The Strength Upgrade (r^2)

Bill the Statypus says: We've been calling r^2 the "Strength" of a relationship. But what is it actually measuring? Today we use our fingerprints to see how a model physically cleans up a messy dataset.

1. The Blind Guess: The Bar ($\bar{y}$)

Imagine you have a bucket of fish. You don't know their heights, but you have to guess their weights. Your "best guess" for every fish is simply the mean weight: $\bar{y}$.

Statypus Insight: Visual Mnemonic

Notice the bar over the y in $\bar{y}$. It is a perfectly horizontal line. In statistics, $\bar{y}$ represents a perfectly flat guess that ignores the explanatory variable entirely.

2. The Informed Guess: The Hat ($\hat{y}$)

Now, imagine you get an upgrade. You are allowed to see the height of the fish before you guess its weight. You now use your prediction machine: $\hat{y}$.

Statypus Insight: Visual Mnemonic

Notice the "hat" over the y in $\hat{y}$. The sides of the hat are slanted. In statistics, $\hat{y}$ represents the slanted regression line that tilts to follow the story of the data.

The Cleanup

In your Desmos exploration, follow these steps:

1. Drag points around until r is exactly **0.80** (or at least as close as you can get it).
2. Record the "Sum of Squared Residuals" (SSR): ____________________
3. Record the "Sum of Squared Deviations": (SSD) ____________________

3. Calculating the Improvement

We want how much better $\hat{y}$ is than $\bar{y}$ at predicting values of y if you know the value of x.

$$\text{Percentage Improvement} = \frac{SSD - SSR}{SSD} = \underline{\hspace{4cm}} \approx \qquad\qquad \%$$

Sally the Statypus says: That is the same as r^2! This is why we call r^2 the **Coefficient of Determination**. It tells us exactly what percentage of the variation in our response is explained by our model.

4. Interpreting the Story

Bill the Statypus says: When a scientist reports r^2, they are explaining how much of the story they have successfully decoded. We often use a specific sentence structure to communicate this.

Reflection: The Universal Template

If x and y have a correlation of r, then r^2 represents the amount of variation that can be explained by or removed from predictions of y if we know the corresponding value of x.
As an example, if $r = 0.9$**,** then 81% of the variation of y can be accounted for (and essentially removed) by using $\hat{y}$ and the corresponding value of x. Or more succinctly as: **81% of the variation of** y **can be attributed to its relationship to** x**.**

5. The Mystery Remaining

Bill the Statypus says: If $r = 0.8$, it means the fish's height explains 64% of the variation of the fish's weight. But that means 36% of the variation is still a mystery. The prediction isn't perfect because the real world is messy.

Reflection: The Residual Mystery

Think back to the human height/weight analogy. If our model has an r^2 of 0.60 (60% explained), list reasons why two people might be the exact same height but have very different weights.

Statypus Insight: Extreme Logic

- If $r = \pm 1$, then 100% of the variation of y is removed if we know the value of x. That is, x perfectly predicts y and the points must all be on a perfect straight line.
- If $r = 0$, then none of the variation of y is accounted for by x. That is, knowing x offers no assistance in understanding y.

Sally the Statypus says: Remember that real science lies in between these extremes. Often one variable does help us improve our estimates of another variable because they are related!