# Linear Regression

Foundations of Data Analysis

March 16, 2021

# Pfizer COVID-19 Vaccine Trial

Pfizer enrolled 43,548 participants, half received the vaccine, half received a placebo. [1]

Of the 18,508 completing vaccination, 9 got COVID-19
Of the 18,435 completing placebo, 169 got COVID-19

Was the vaccine effective?

[1] https://www.nejm.org/doi/full/10.1056/NEJMoa2034577

# Pfizer COVID-19 Vaccine Trial

Risk ratio:

$$\begin{aligned} \mathrm{RR} &= \frac{\text{risk of COVID-19 with vaccine}}{\text{risk of COVID-19 with placebo}} \\ &= \frac{9/18508}{169/18435} \\ &\approx 0.053 \end{aligned}$$

Vaccine Efficacy $= 1 - \mathrm{RR} \approx 0.947$

# Pfizer COVID-19 Vaccine Trial

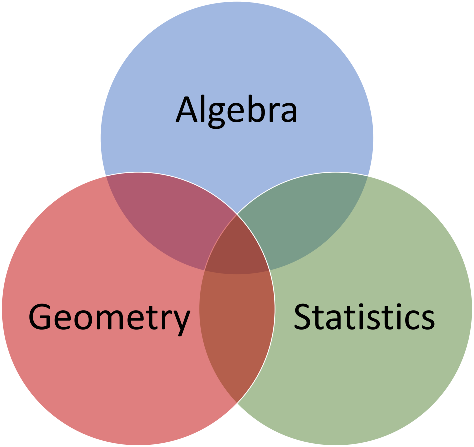
Contingency table:

| | Vaccine | Placebo |
|---|---|---|
| Positive | 9 | 169 |
| Negative | 18,499 | 18,266 |

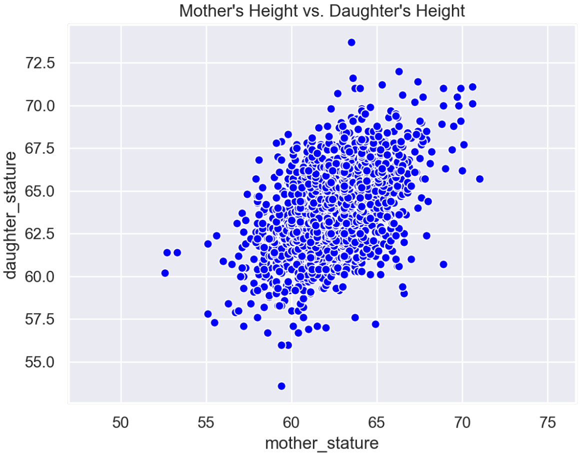
Using hypergeometric probability, $p(k)$, the $p$-value is:

$$P(X \leq 9) = \sum_{k=0}^{9} p(k) < 2 \times 10^{-16}$$

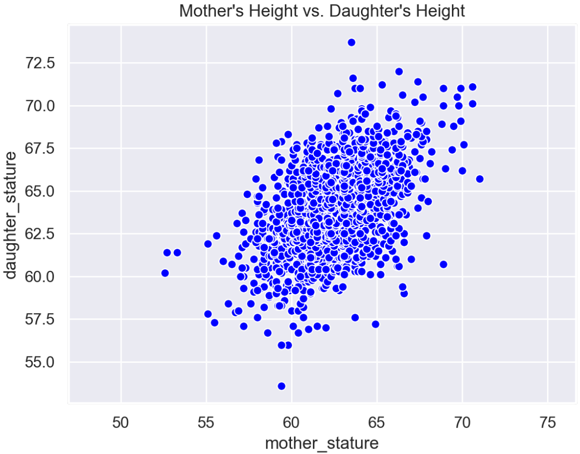
This is the probability for this result, or better, by random chance if the vaccine were not effective.


Algebra
Geometry
Statistics

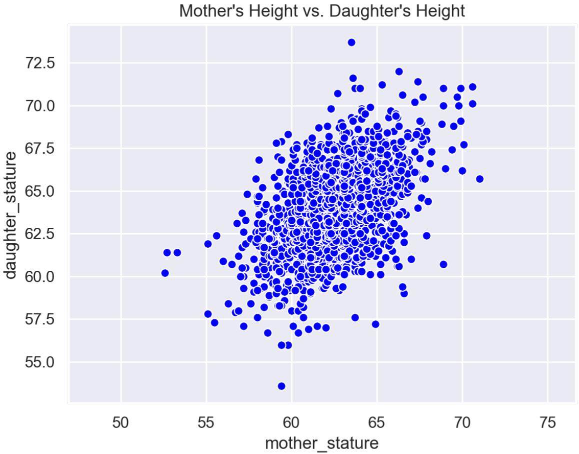
# Is there a relationship between the heights of mothers and their daughters?



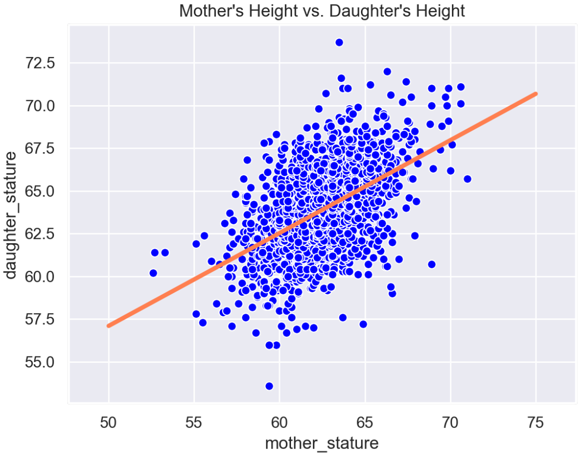
# If you know a mother's height, can you predict her daughter's height with any accuracy?



**Linear regression** is a tool for answering these types of questions.



# It models the relationship as a straight line.



# Regression Setup

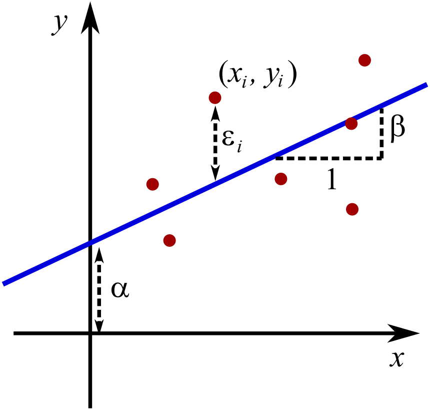
When we are given real-valued data in pairs:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n) \in \mathbb{R}^2$$

Example:
$x_i$ is the height of the $i$th mother
$y_i$ is the height of the $i$th mother's daughter

# Linear Regression

Model the data as a line:



$$y_i = \alpha + \beta x_i + \epsilon_i$$

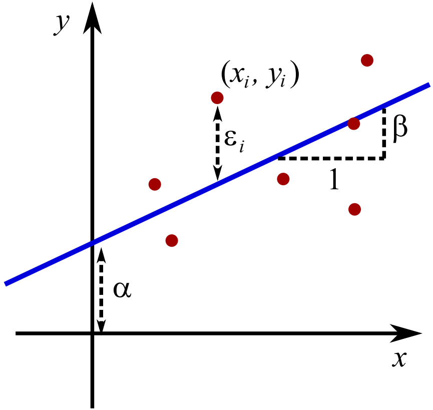
$\alpha$ : intercept

$\beta$ : slope

$\epsilon_i$ : error

# Geometry: Least Squares

We want to fit a line as close to the data as possible, which means we want to **minimize the errors**, $\epsilon_i$.



$$y_i = \alpha + \beta x_i + \epsilon_i$$

$\alpha$ : intercept

$\beta$ : slope

$\epsilon_i$ : error

# Geometry: Least Squares

Taking the line equation: $y_i = \alpha + \beta x_i + \epsilon_i$

Rearrange to get: $\epsilon_i = y_i - \alpha - \beta x_i$

We want to minimize the **sum-of-squared errors (SSE)**:

$$\mathrm{SSE}(\alpha, \beta) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

# Least Squares: Step 1

Center the data by removing the mean:

$$\tilde{y}_i = y_i - \bar{y}$$

$$\tilde{x}_i = x_i - \bar{x}$$

Note: $\sum_{i=1}^{n} \tilde{y}_i = 0$ and $\sum_{i=1}^{n} \tilde{x}_i = 0$

We'll first get a solution: $\tilde{y} = \alpha + \beta\tilde{x}$, then shift it back to the original (uncentered) data at the end

# Least Squares: Step 2

Take derivative of $\mathrm{SSE}(\alpha, \beta)$ wrt $\alpha$ and set to zero:

$$
\begin{aligned}
0 = \frac{\partial}{\partial \alpha} \mathrm{SSE}(\alpha, \beta) &= \frac{\partial}{\partial \alpha} \sum_{i=1}^{n} (\tilde{y}_i - \alpha - \beta \tilde{x}_i)^2 \\
&= -2 \sum_{i=1}^{n} (\tilde{y}_i - \alpha - \beta \tilde{x}_i) \\
&= -2 \sum_{i=1}^{n} \tilde{y}_i + 2n\alpha + 2\beta \sum_{i=1}^{n} \tilde{x}_i
\end{aligned}
$$

Using $\sum \tilde{y}_i = \sum \tilde{x}_i = 0$, we get

$$\hat{\alpha} = 0$$

# Least Squares: Step 3

With $\alpha = 0$, we are left with

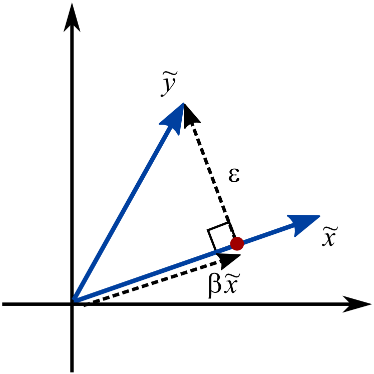
$$\tilde{y}_i = \beta \tilde{x}_i + \epsilon_i$$

Or, in vector notation:

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_n \end{bmatrix} = \beta \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

# Least Squares: Step 3

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \vdots \\ \tilde{y}_n \end{bmatrix} = \beta \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \vdots \\ \tilde{x}_n \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$



Minimizing $\mathrm{SSE}(\alpha, \beta) = \sum \epsilon_i^2 = \|\epsilon\|^2$ is projection!

Solution is $\hat{\beta} = \frac{\langle \tilde{x}, \tilde{y} \rangle}{\|\tilde{x}\|^2}$

# Shifting Back to Uncentered Data

So far, we have:

$$\tilde{y}_i = \hat{\beta}\tilde{x}_i + \epsilon_i$$

Expanding out $\tilde{x}_i$ and $\tilde{y}_i$ gives

$$(y_i - \bar{y}) = \hat{\beta}(x_i - \bar{x}) + \epsilon_i$$

Rearranging gives

$$y_i = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta}x_i + \epsilon_i$$

So, for the uncentered data, $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

# Probability: Maximum Likelihood

So far, we have only used geometry, but if our data is random, shouldn't we be talking about probability?

To make linear regression probabilistic, we model the errors as Gaussian:

$$\epsilon_i \sim N(0, \sigma^2)$$

The likelihood is

$$L(\alpha, \beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

# Probability: Maximum Likelihood

The log-likelihood is then

$$\log L(\alpha, \beta) = -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \epsilon_i^2 + \text{const.}$$

Maximizing this is equaivalent to minimizing SSE!

$$\max \log L = \min \sum \epsilon_i^2 = \min \text{SSE}$$