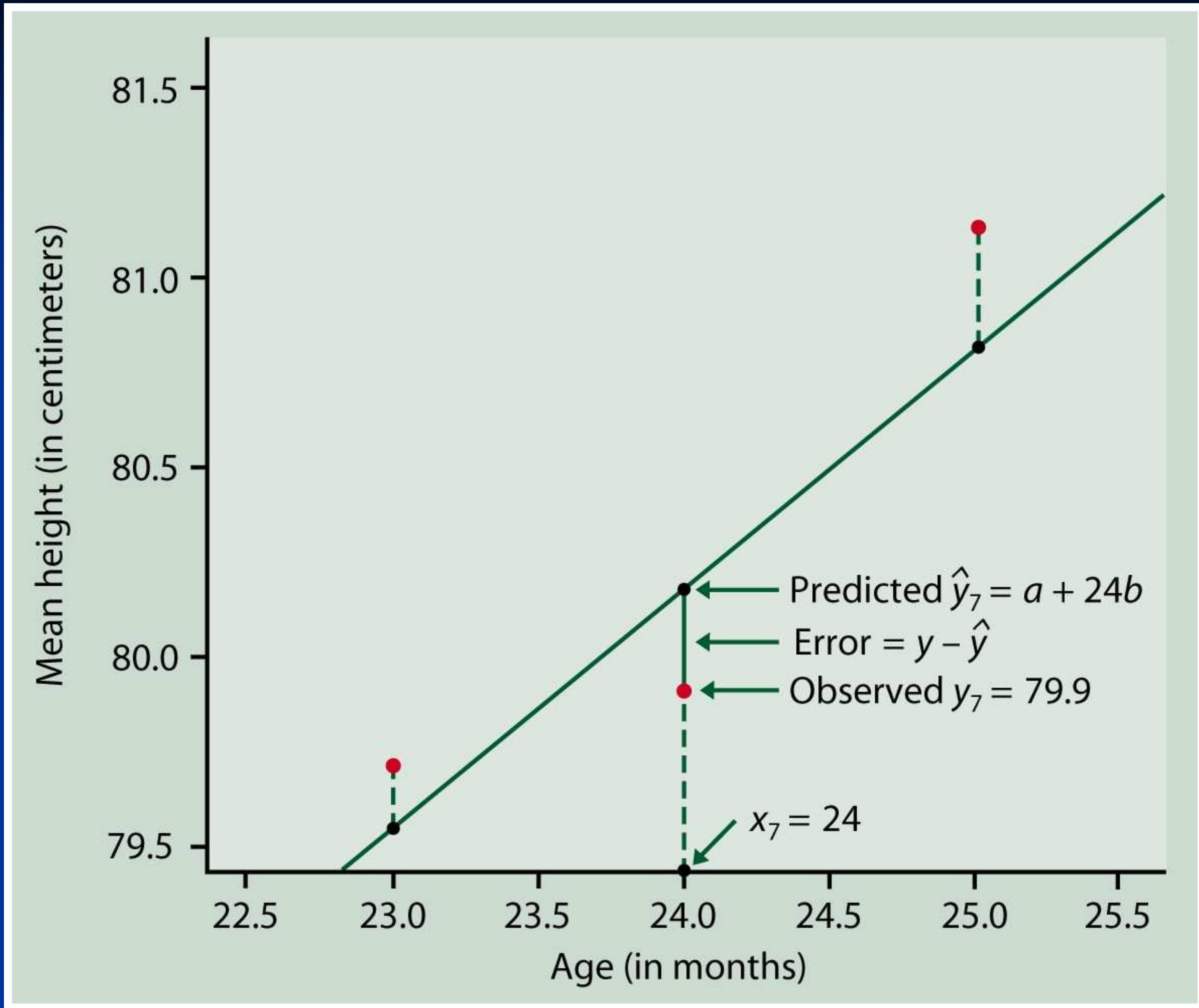
Mean height (in centimeters)
86
84
82
80
78
76
74
Predicted height
16
18
20
22
24
26
28
30
32
Age (in months)

Definition 2

- The vertical deviation from a point to the line (the difference between the observed and predicted values of *Y*, or the error) is called the **residual**.

$$y_i - \hat{y}_i = y_i - (a + bx_i)$$

$$\text{residual} = \varepsilon_i = \text{observed } y_i - \text{ predicted } y_i$$



Mean height (in centimeters)
81.5
81.0
80.5
80.0
79.5
22.5
23.0
23.5
24.0
24.5
25.0
25.5
Age (in months)
Predicted $\hat{y}_7 = a + 24b$
Error = $y - \hat{y}$
Observed $y_7 = 79.9$
$x_7 = 24$

Least Squares Criterion

- The "best fit" line is defined as the line that minimizes the sum of the squared residuals.
- We want the values of *a* and *b* that minimizes the following quantity,

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

Least Squares Intercept and Slope

- The values of *a* and *b* that minimize this quantity are,

$$b = r\frac{s_y}{s_x}$$

$$a = \overline{y} - b\overline{x}$$

where r is the correaltion coefficient.

Assumptions

- The regression line estimates the conditional mean of Y given $X=x$ for any point x if the following assumptions are met.
  1. Conditional mean of Y is a linear function of X.
  2. Conditional SD of Y is constant for all X.
- We often make an additional assumption:
  3. The conditional distribution of Y is a normal distribution for any value of x.

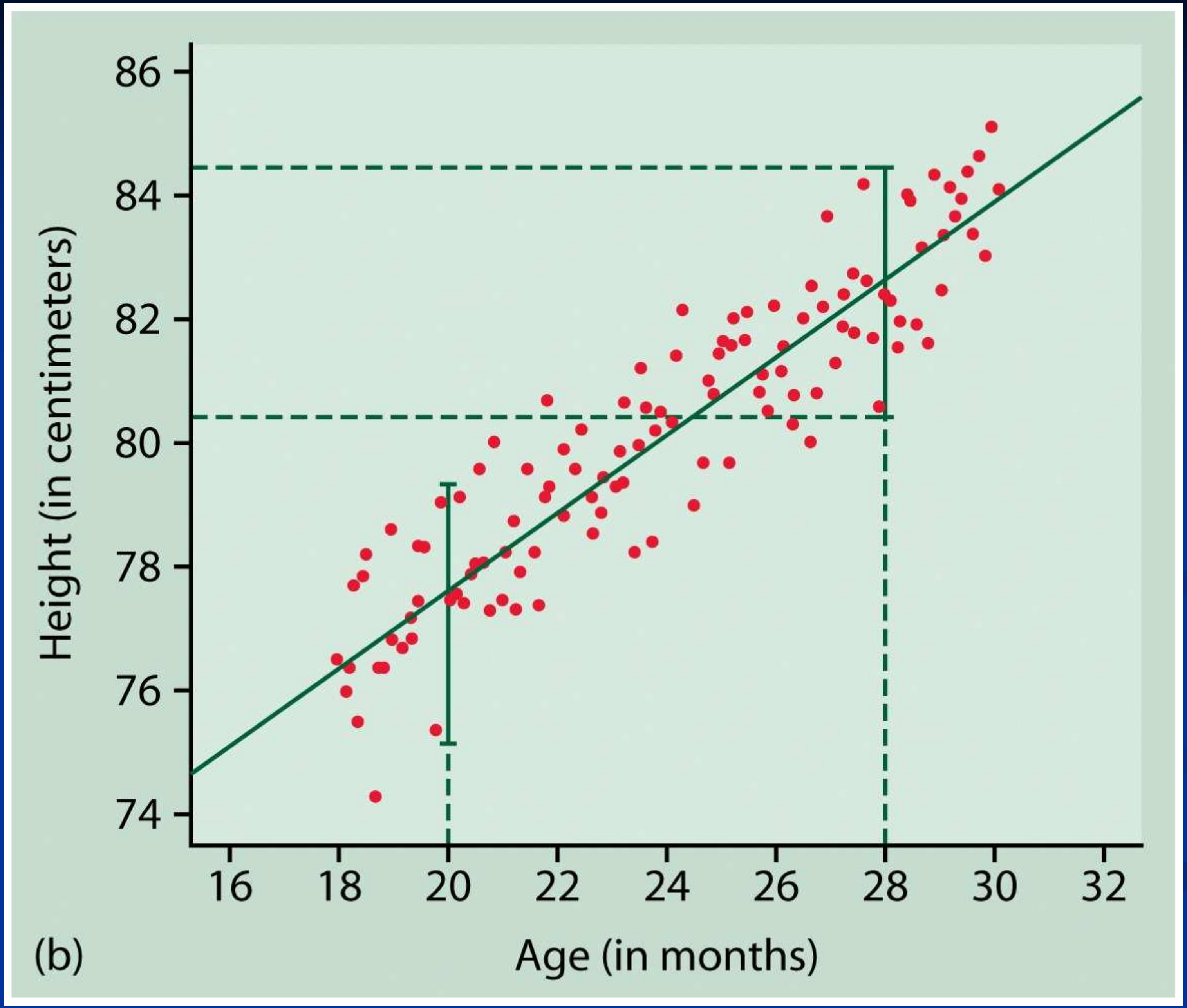

86
84
82
80
78
76
74
Height (in centimeters)
16
18
20
22
24
26
28
30
32
Age (in months)
(b)

TABLE 2.7 Mean height of Kalama children

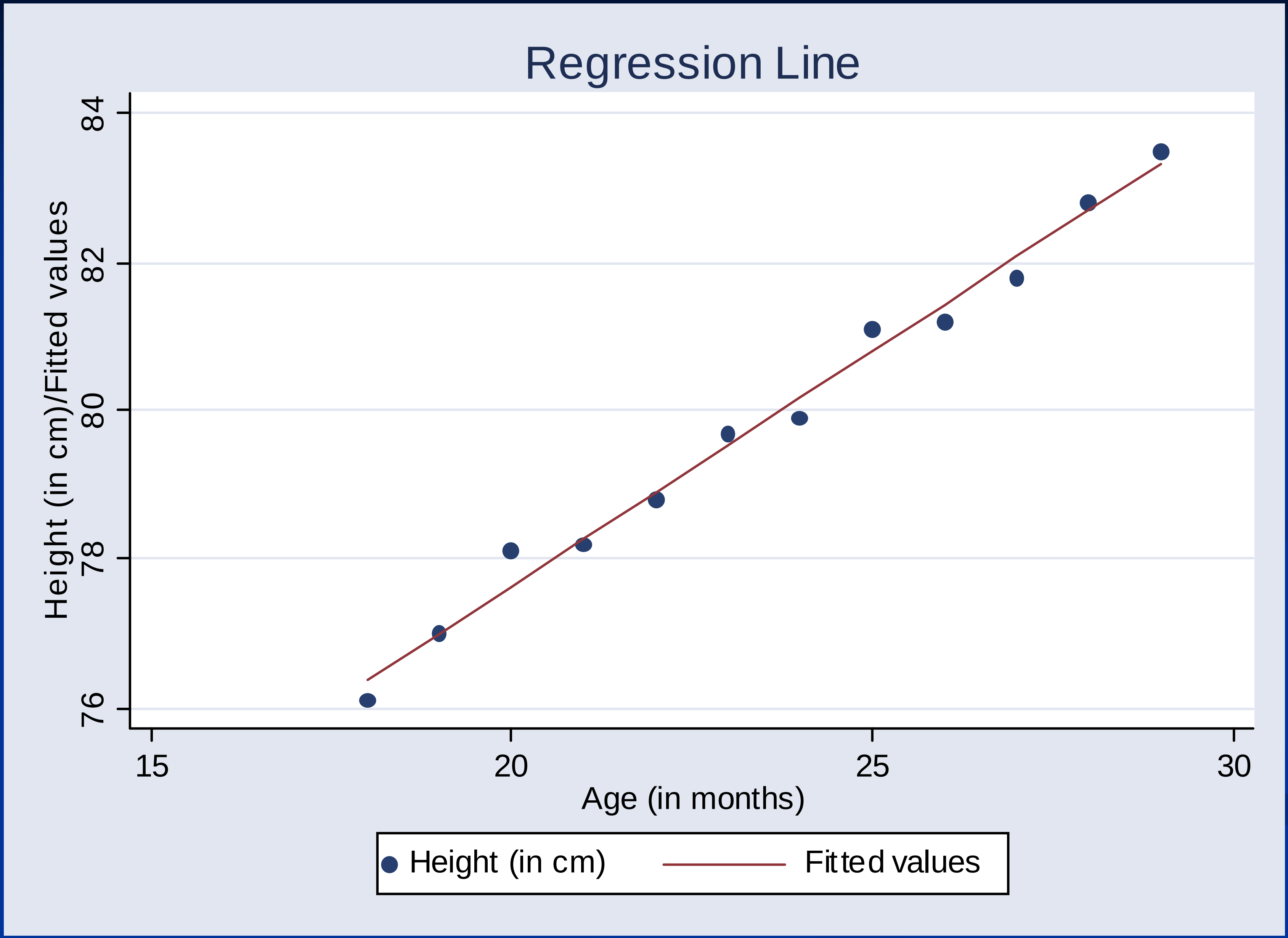
Age x in months	Height y in centimeters
18	76.1
19	77.0
20	78.1
21	78.2
22	78.8
23	79.7
24	79.9
25	81.1
26	81.2
27	81.8
28	82.8
29	83.5

Example: Height and Age

```
. regress height age

      Source |       SS       df       MS              Number of obs =      12
-------------+------------------------------           F(  1,    10) =  880.00
       Model |  57.6548678     1  57.6548678           Prob > F      =  0.0000
    Residual |  .655171562    10  .065517156           R-squared     =  0.9888
-------------+------------------------------           Adj R-squared =  0.9876
       Total |  58.3100394    11  5.30091267           Root MSE      =  .25596

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .6349653   .0214047    29.66   0.000     .5872726     .682658
       _cons |   64.92832    .508409   127.71   0.000     63.79551    66.06112
------------------------------------------------------------------------------
```


Regression Line
84
82
80
78
76
Height (in cm)/Fitted values
15
20
25
30
Age (in months)
Height (in cm)
Fitted values