Lady Drinking Tea - Fisher's Exact Test

Present the lady with eight cups of tea with milk, where four cups have milk added first and the rest cups have tea added first. She then tasted each cup and reported which four she thought had milk added first.

Question: how do we test whether she really is skilled at this or if she's just guessing?

Null Hypothesis Testing

Null hypothesis H_0: the lady could not really tell the difference between teas, and she is just guessing.

Hypothesis testing: to disprove or reject H_0 with p-value.

Null Hypothesis Testing

Assuming the null hypothesis is true, we can compute the probability of her performing as well as she did or better.

Suppose we have *K* cups with milk first that she gets correct:

Contingency table: provides interrelation with two or more variables.

		Lady's Answer	
		Milk First	Tea First
Truth	Milk First	k	4 - k
	Tea First	4 - k	k

A value of *k means the lady gets 2k correct.*

Null Hypothesis Testing

		Lady's Answer	
		Milk First	Tea First
Truth	Milk First	k	4 - k
	Tea First	4 - k	k

$$p('2k\,\text{correct}') = \frac{C_4^k \cdot C_4^{4-k}}{C_8^4}$$

Null Hypothesis Testing

What is the probability of her getting this outcome *or better*?

$$p(\text{'}2k \text{ correct or better'}) = P(X \geq 2k) = \sum_{i=k}^{4} P(i)$$

Null Hypothesis Testing

In general, if we had *N* cups of tea, *K* cups with milk first. The lady selects *n* cups as ones she thinks are with milk first, getting *k* correct. The chances of getting these k "successes" by just choosing at random with uniform probability is

$$\boxed{p(X=k)=\frac{C_K^k \cdot C_{N-K}^{n-k}}{C_N^n}}$$

Hypergeomeric distribution

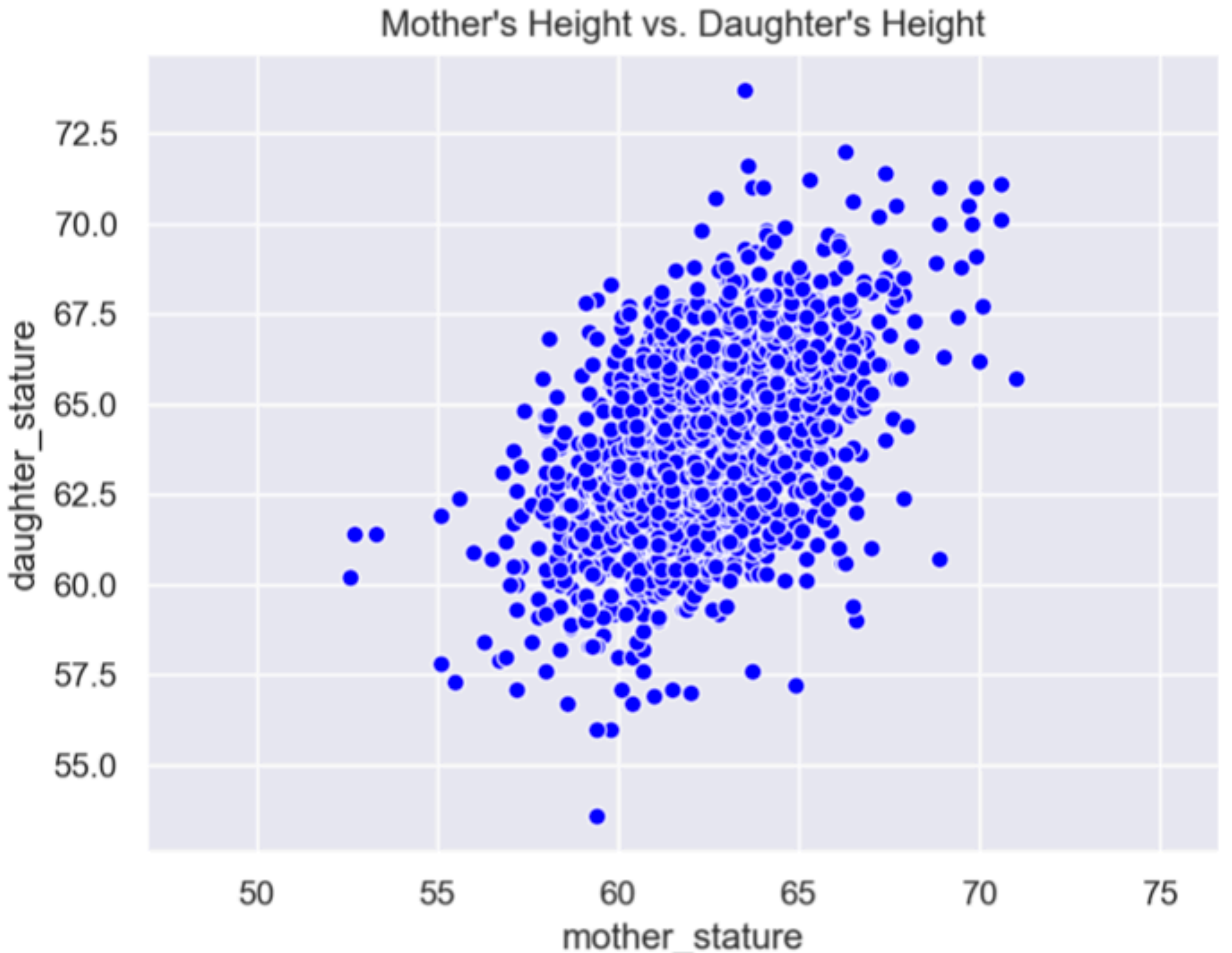
https://en.wikipedia.org/wiki/Hypergeometric_distribution

Linear Regression

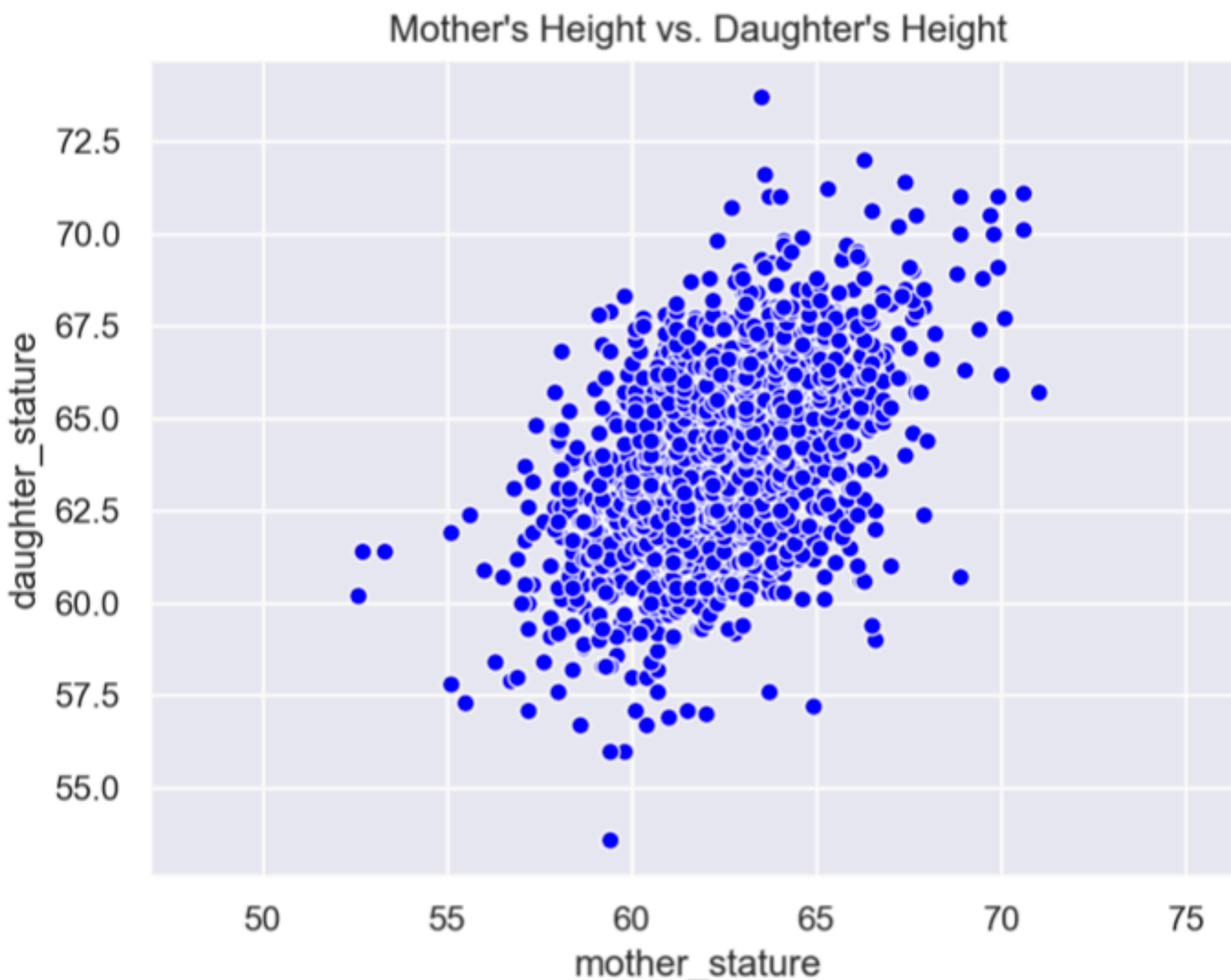
Foundations of Data Analysis

February 22, 2023

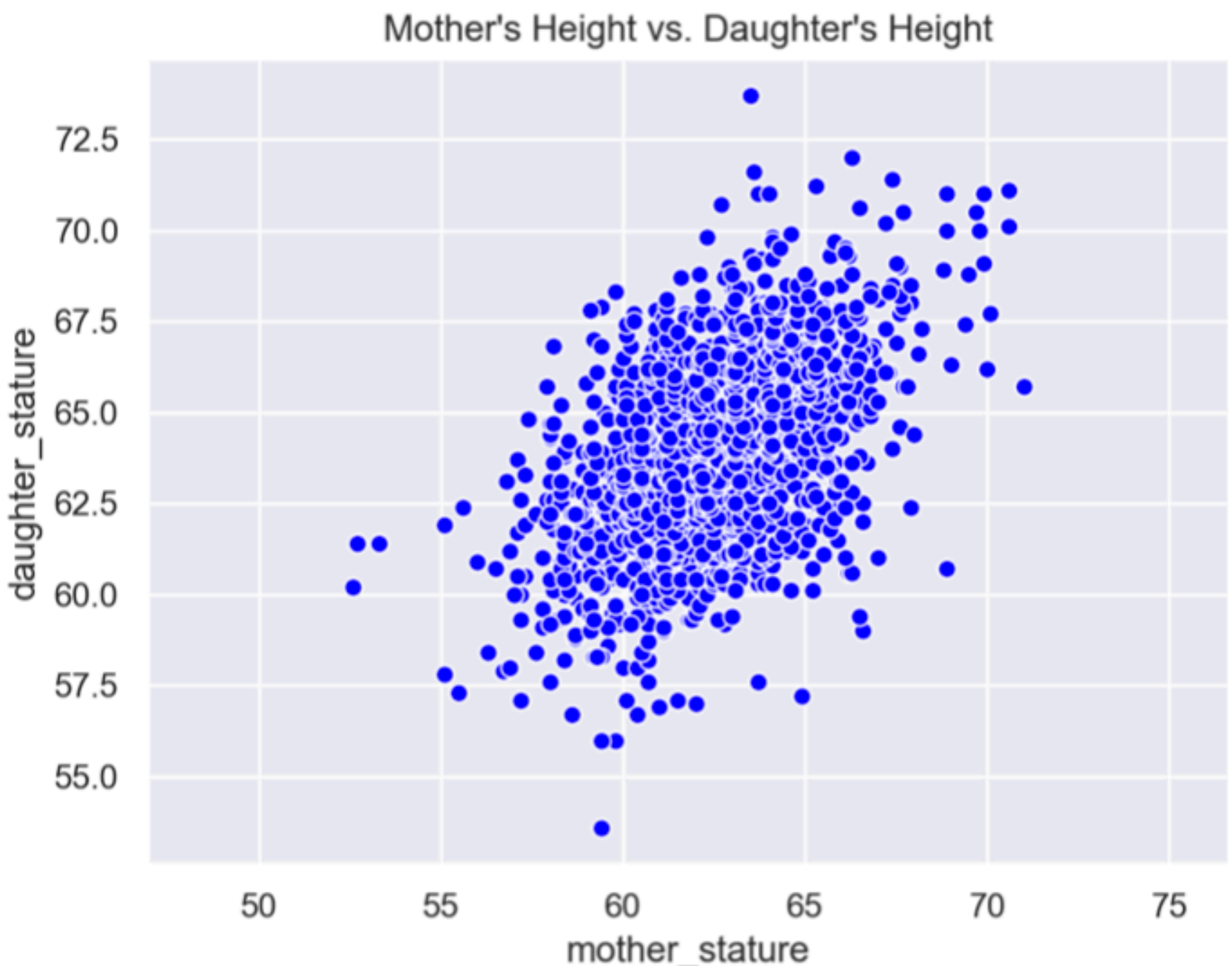
Is there a relationship between the heights of mothers and their daughters?



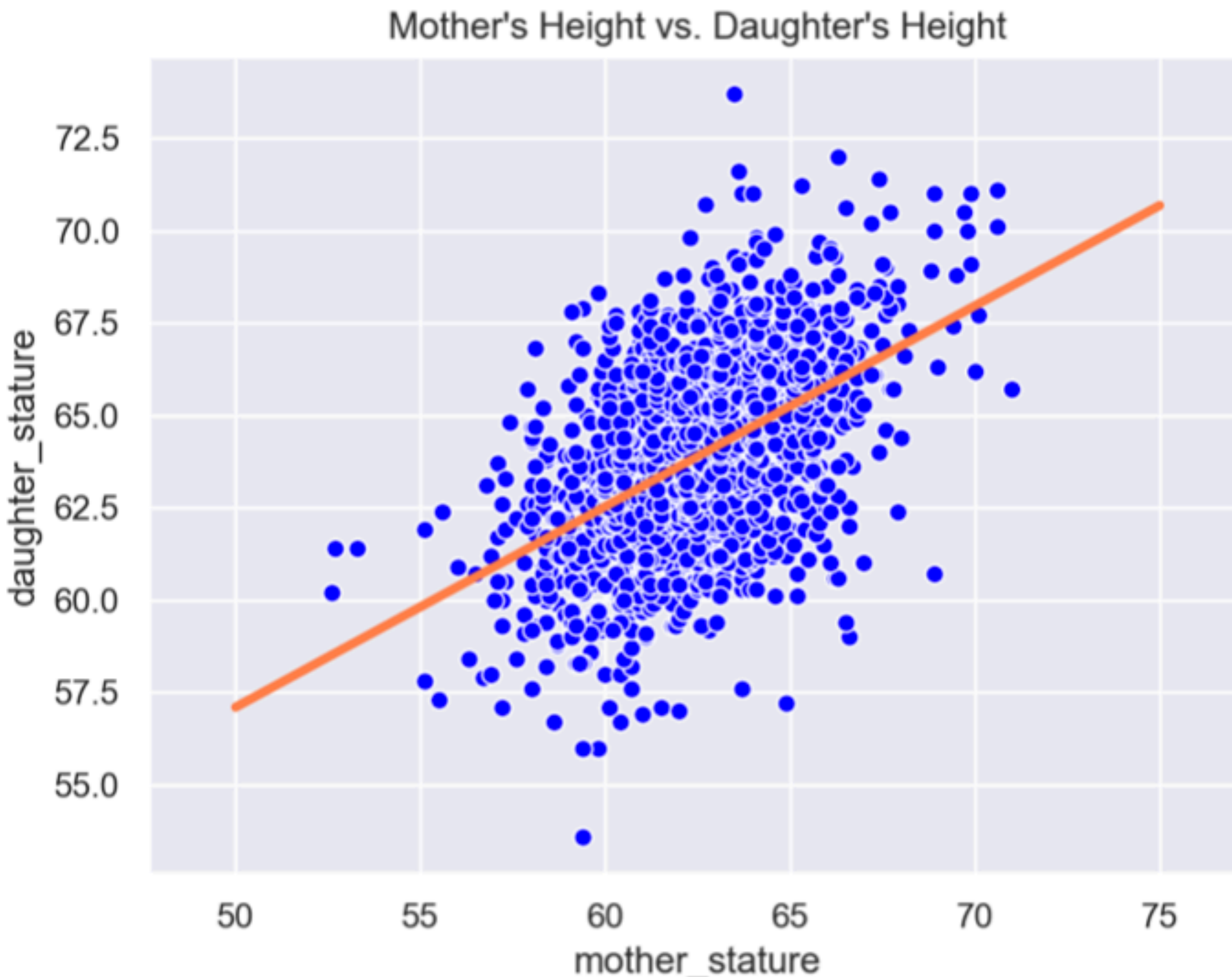
If you know a mother's height, can you predict her daughter's height with any accuracy?



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

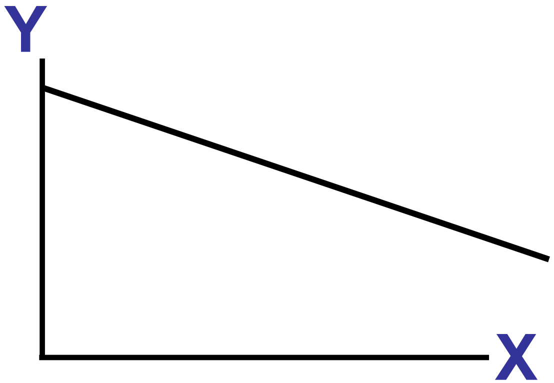


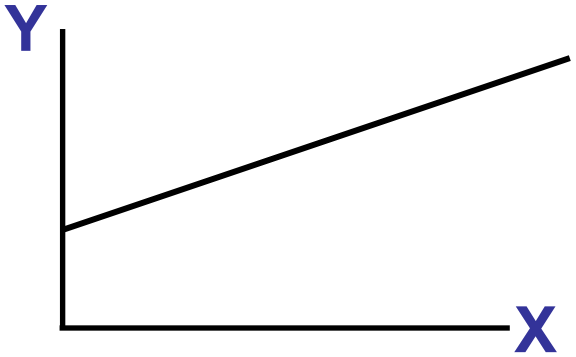
It models the relationship as a straight line.



Types of Relationships Between Two Variables

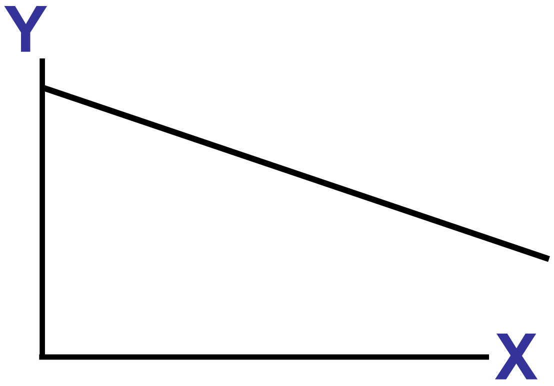
- Positive and negative relationships

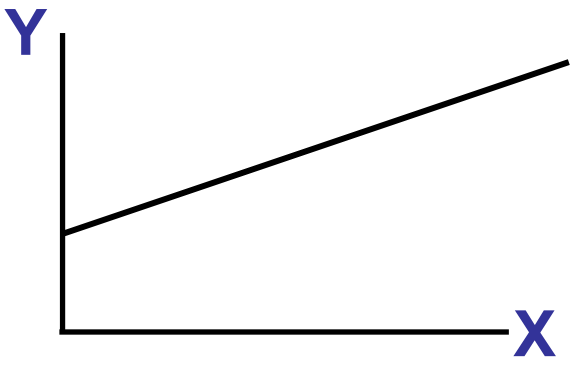




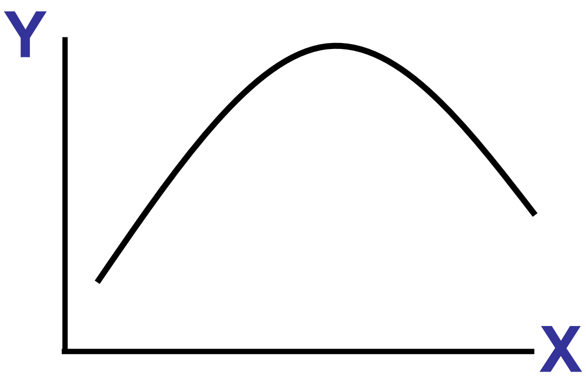
Types of Relationships Between Two Variables

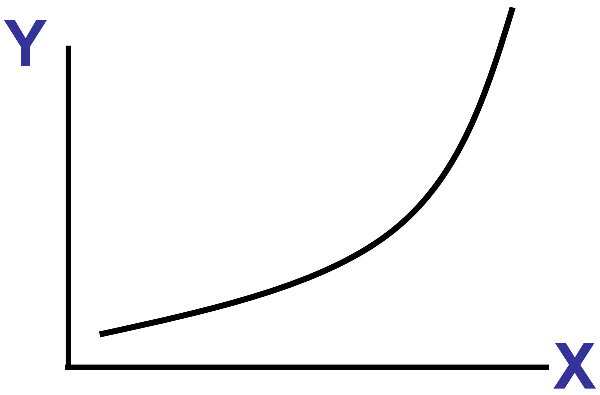
- Positive and negative relationships





- Curve relationships





Regression Setup

When we are given real-valued data in pairs:

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

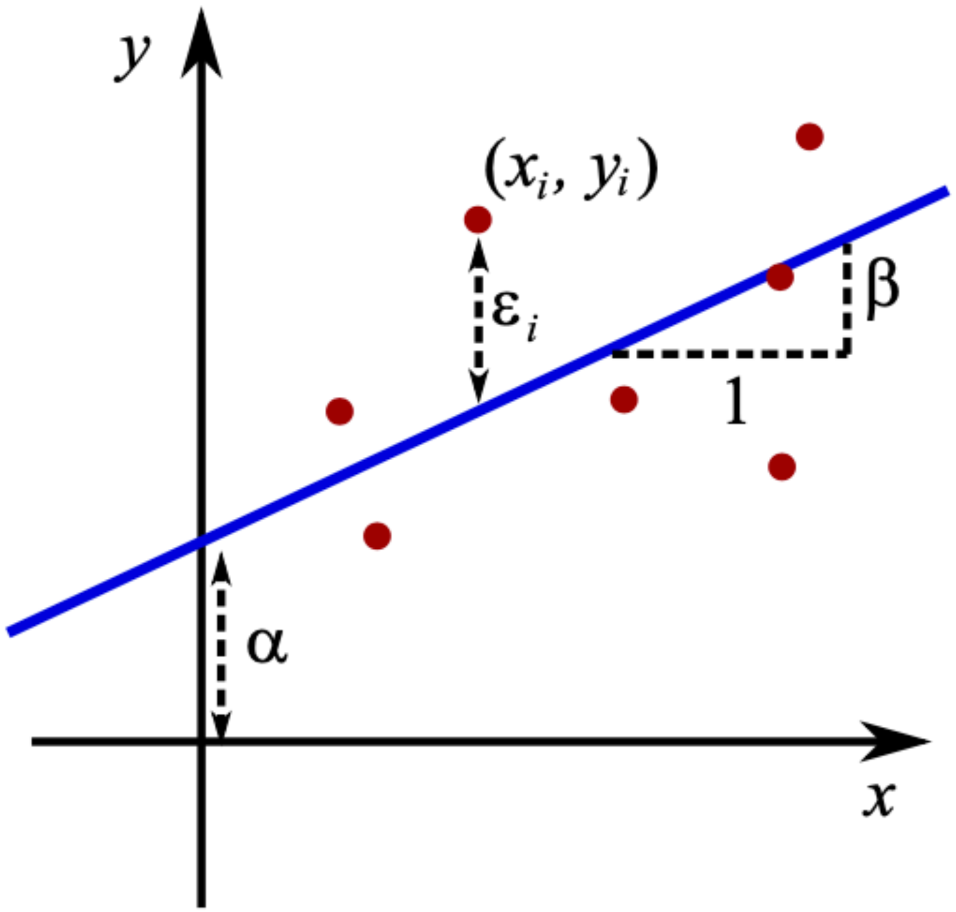
Example:

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:



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

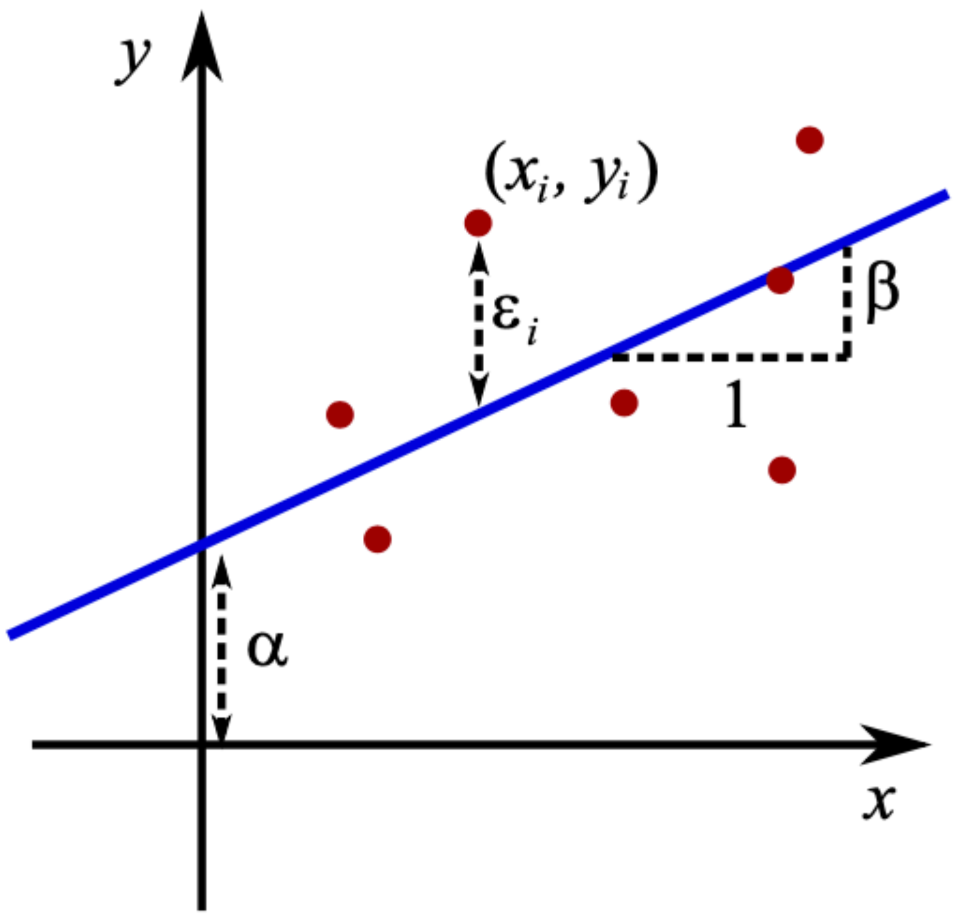
α : intercept

β : slope

ϵ_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**, ϵ_i



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

α : intercept

β : slope

ϵ_i : *error*

Geometry: Least Squares

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

Geometry: Least Squares

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

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

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)**:

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

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 $SSE(\alpha, \beta)$ wrt α and set to zero:

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

Least Squares: Step 2

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

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

$$= -2 \sum_{i=1}^{n} (\tilde{y}_i - \alpha - \beta \tilde{x}_i)$$

Least Squares: Step 2

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

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

$$= -2 \sum_{i=1}^{n} (\tilde{y}_i - \alpha - \beta \tilde{x}_i)$$

$$= -2 \sum_{i=1}^{n} \tilde{y}_i + 2n\alpha + 2\beta \sum_{i=1}^{n} \tilde{x}_i$$

Least Squares: Step 2

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

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

$$= -2 \sum_{i=1}^{n} (\tilde{y}_i - \alpha - \beta \tilde{x}_i)$$

$$= -2 \sum_{i=1}^{n} \tilde{y}_i + 2n\alpha + 2\beta \sum_{i=1}^{n} \tilde{x}_i$$

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

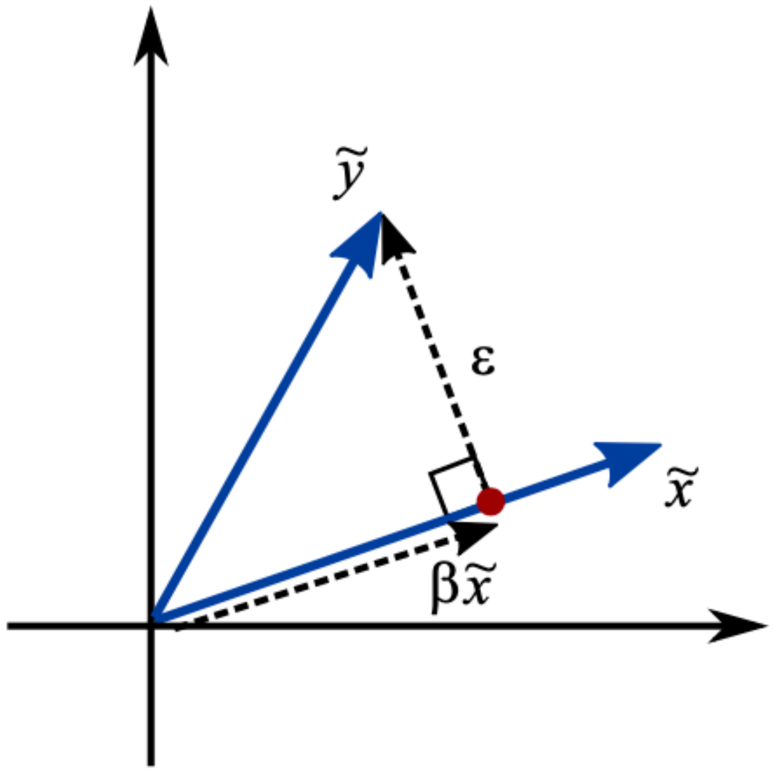
$$\tilde{y}_i = \beta \tilde{x}_i + \epsilon_i$$

Or, in vector notation:

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \cdots \\ \tilde{y}_n \end{bmatrix} = \beta \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \cdots \\ \tilde{x}_n \end{bmatrix} + \begin{bmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \cdots \\ \tilde{\epsilon}_n \end{bmatrix}$$

Least Squares: Step 3

$$\begin{bmatrix} \tilde{y}_1 \\ \tilde{y}_2 \\ \cdots \\ \tilde{y}_n \end{bmatrix} = \beta \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \cdots \\ \tilde{x}_n \end{bmatrix} + \begin{bmatrix} \tilde{\epsilon}_1 \\ \tilde{\epsilon}_2 \\ \cdots \\ \tilde{\epsilon}_n \end{bmatrix}$$



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

Solution is $\hat{\beta} = \dfrac{< \tilde{x}_i, \tilde{y}_i >}{||\tilde{x}||^2}$

Shifting Back to Uncentered Data

So for, we have:

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

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

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

Shifting Back to Uncentered Data

So for, 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

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

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 equivalent to minimizing SSE!

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