Week 3: Basic regression
3. Linear model interpretation

Stat 140 - 04
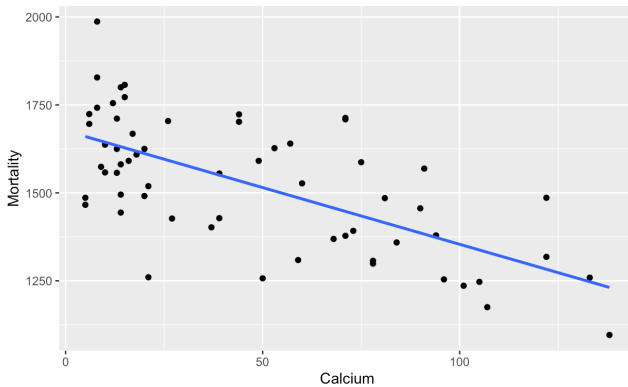
Mount Holyoke College

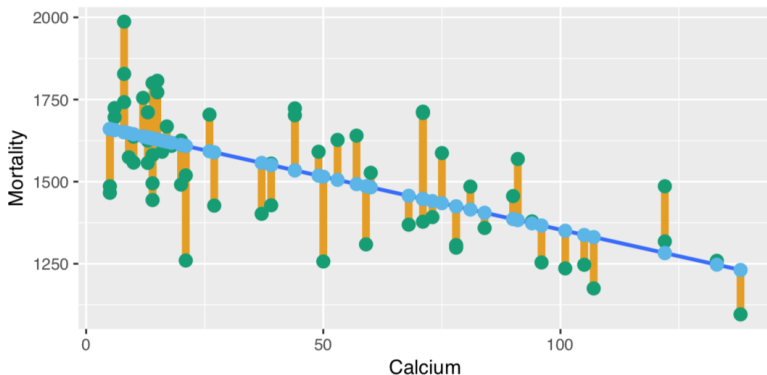The line of best fit is the line for which the sum of the squared residuals is smallest, the **least squares line**.

The algebraic equation for a line is

$$Y = b_0 + b_1 X$$

The use of coordinate axes to show functional relationships was invented by Rene Descartes (1596-1650). He was an artillery officer, and probably got the idea from pictures that showed the trajectories of cannonballs.
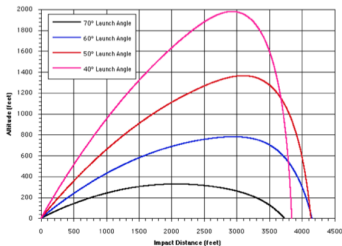
- x: the explanatory variable (calcium concentration)
- y: the response variable (mortality rate)
- $\bar{x}, \bar{y}$: sample mean of x and y
- s_x, s_y: sample standard deviation of x and y
- R: correlation between x and y

The least square line has

▶ slope:

$$b_1 = \frac{s_y}{s_x} R$$

▶ intercept (the value at $x = 0$):

$$b_0 = \bar{y} - b_1 \bar{x}$$

Suppose we have the following information.

	mortality rate (y)	calcium concentration (x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

Suppose we have the following information.

	mortality rate (y)	calcium concentration (x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

1. Calculate the slope.
$$b_1 = R \times \frac{s_y}{s_x} = -0.65 \times \frac{188}{38} = -3.21$$

Suppose we have the following information.

	mortality rate (y)	calcium concentration (x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

1. Calculate the slope.
$$b_1 = R \times \frac{s_y}{s_x} = -0.65 \times \frac{188}{38} = -3.21$$

2. Calculate the intercept.
$$b_0 = \bar{y} - b_1 \times \bar{x} = 1524 + 3.21 \times 47 = 1674.87$$

Suppose we have the following information.

	mortality rate (y)	calcium concentration (x)
mean	$\bar{y} = 1524$	$\bar{x} = 47$
sd	$s_y = 188$	$s_x = 38$
correlation		$R = -0.65$

1. Calculate the slope.
$$b_1 = R \times \frac{s_y}{s_x} = -0.65 \times \frac{188}{38} = -3.21$$

2. Calculate the intercept.
$$b_0 = \bar{y} - b_1 \times \bar{x} = 1524 + 3.21 \times 47 = 1674.87$$

3. Write out the linear model.
$$\widehat{\text{Mortality}} = 1675 - 3 \text{ Calcium}$$

In general, the regression line is

$$\hat{y} = b_0 - b_1 x$$

1. **Slope** b_1: Slopes are always expressed in y-units per x-unit. They tell how the y-variable changes (in its units) for a one-unit change in the x-variable.
2. **Intercept** b_0: the value the line takes when x is zero

How to interpret intercept and slope in the context of data?

Units:

- x-variable: calcium concentration (parts per million)
- y-variable: mortality rate (deaths per 100,000 population)

The slope, -**3**, says that for 1 unit increase in x-variable, we can expect, on average, to have 3 units less in y-variable.

Units:

- x-variable: calcium concentration (parts per million)
- y-variable: mortality rate (deaths per 100,000 population)

The slope, **-3**, says that for 1 unit increase in x-variable, we can expect, on average, to have 3 units less in y-variable.

This means, for 1 part per million increase in calcium concentration, we can expect, on average, to have 3 deaths per 100,000 population less in mortality rate.

Units:

- x-variable: calcium concentration (parts per million)
- y-variable: mortality rate (deaths per 100,000 population)

The slope, **-3**, says that for 1 unit increase in x-variable, we can expect, on average, to have 3 units less in y-variable.

This means, for 1 part per million increase in calcium concentration, we can expect, on average, to have 3 deaths per 100,000 population less in mortality rate.

Less formally, for each additional parts per million increase in calcium concentration, the predicted number of mortality rate decreases by 3 deaths per 100,000 population.

Algebraically, that's the value the line takes when x is zero.

Here, our model predicts that when the water does not have any calcium, on average, the mortality rate is 1676 deaths per 100,000 population.

Note that the intercept serves only as a starting value for our predictions, and we don't interpret it as a meaningful predicted value.

One of the towns in our sample had a measured Calcium concentration of 71. What is the predicted value for the mortality rate in that town?

$$\widehat{\text{Mortality}} = 1675 - 3 \text{ Calcium}$$

One of the towns in our sample had a measured Calcium concentration of 71. What is the predicted value for the mortality rate in that town?

$$\widehat{\text{Mortality}} = 1675 - 3 \text{ Calcium}$$

By hand:

$$\widehat{\text{Mortality}} = 1675 - 3 \times 71 = 1462$$

The predicted value for the mortality rate in that town is 1462 deaths per 100,000 population.

The calculations in this unit are particularly sensitive to the rounding. If your answers differ from what I provided here, they may still be correct in the sense that you calculated them by a correct method, buy you may have used values that were rounded differently than those used to find these answers.

In general, we use the original data and do no intermediate rounding.

Don't be concerned about minor differences — what is important is your interpretations of the results.

1. Find the least square line by hand
2. Interpret intercept and slope