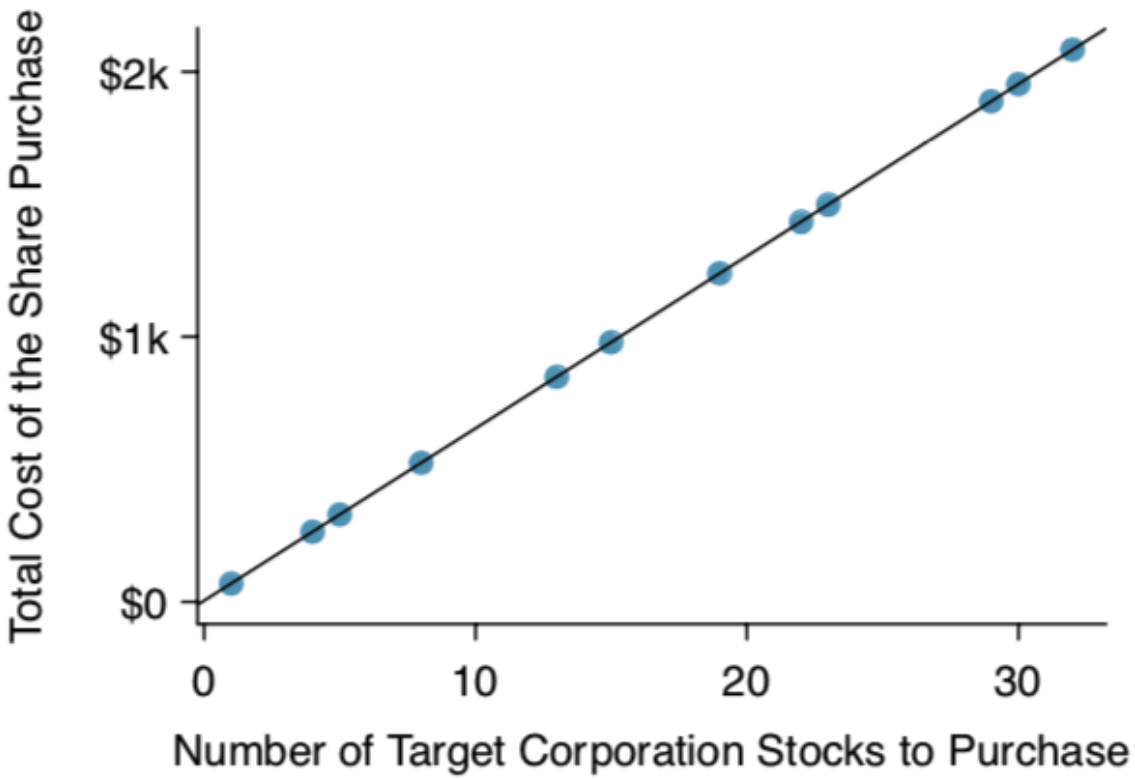
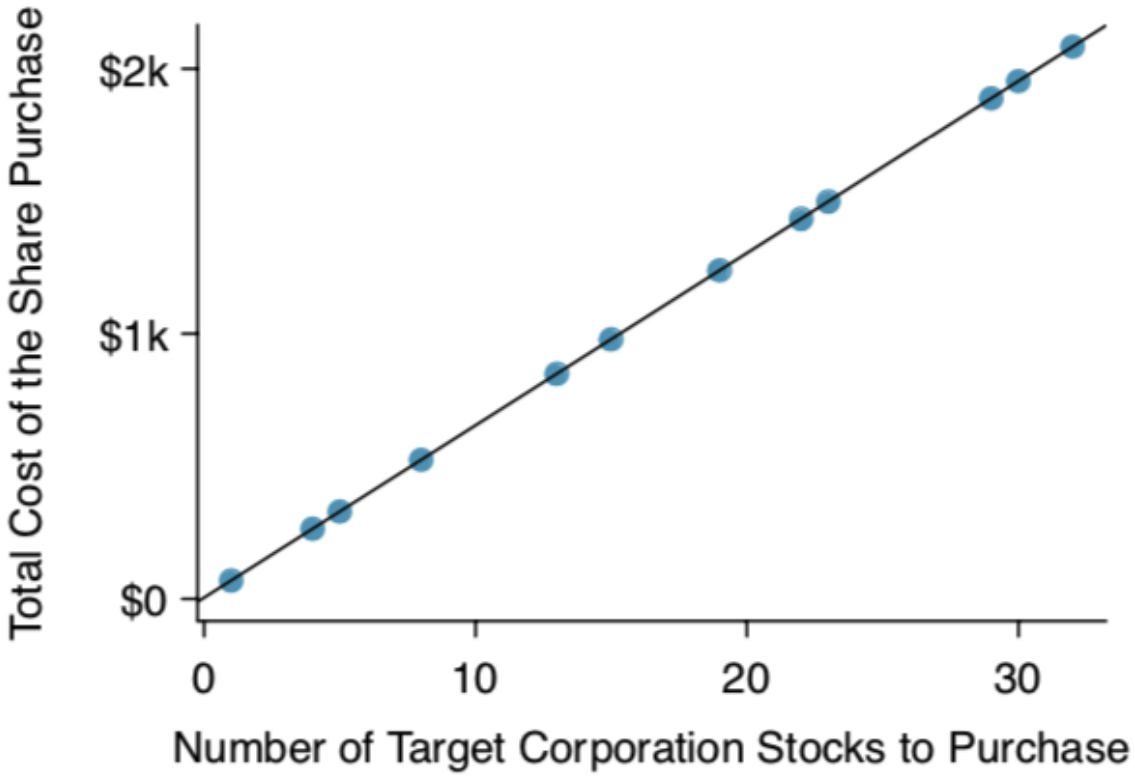
Scatterplots

- **Scatterplots** allow us to plot pairs of points (x, y) on a graph.
- If x and y are variables, we can use this to visualize their relationship.

Scatterplots



Fitting a Line to Data



This relationship can be modeled perfectly with a straight line:

$$y = 5 + 64.96x$$

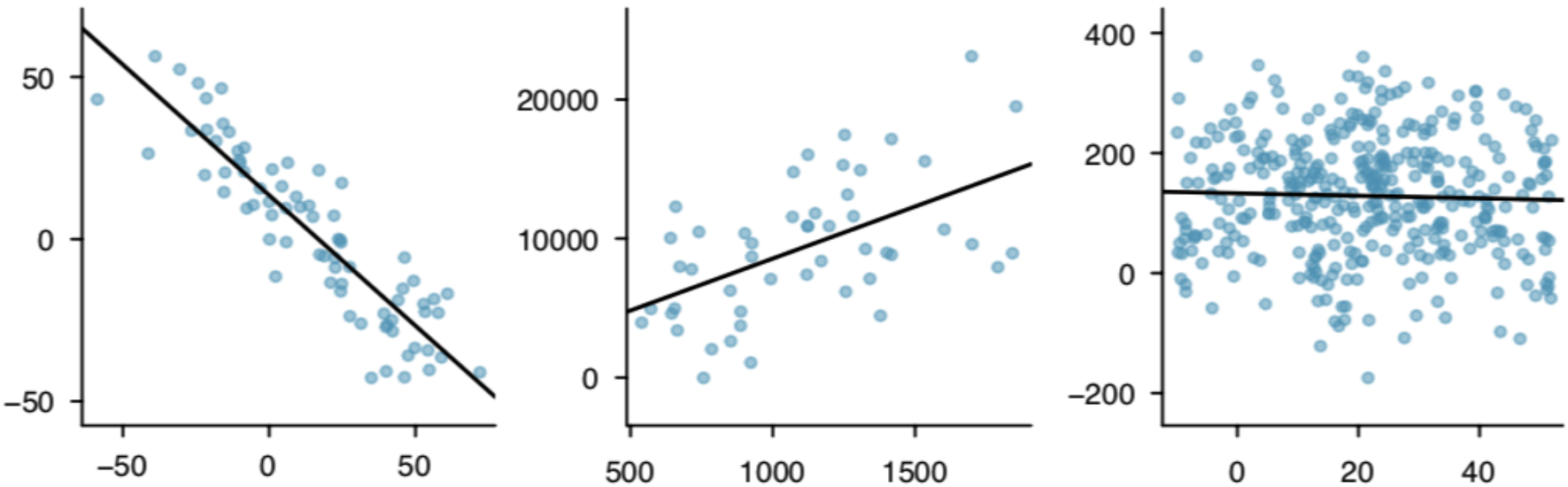
Fitting a Line to Data

When we can model a relationship *perfectly*,

$$y = 5 + 64.96x,$$

we know the exact value of y just by knowing the value of x.

Linear Regression



Linear regression takes this idea of modeling data with a line and allows the relationship to be imperfect.

Clouds of Points

Think of this like the 2-dimensional version of using $\bar{x}$ to estimate μ.

- We use sample data to estimate relationship between two variables.
- The true (population) relationship is unknown.

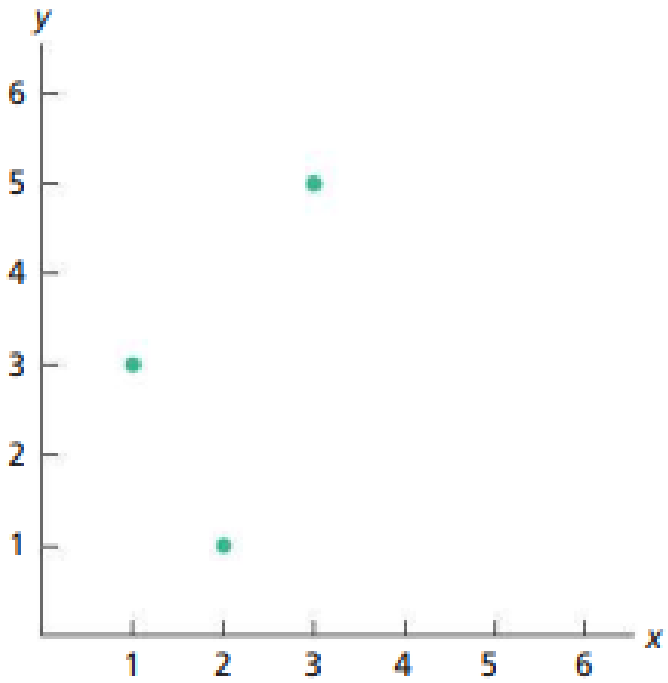
Linear Regression

For a regression line

$$y = b_0 + b_1 x$$

we make predictions about y using values of x.

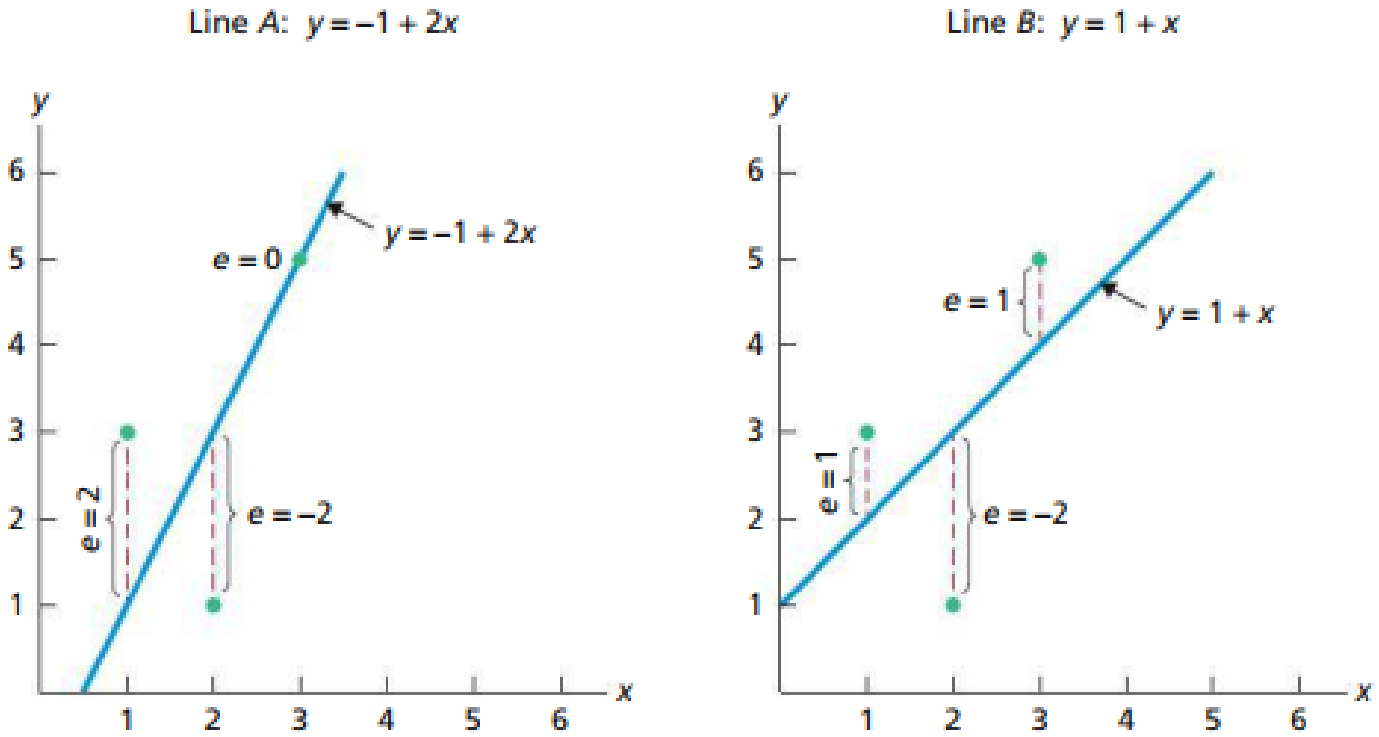
- y is called the **response** or **dependent variable**.
- x is called the **independent** or **predictor variable**.



x	y
1	3
2	1
3	5

There are lots of potential ways to include a line!

Here are two possible lines:



We let e represent "error". This is the difference between the *predicted* y-value and the actual y-value from the data.

Error

We use $\hat{y}$ to denote a predicted y-value.

$$e = y - \hat{y}$$

Example: Find the error for Lines A and B at the data point where $x = 1$, $y = 3$.

Example Solution

At the data point where $x = 1$, $y = 3$:

- Line A ($y = -1 + 2x$) has error

$$3 - (-1 + 2 \times 1) = 2$$

- and Line B ($y = 1 + x$) has error

$$3 - (1 + 1) = 1$$

Which is better?

For Line A:

x	y	$\hat{y}$	e
1	3	1	2
2	1	3	-2
3	5	5	0

For Line B:

x	y	$\hat{y}$	e
1	3	2	1
2	1	3	-2
3	5	4	1

Hint: think back to standard deviation!

Least Squares Criterion

For Line A:

x	y	$\hat{y}$	e
1	3	1	2
2	1	3	-2
3	5	5	0

For Line B:

x	y	$\hat{y}$	e
1	3	2	1
2	1	3	-2
3	5	4	1

What now? Think back to standard deviation!

Least-Squares Criterion

Def: The **least-squares criterion** is that the line that best fits a set of data points is the one having the smallest possible sum of squared errors.

Def: A **regression line** is the line that best fits a set of data points according to this criterion.

Def: A **regression equation** is the equation of the regression line.

Some Useful Formulas

Quantity	Formula	Computing Formula
S_{xx}	$\sum(x_i - \bar{x})^2$	$\sum x_i^2 - (\sum x_i)^2/n$
S_{xy}	$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum x_i y_i - (\sum x_i)(\sum y_i)/n$
S_{yy}	$\sum(y_i - \bar{y})^2$	$\sum y_2^2 - (\sum y_i)^2/n$

The Regression Equation

The regression equation for a set of n data points is $\hat{y} = b_0 + b_1 x$ where

$$b_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad b_0 = \bar{y} - b_1 \bar{x}$$

Ex:

Calculate the regression equation for the data:

x	y
1	3
2	1
3	5

Solution

First, $\bar{x} = 2$ and $\bar{y} = 3$.

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	3	-1	0	1	0
2	1	0	-2	0	0
3	5	1	2	1	2
				S_{xx}: 2	S_{xy}: 2

So

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{2}{2} = 1 \quad \text{and} \quad b_0 = \bar{y} - b_1\bar{x} = 3 - 1 \times 2 = 1$$

Then the regression equation is

$$\hat{y} = b_0 + b_1 x = 1 + x.$$

(Note: we almost never do this by hand!)

Prediction

Often, when we build a regression model our goal is prediction.

- We want to use information about the predictor variable to make predictions about the response variable.

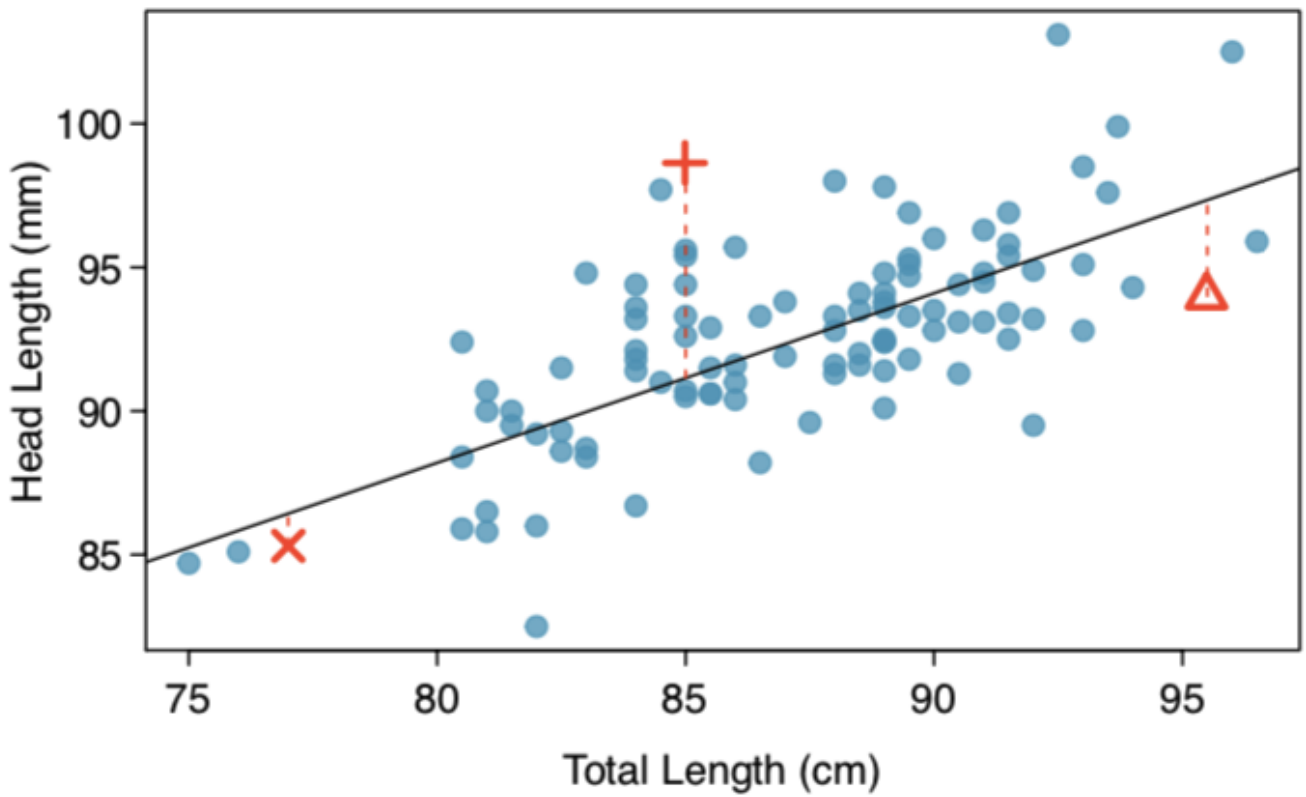
Example: Possum Head Lengths

Researchers captured 104 brushtail possums and took a variety of body measurements on each before releasing them back into the wild.

We consider two measurements for each possum:

- total body length (cm).
- head length (mm).

Example: Possum Head Lengths



$$\hat{y} = 41 + 0.59x$$

Predict the head length for a possum with a body length of 80 cm.

Example Solution

Plugging $x = 80$ into the regression equation,

$$\hat{y} = 41 + 0.59 \times 80 = 88.2.$$

The predicted head length is 88.2 mm.

Extrapolation

- When we make predictions, we plug in values of x to estimate values of y.
- However, this has limitations!
- We don't know how the data outside of our limited window will behave.

Extrapolation

Applying a model estimate for values outside of the data's range for x is called **extrapolation**.

- The linear model is only an approximation.
- We don't know anything about the relationship outside of the scope of our data.
- Extrapolation assumes that the linear relationship holds in places where it has not been analyzed.

Example

For the possum data, body lengths range from 75 to 97 cm.

- We predicted a head length of 88.2 mm for a body length of 80 cm.
- Now let's predict head length for a body length of 150 cm.
  How confident are we in this prediction?

Example Solution

For a body length of 150 cm,

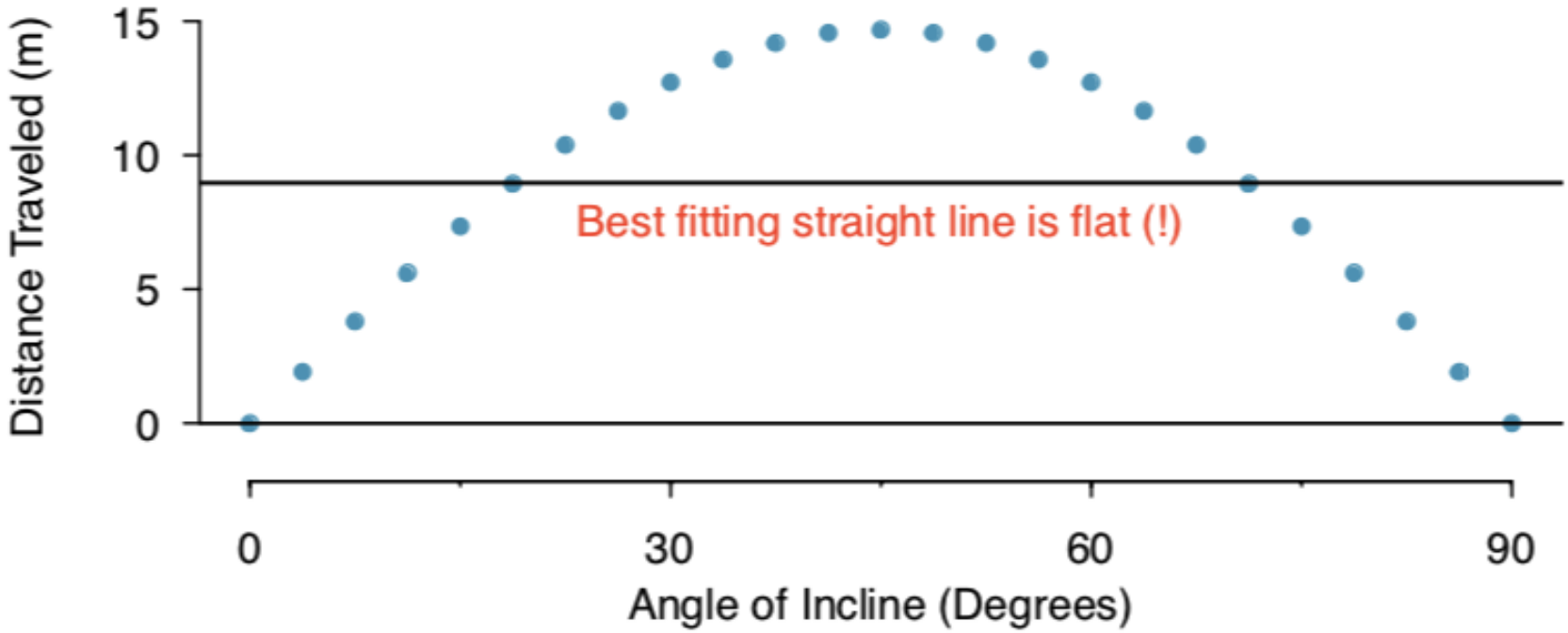
$$\hat{y} = 41 + 0.59 \times 150 = 129.5.$$

The predicted head length is 129.5 mm... but since we are *extrapolating*, we shouldn't be very comfortable making this prediction.

Outliers and Influential Observations

- An **outlier** is a data point that falls from from the regression line, relative to the other data points.
- An **influential observation** is a data point whose removal causes the regression equation to change considerably.

A Final Word of Caution



Sometimes, there is a clear relationship but simple linear regression won't work! Always plot your data before doing a linear regression.