

Linear Regression

Modeling two numeric variables

Download the section 15 .Rmd handout to
STAT240/lecture/sect15-regression-intro.

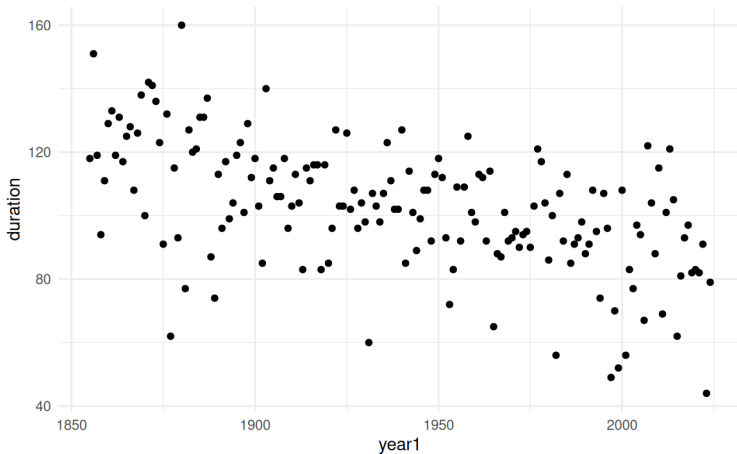
Material in this section is covered by Chapter 14 on
the notes website.

Download lake-monona-winters-2025.csv and
riley.txt to STAT240/data.

Linear regression models two continuous variables.

- Son's height vs father's height
- Sales vs advertising spending
- Anything with (x_i, y_i) pairs

We'll work with Lake Monona year (x) and freeze duration (y).



A scatterplot shows a downward trend.

Correlation quantifies the *linear* relationship.

Let $(x_1, y_1), \dots, (x_n, y_n)$ be a sample of n points and let $\bar{x}, s_x, \bar{y}, s_y$ be the sample means and SDs.

Correlation r is calculated by adding the product of the deviations, and dividing by $(n - 1)s_x s_y$.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n - 1)s_x s_y}$$

For the Monona freeze data, we see a correlation of -0.543, a negative correlation.

How do we interpret this?

Correlation is a unitless measure from -1 to 1.

It is -1 or 1 when the points are in a perfect line.

- $|r| > 0.8$ is considered strong correlation
- $0.5 < |r| \leq 0.8$ is considered moderate
- $|r| \leq 0.5$ is considered weak

Correlation measures linear relationship.

- Basis of our linear model
- Does not pick up on other relationships
- Graph your data!

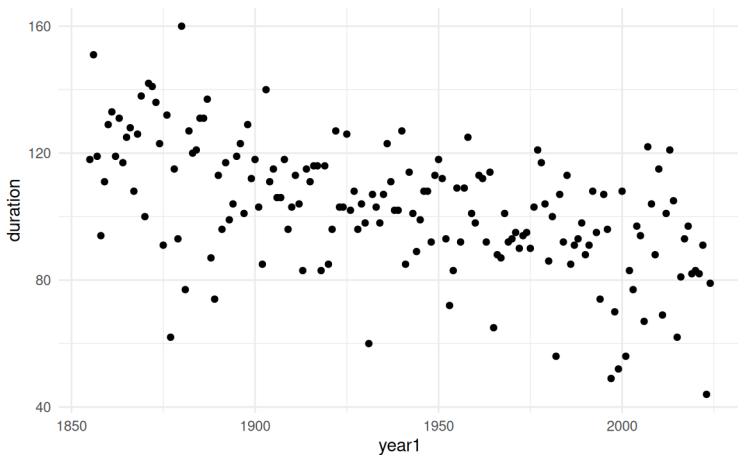
Remember correlation does not equal causality!

We use correlation to build a **linear model**, which has a slope and a y-intercept:

$$y = mx + b$$

In statistics, we use the notation

$$y = \beta_0 + \beta_1 x$$



What is the true relationship between year/duration?

A linear model looks like

$$\text{Duration} = \beta_0 + \beta_1 (\text{Year}) + \text{Random error}$$

- β_0 is the duration when the year is 0 (?)
- β_1 is the change in duration one year later.

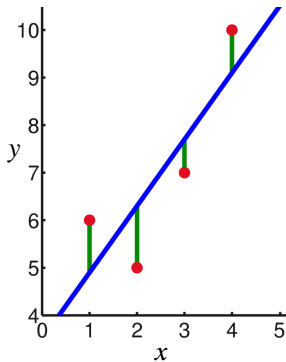
$$\text{Generally: } y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- β_0 and β_1 are parameters of interest
- ϵ_i are unknown errors

How to estimate β_0 and β_1 ?

Minimize the *vertical* distance from the observed data to the estimated line.



(Wikipedia)

The difference between the observed and estimated y 's: $(y_i - \hat{y}_i)$ is called the **residual**.

The estimated y for a given x_i is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Pick $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the
sum of squared residuals.

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

$$\hat{\beta}_1 = r \left(\frac{s_y}{s_x} \right)$$

- Related to r
- $\frac{s_y}{s_x}$ tells us whether the data is tall or wide

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- Line goes through (\bar{x}, \bar{y})

riley.txt records a boy's height and age.

- Filter from age 24 months to 96 months

Find the slope and intercept for the least-squares regression line.

$$\hat{\beta}_1 = r\left(\frac{s_y}{s_x}\right), \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

We predict Riley's height at age x months to be

$$\hat{y} = 30.25 + 0.25(x)$$

We can also use R's `lm()` function.

We've seen how to estimate a linear model on (x_i, y_i) data pairs.

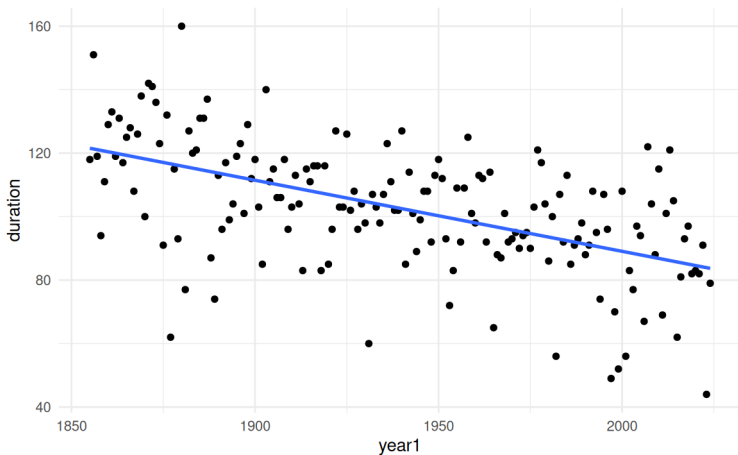
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

We can do this on any set of data. Is this model actually valid?

A model relates our data to the parameter(s).

- $\text{Binom}(8, p)$ in “Lady tasting tea”
- $X_i \sim D(\mu, \sigma)$ for mean inference
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ for a linear model

Let's look at our data points around the line.



The points should look normally distributed around the line.

Our full model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma)$$

Note:

- Errors centered at 0
- σ is constant

Three assumptions make this model accurate:

- X and Y have a linear relationship.
- The errors are normal with mean 0
- The spread around the line is constant for all x .

We evaluate these assumptions with the residuals.

Make a plot of residuals $y_i - \hat{y}_i$ versus x .

The points should be scattered in a random “cloud”.

In R, get the residuals from our `lm` with `resid()` and the predicted values with `predict()`.

Linear models are commonly used for prediction.

The Riley dataset does not have a point for $x = 78$ (six and a half). How tall was Riley at this point?

$$\hat{y} = 30.25 + 0.25(x)$$

We predict that Riley was

$$30.25 + 0.25(78) = 49.75$$

inches tall at six and a half. This is the height of the line at $x = 78$.

Later, we'll learn about the error in this estimate.

Let's go back to the Lake Monona data

- Predict the duration of the freeze in a future year.

What assumption are we making for this prediction?

We have only validated the linear model in the original range of our data.

We know that it's reasonable for years 1855-2025, but what about afterwards?

Trying to predict outside of the observed x values is called **extrapolation**.

Don't extrapolate too far from the original data.

When does the model stop being valid?

- No specific point
- Are we still willing to accept that the model is appropriate?