# The Prediction Machine

**Statypus Insight:**

In Section 5.1, we manually drew lines through data clouds. In Section 5.4, we formalize this. A truly balanced regression line isn't just "in the middle"—it is mathematically tethered to the center of gravity of your data: the point $(\bar{x}, \bar{y})$.

## 1. Finding the Anchor

Before the machine can calculate a slope, it must know its fixed point. Use R and the `FishMarket` dataset to find the coordinates of the center of your data.

Mean Height ($\bar{x}$): _____ Mean Weight ($\bar{y}$): _____

## 2. Engineering the "Best" Line

Open your textbook to **Section 5.4**. Read through the opening paragraphs and Section **5.4.1** to get a handle on what we mean by the "sum squared error." **Where have we seen the technique of squaring values to avoid cancellation before?**

**Reflection: The "On Average" Logic**

In everyday English, we use a specific phrase that perfectly describes what a regression line is doing. We might say, "Taller people are heavier," but we know that's not true for everyone. What we really mean is that taller people are heavier **on average**.

In statistics, we use the symbol $\hat{y}$ (read: "y-hat") to represent this "average" story.

- A **positive correlation** shows that the average $y$, which is $\hat{y}$, increases as $x$ increases.
- A **negative correlation** shows that the average $y$ decreases as $x$ increases.

**Human Analogy:** If we look at 1,000 people who are all 5'9", they will have a distribution of weights (some light, some heavy). Our model predicts the **mean** of that distribution. If we move to the 6'3" group, we expect the entire distribution to shift higher.

**The Reality Check:** If our model predicts a weight of 500g for a 15cm fish, but we find a real fish that weighs 600g, did our model "fail"? Use the logic above to explain.

## 2. Assembling the Machine: `LinearRegression()`

**Bill the Statypus says:** You've already seen that `plot()` and `cor()` use the same syntax. To build the actual machine, we just use one more tool that follows the exact same pattern.

**Sally the Statypus says:** In this class, we define our prediction machine as $\hat{y} = a + bx$.
$a$ is the **Intercept** (where the machine starts).
$b$ is the **Slope** (how fast the response changes for every 1 unit of $x$).

**Coding Corner:**

Turn to Section 5.4.2 to find the custom function used to build the regression line. Use it to predict Weight based on Height in the FishMarket data.

**Bill the Statypus says:** (blank stare) If you hit an error... load... the... .RData...

**Recording the Blueprints**
Use your new tool on the `Height` and `Weight` variables. The machine will hand you two numbers.

Intercept ($a$): _____ Slope ($b$): _____

## 3. The Manual Prediction

Substitute your blueprints into the formula below to complete your fish-predicting machine:

$$\widehat{\text{Weight}} = \_\_\_\_\_ + (\_\_\_\_\_) \times \text{Height}$$

**Manual Test**
A new fish arrives at the market with a Height of **15 cm**. Use your formula and a calculator to predict its weight.

Manual Prediction: _____ grams.

**Reflection: But what does that mean?!**

Write out, in full sentences, the meaning of your above calculation in a way that would make sense to a student first walking into a statistics classroom.

## 4. Automated Tooling: The $x_0$ Upgrade

**Precision Arguments**
In R, your custom function has an optional third argument called `x0`. This allows you to tell the machine exactly which $x$ value you want a prediction for. See Section 5.7.1 for the help file for `LinearRegression`.

**Coding Corner: Automated prediction for a Height of 15cm**

Find the mean weight of a fish that has a height of 15cm. Record the answer here.

**Reflection: The Reality Check**

1. Compare your answer from Page 2 to the computer's answer. Are they identical? **Y / N**
2. We find a real fish in the market that is 15cm tall and weighs **100g more** than your machine predicted. Does this mean your math is wrong, or is something else happening? (Think back to the human height/weight analogy).
3. The value of **Strength ($r^2$)**, which we calculated in the last block, shows us how much of the "story" of a fish's weight is being explained by its height. In this case, how much of the variation of a fish's weight is explained by its height?

**Sally the Statypus says:** You likely already read this in the book, but the value of $a$ may have no practical meaning if the variable used as $x$ cannot be 0. This is why a negative intercept shouldn't necessarily worry you!

**Bill the Statypus says:** Yeah, yeah... but we forgot to remind them in the last block that **Correlation Does Not Imply Causation**. If that point isn't already clear in your head, **PLEASE** go back to section 5.3 and review that concept.

**Sally the Statypus says:** (smacking her own forehard) Yeah... what Bill said.