Linear Regression

Modeling two numeric variables

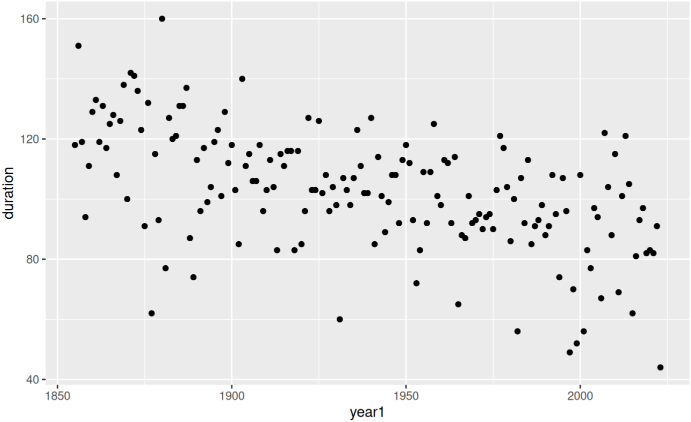
Download the section 12 .Rmd handout to `STAT240/lecture/sect12-regression-intro`.

Download the files `lake-monona-winters-2024.csv` and `riley.txt` to `STAT240/data`.

Linear regression is a tool used to model the relationship between 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 (strength and direction) between two variables.

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a sample of n points and let $\bar{x}, s_x, \bar{y}, s_y$ be the sample means and SDs of the X and Y values.

The 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 unitless and is always within $[-1, 1]$.

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

Typically, $|r| > 0.8$ is considered strong correlation, $0.5 < |r| \leq 0.8$ is considered moderate, and $|r| \leq 0.5$ is considered weak.

Correlation is the measure of the linear relationship between two variables.

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

Remember correlation does not equal causality!

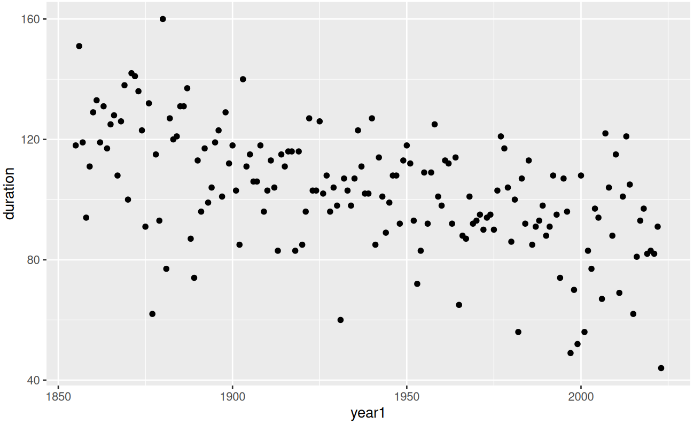
We use correlation to build a **linear model**, which has a slope and a y-intercept:

$$y = mx + b$$

In statistics, we use the notation

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

β_1 is the linear relationship between x and y.



What is the straight line that best describes the relationship between year and duration?

A linear model would look 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.
In general,

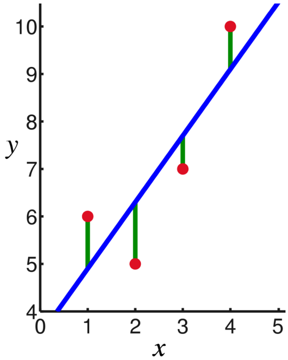
$$y_i \;=\; \beta_0 + \beta_1 x_i + \epsilon_i$$

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

- β_0 and β_1 are unknown coefficients
- ϵ_i are unknown errors

How should we estimate β_0 and β_1?

We minimize the *vertical* distance from the points to the 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

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

We pick the $\hat{\beta}_0$ and $\hat{\beta}_1$ that give us the smallest **sum of squared** residuals.

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

We use calculus to find formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize SS_E.

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

- 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` gives a boy's height in inches and age in months over several years.

- Filter from age 2 years to 8 years

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.

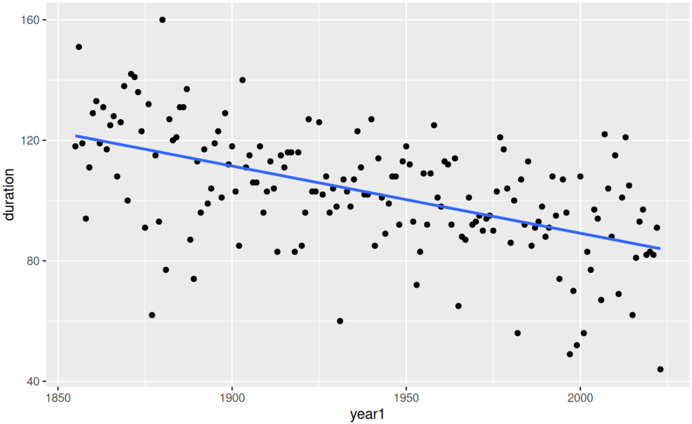
$$\hat{y}_i \;=\; \hat{\beta}_0 + \hat{\beta}_1 x_i$$

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

Formally, a **model** relates our observed data to an unknown parameter.

- e.g. *Binom*$(8, p)$ in "Lady tasting tea"
- For a linear model, β_0, β_1, and ϵ_i

Let's look closer at the errors. How are the points "distributed" around the line?



We assume the points are normally distributed around the line.

Our full model is

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

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

Errors centered at 0 means points vary equally above and below the line.

Three assumptions make this model accurate:

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

We evaluate these assumptions by looking at the residuals.

We can obtain the residuals from our lm with `resid` and the predicted values with `predict`.

Make a plot of residuals with x on the x-axis and residuals $y_i - \hat{y}_i$ on the y-axis.

The points should be scattered in a random "cloud".

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 of the 2026-2027 winter.

For this prediction, we are making a very strong assumption about our model.

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

We know that it's reasonable for years 1855-2024, 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.

What does the model stop being valid?

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