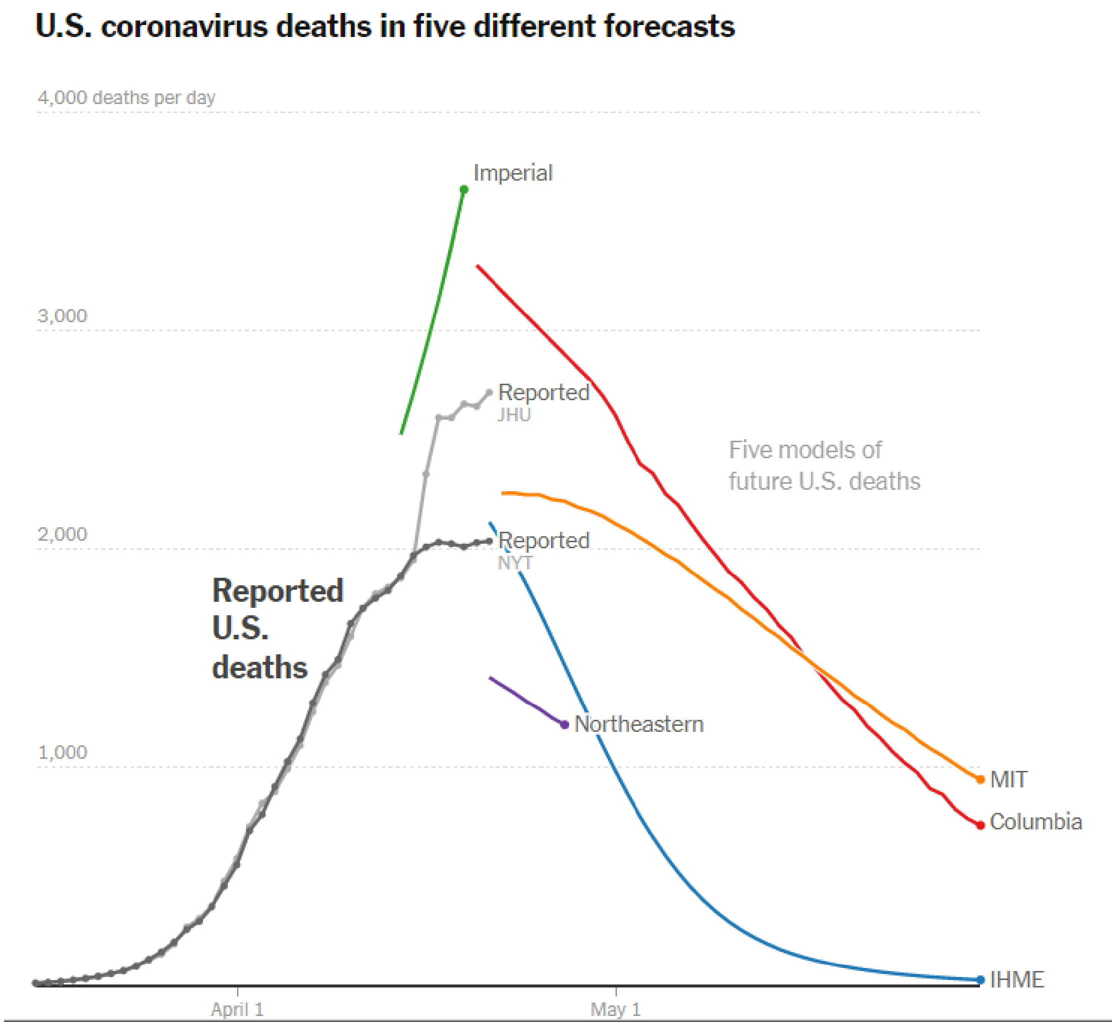
Probability and Mathematical Statistics in Data Science

Lecture 31: Section 11.3: Least Squares Regression

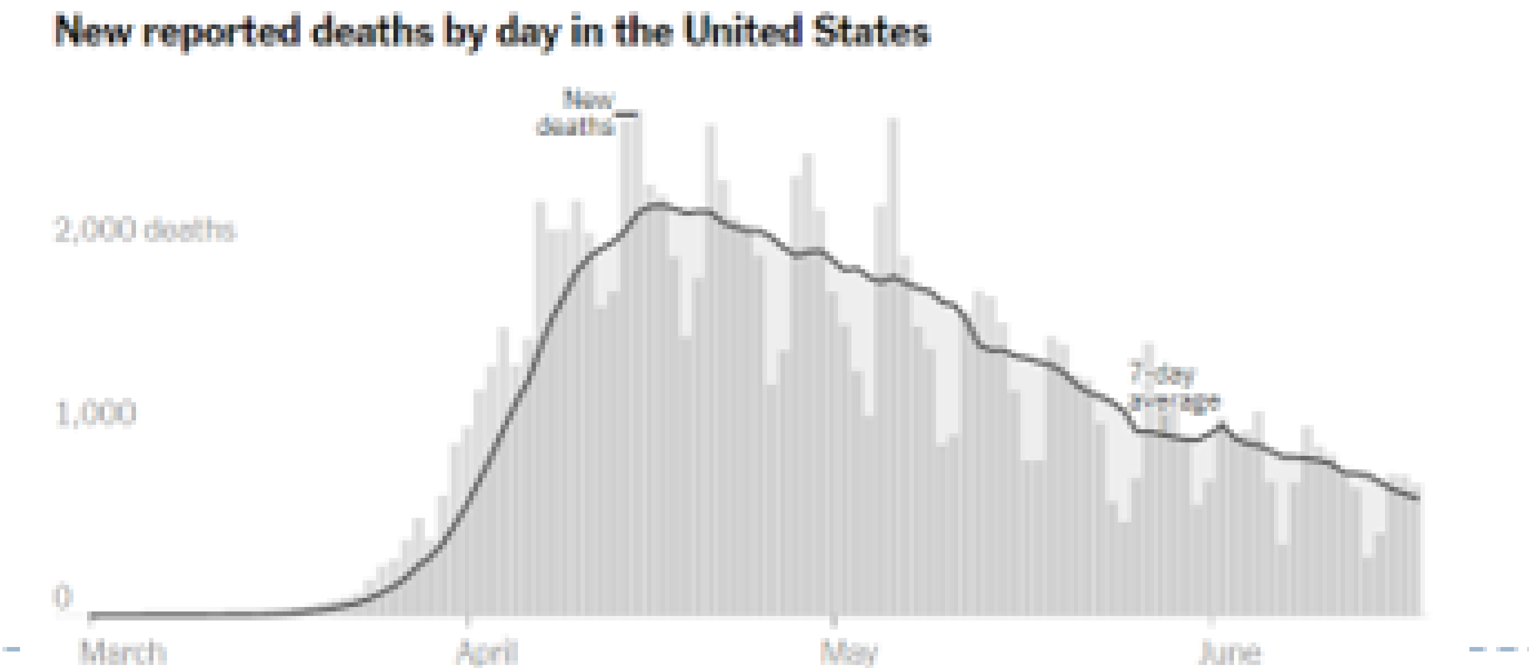
All models are wrong but some are useful

—George E. Box

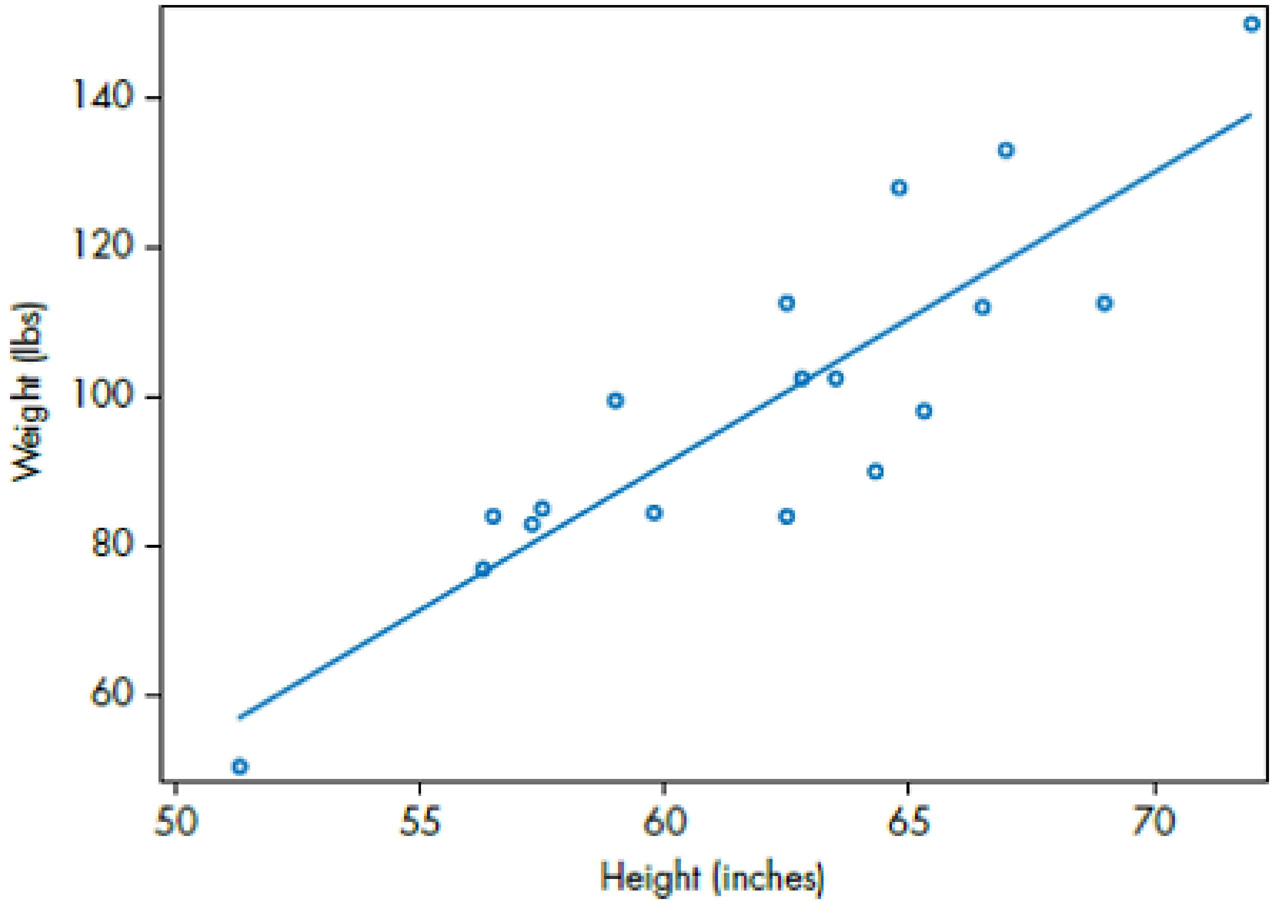
What 5 Coronavirus Models Say the Next Month Will Look Like – NY Times – April 22nd, 2020



Coronavirus in the U.S.: Latest Map and Case Count – NY Times – June 20th



Specifying Linear Relationships with Linear Regression



Modeling Relationships: Linear Regression

- We can summarize the linear relationship between two quantitative variables by fitting a line to the scatterplot of data points

- In this context, the x-axis variable is known as the **explanatory variable.** The y-axis variable is known as the **response variable**

- In our example of 19 children, height is our explanatory variable and weight is our response variable

- We are using the variable height to try and explain (at least some) of the variability in weight measurements

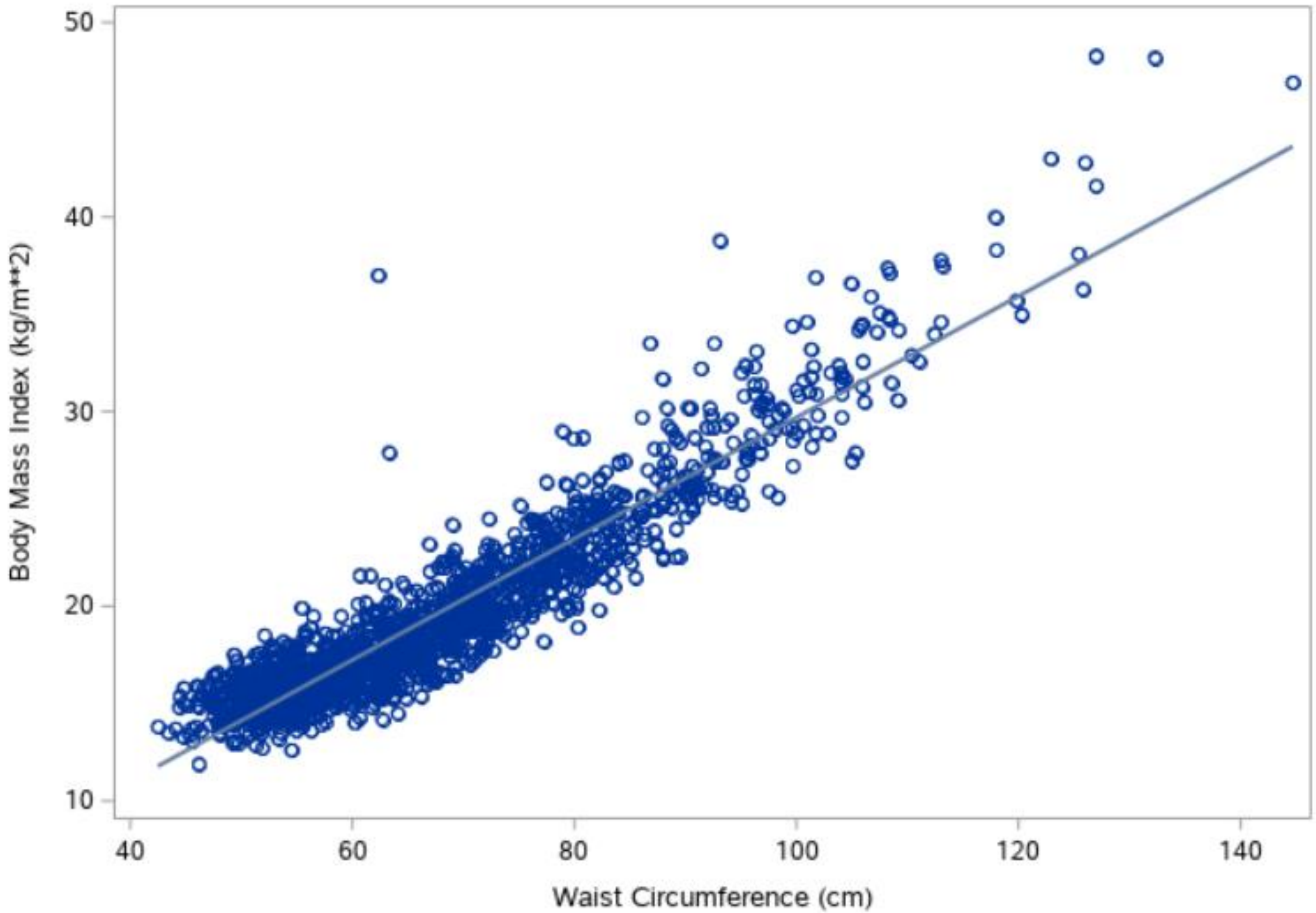
Modeling Relationships: Linear Regression

- **Regression analysis** is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

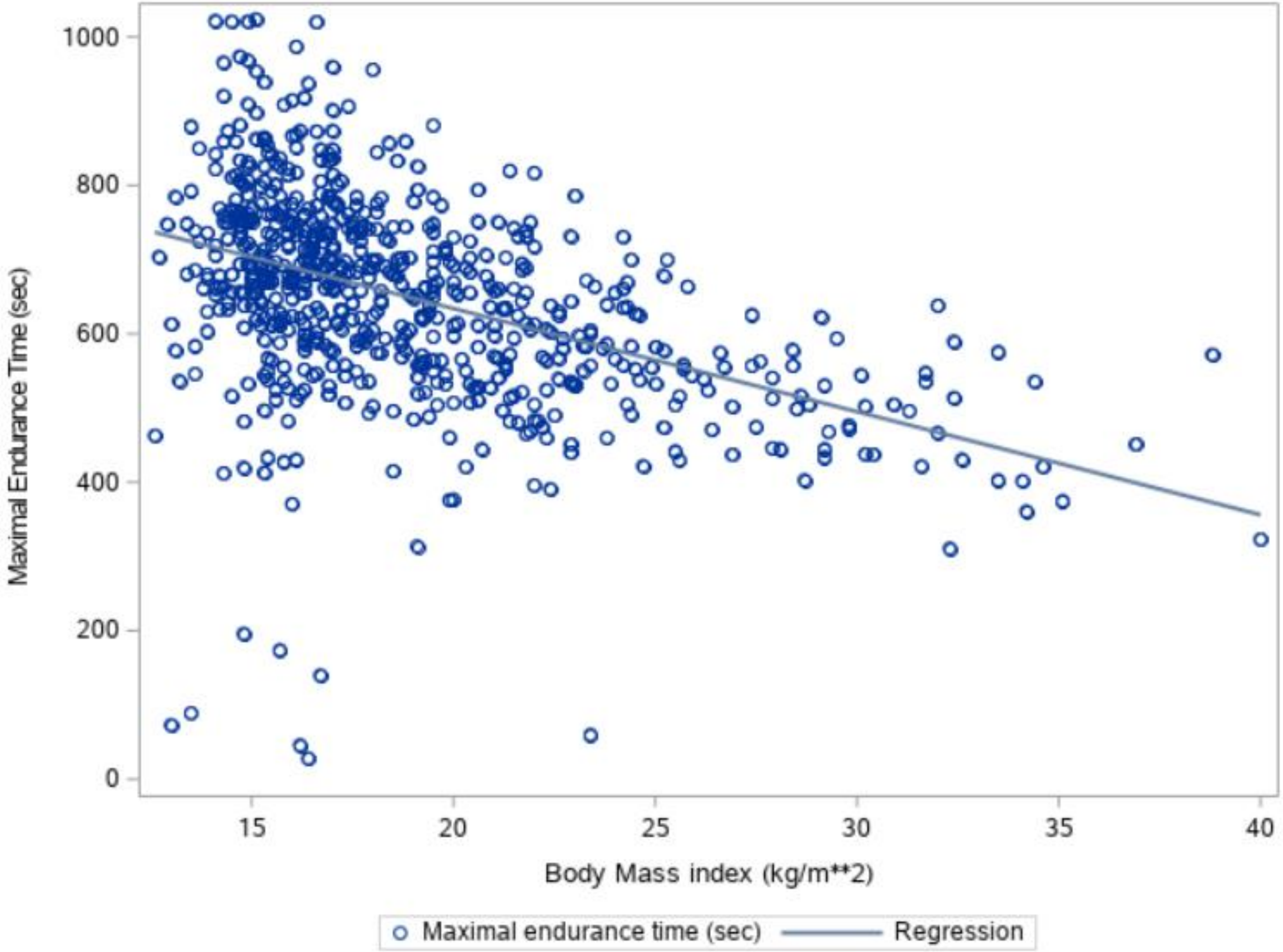
Dependent variable: the variable we wish to predict or explain

Independent variable: the variable used to predict or explain the dependent variable

Specifying Linear Relationships with Linear Regression



Specifying Linear Relationships with Linear Regression



Least Squares Regression

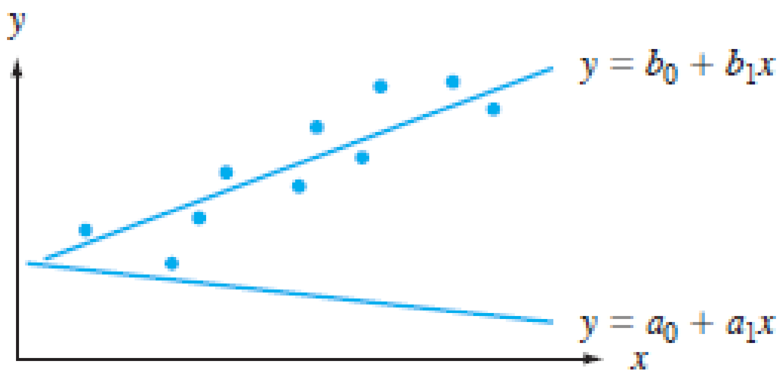
- Given a random variable pair X, Y , we want a model that describes the relationship between the predictor (X) and response (Y) variables. That is, can we express the relationship mathematically?

- Perhaps as $Y = f(X)$ or $Y = f(X) + random\ error$

- Want to use a linear function of X to estimate Y, say $aX + b$

- What is the "Best" line these these data.

Least Squares Regression



- What is the best line to use?

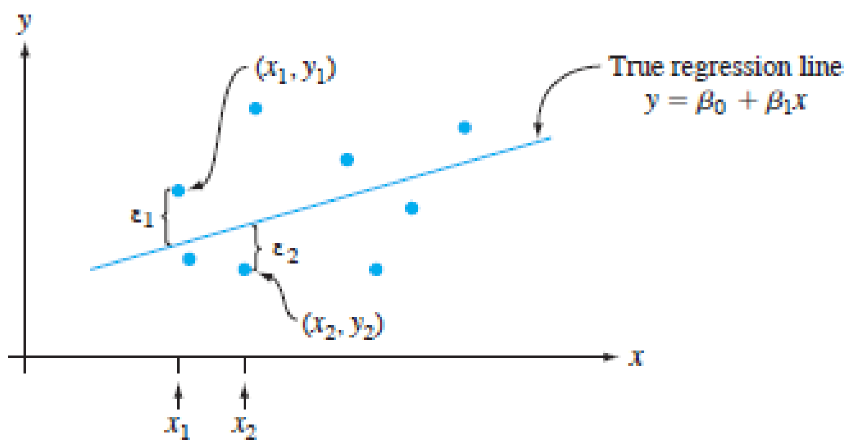
The Simple Linear Regression Model

The Simple Linear Regression Model

There are parameters β_0, β_1, and σ^2, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the **model equation**

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{12.1}$$

The quantity ϵ in the model equation is a random variable, assumed to be normally distributed with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.



Modeling Relationships: Linear Regression

Population Y intercept

Population Slope Coefficient

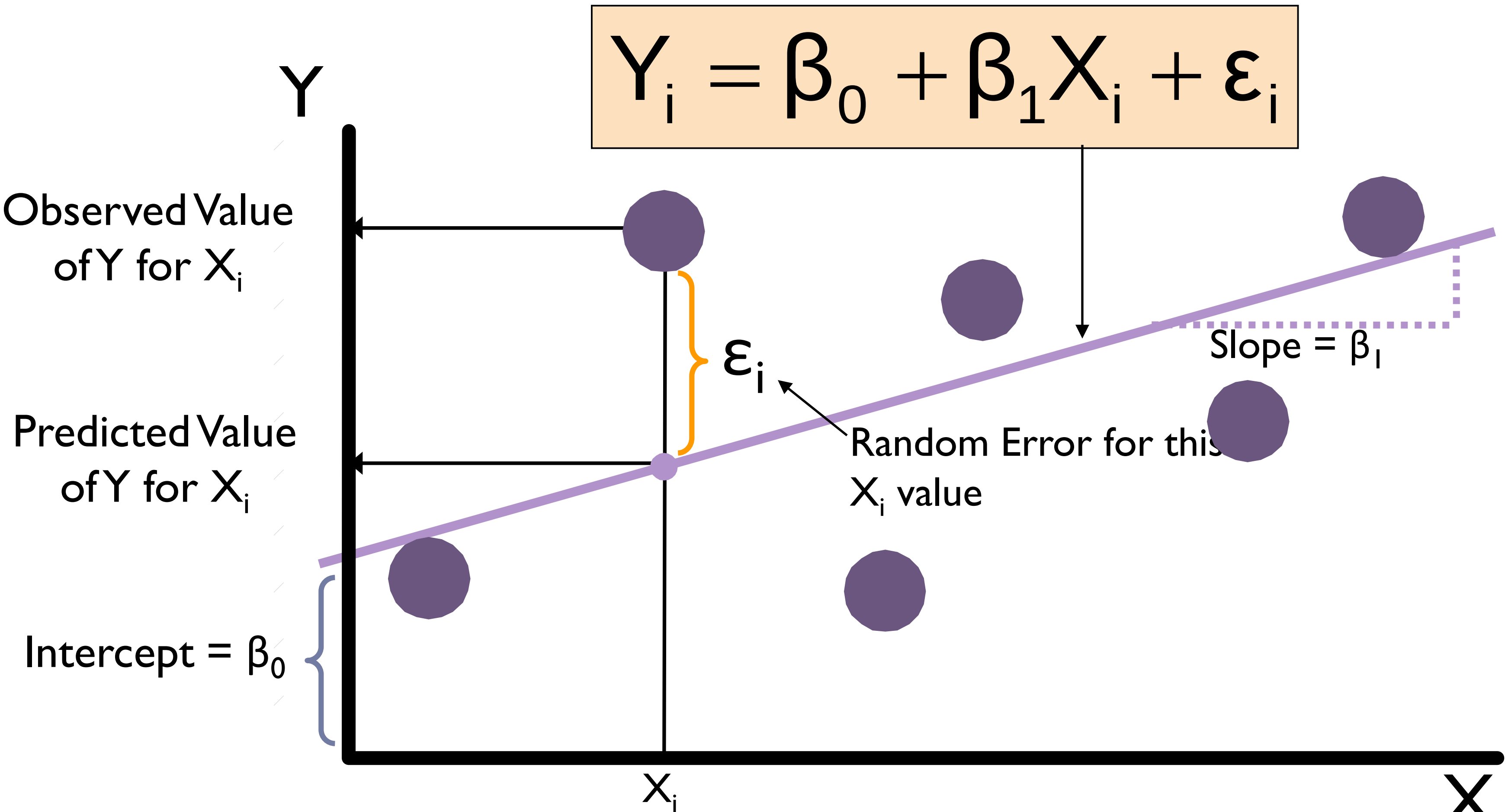
Independent Variable

Random Error term

Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

Modeling Relationships: Linear Regression



Least Squares Regression

- The regression method is used to draw the regression line which can be used for prediction.

- It is also called the **least squares line** because it minimizes mean squared error. By *error* we mean the vertical difference between the y-value for some x, and the height of the regression line at that x.

- $e_i = y_i - (b_0 + b_1 x)\ ,\ i = 1, 2, \ldots, n$

Principle of Least Squares

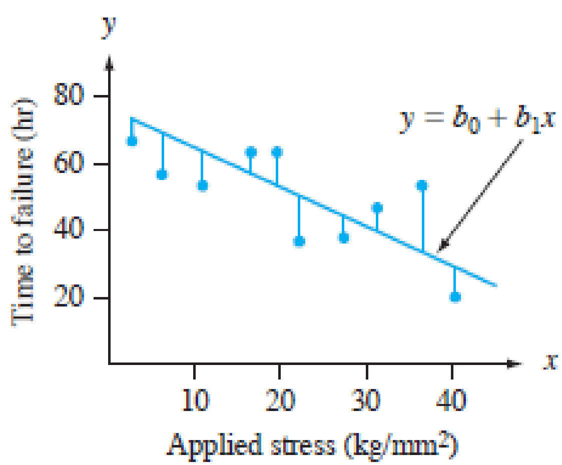
The vertical deviation of the point (x_i, y_i) from the line $y = b_0 + b_1x$ is

$$\text{height of point} - \text{height of line} = y_i - (b_0 + b_1x_i)$$

The sum of squared vertical deviations from the points $(x_1, y_1), \ldots, (x_n, y_n)$ to the line is then

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1x_i)]^2$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the **least squares estimates**, are those values that minimize $f(b_0, b_1)$. That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are such that $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ for any b_0 and b_1. The **estimated regression line** or **least squares line** is then the line whose equation is $y = \hat{\beta}_0 + \hat{\beta}_1x$.



Taking the Derivatives

The minimizing values of b_0 and b_1 are found by taking partial derivatives of $f(b_0, b_1)$ with respect to both b_0 and b_1, equating them both to zero [analogously to $f'(b) = 0$ in univariate calculus], and solving the equations

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

The Least Squares Intercept and Slope

The least squares estimate of the slope coefficient β_1 of the true regression line is

$$b_1 = \hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad (12.2)$$

Computing formulas for the numerator and denominator of $\hat{\beta}_1$ are

$$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n \qquad S_{xx} = \sum x_i^2 - (\sum x_i)^2/n$$

The least squares estimate of the intercept β_0 of the true regression line is

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (12.3)$$

Modeling Relationships: Linear Regression

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

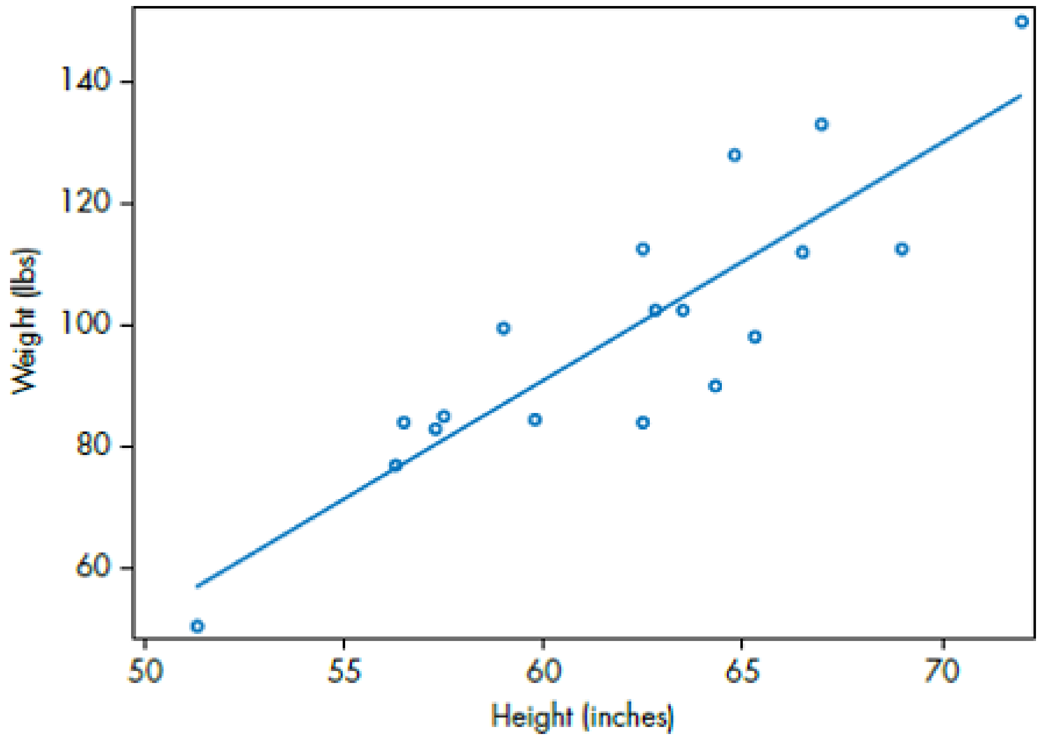
Estimate of the regression intercept

Estimate of the regression slope

Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

19 Children Height-Weight Example



- Our aim is to fit a line to the data that gets as close to the data points as possible.
- For this reason, the line is often called the **line of best fit**.

19 Children Height-Weight Example

There are two children in our sample, Janet and Jeffrey, with a height of 62.5 inches. The individual observed weights for Janet and Jeffrey are 112.5 lbs. and 84 lbs., respectively.

Predicted Weight = –143 + 3.9 x (62.5)
= –143 + 243.75
= 100.75 lbs.

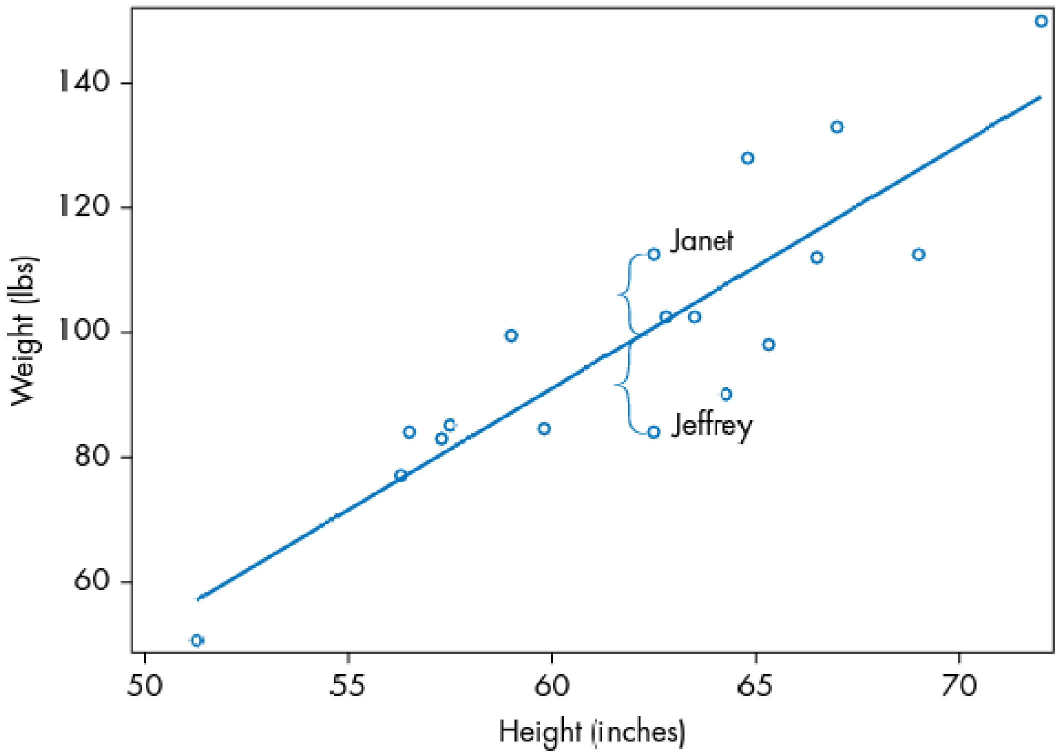
Therefore, the individual residual deviations for Janet and Jeffrey are as follows:

$$\boxed{e_i = y_i - \hat{y}_i}$$

residual deviation = observed weight – predicted weight

Janet: 112.5 lbs.: residual deviation = 112.5 – 100.75 = 11.75 lbs.
Jeffrey: 84 lbs.: residual deviation = 84 – 100.75 = –16.75 lbs.

19 Children Height-Weight Example



$$\boxed{\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

- Line of Best Fit -> Minimize the Sum of the Squared Residuals

The Least Squares Method

b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between Y and $\hat{Y}$:

$$\boxed{\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2}$$

Example: From STATS: Data and Models

57. **Body fat.** It is difficult to determine a person's body fat percentage accurately without immersing him or her in water. Researchers hoping to find ways to make a good estimate immersed 20 male subjects, then measured their waists and recorded their weights.

Waist (in.)	Weight (lb)	Body Fat (%)	Waist (in.)	Weight (lb)	Body Fat (%)
32	175	6	33	188	10
36	181	21	40	240	20
38	200	15	36	175	22
33	159	6	32	168	9
39	196	22	44	246	38
40	192	31	33	160	10
41	205	32	41	215	27
35	173	21	34	159	12
38	187	25	34	146	10
38	188	30	44	219	28

a) Create a model to predict *%Body Fat* from *Weight*.
b) Do you think a linear model is appropriate? Explain.
c) Interpret the slope of your model.
d) Is your model likely to make reliable estimates? Explain.
e) What is the residual for a person who weighs 190 pounds and has 21% body fat?

Textbook Body Fat Example: Excel Output

The regression equation is:

$$\widehat{\text{Body Fat}(\%)} = -27.376 + 0.2499\,(\text{weight})$$

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.69663276
R Square	0.485297203
Adjusted R Square	0.456702603
Standard Error	7.049132279
Observations	20

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	843.325214	843.3252	16.97164	0.000643448
Residual	18	894.424786	49.69027		
Total	19	1737.75			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-27.37626233	11.54742832	-2.37077	0.029119	-51.63650899	-3.116015659
Weight	0.249874137	0.060653997	4.119665	0.000643	0.122444818	0.377303457

Textbook Body Fat Example: Interpretation of b_o

$$\widehat{\text{Body Fat}}(\%) = -27.376 + 0.2499\,(\text{weight})$$

- b_0 (-27.376) is the estimated average value of body fat(%) when the value of weight(lb) is zero (if weight = 0 is in the range of observed X values)

- Because we can't have a weight of 0, b_0 has no practical application

Textbook Body Fat Example: Interpreting b_1

$$\widehat{\text{Body Fat}}(\%) = -27.376 + 0.2499\,(\text{weight})$$

- b_1 (0.2499) estimates the change in the average value of body fat(%) as a result of a one-unit increase in weight(lb)

Here, $b_1 = 0.2499$ tells us that the mean value of body fat(%) increases by 0.2499, on average, for each additional one pound increase in weight

Textbook Body Fat Example: Making Predictions

Predict the body fat(%) for a person whose weight is190 lbs:

$$\widehat{\text{Body Fat}(\%)} = -27.376 + 0.2499\,(\text{weight})$$

$$\widehat{\text{Body Fat}(\%)} = -27.376 + 0.2499\,(190)$$

$$= 20.1$$

What is the residual for someone who weighs 190 lbs and has a body fat content of 21%?