UNIT 7 · BLOCK 2

Linear Regression

Drawing the one line that lets you predict.

Why this matters

Last unit, correlation told us a relationship **exists** — social media and wellbeing move together. Nice. But it can't answer the question everyone actually asks: "*If a student is on social media 6 hours a day, what's their wellbeing likely to be?*" For that you need a **line**. Regression draws the single best line through your dots so you can plug in one number and **predict** the other. Same data, one step further — and it's the engine behind weather forecasts, insurance pricing, GPA predictors, and most of the "data science" you'll ever meet.

By the end you can…

- Explain how **regression** goes beyond correlation — from "a relationship exists" to "predict one variable from another."
- Read the regression equation $\hat{Y} = a + bX$ and say in plain words what the **slope** and **intercept** mean.
- Explain what the **line of best fit** (least squares) minimizes, and what a **residual** is.
- Build the line in Google Sheets and **make a prediction** for any X value.

Interpret r^2 as the percent of variation one variable explains.

1. The big idea — from a number to a line

Correlation gives you a single number, ***r***, that says *how tightly* two things move together and in *which direction*. Useful — but it stops there. **Regression** takes that same relationship and lays a **straight line** right through the middle of the scatterplot. That line is the whole payoff: once you have it, **prediction is just plugging in a number** and reading off what the line says. Correlation describes; regression predicts.

Step	Question it answers	What you get
Correlation (last unit)	*Is* there a relationship, and how strong?	One number, *r*
Regression (today)	What's the best **line**, and what does it **predict**?	An equation, $\hat{Y} = a + bX$

The two are deeply linked — regression is built directly out of the correlation and the standard deviations from Unit 6. They're the same relationship looked at two ways: correlation is the snapshot, regression is the formula you can actually *use*.

2. The line of best fit

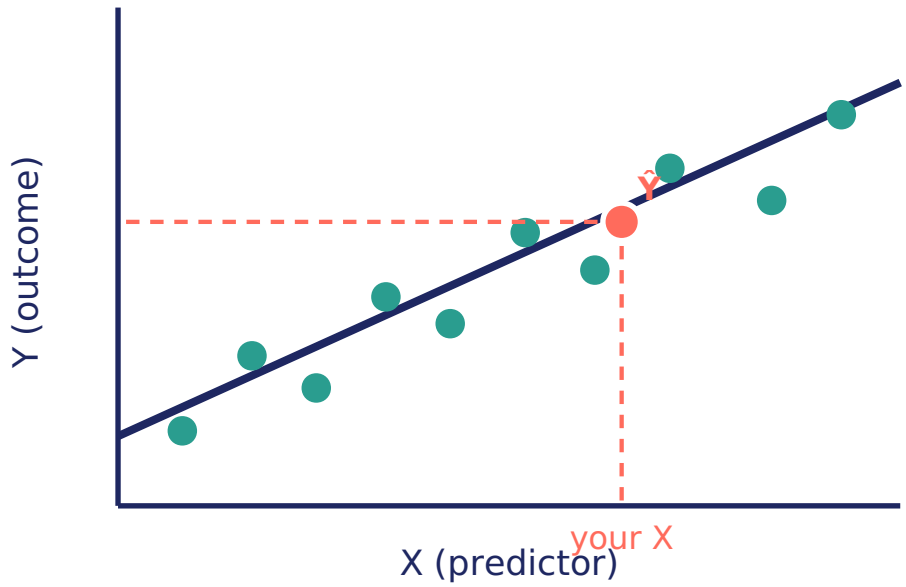
Picture the scatterplot of wellbeing vs. social media — 200 dots scattered in a cloud. You *could* draw a hundred different straight lines through that cloud. Which one is "best"? We need a rule, not a vibe.

The rule is this: for any line you draw, every dot sits some distance **above or below** it — that vertical gap is the line's **error** for that dot. The best line is the one that makes those errors, taken all together, as

small as possible. Because we don't want positives and negatives to cancel out, we **square** each gap before adding them up — which is exactly why this is called the **least-squares line**.

The least-squares line, in one sentence

The **least-squares line** is the single straight line that sits **closest to all the dots at once** — it makes the total of the squared vertical gaps between the dots and the line as small as it can possibly be. You don't compute it by hand; Sheets finds it for you. Just know what it is: **the one best line through the data.**



Each **teal dot** is one student. The **navy line** is the least-squares line of best fit. To predict, pick an X on the bottom axis, go up to the line, and slide across to read **Ŷ**, the predicted Y (the coral point and dashed guides).

3. The equation: $\hat{Y} = a + bX$

Here's the beautiful part — that whole line is captured by just **two numbers**. The equation is:

$$\hat{Y} = a + bX$$

- **$\hat{Y}$** ("y-hat") = the **predicted** value of Y. The little hat means "estimated," not actually observed.
- **X** = the value you plug in (the predictor you know).
- **b** = the **slope** — how much $\hat{Y}$ changes for **each 1-unit increase in X**. This is the headline of any regression.
- **a** = the **intercept** — the predicted $\hat{Y}$ when **X = 0** (where the line crosses the left axis).

Two numbers and you can predict. The **slope b** is the rate: "*for every one step right on X, the line rises (or falls) by b.*" The **intercept a** is just the line's starting height — its value when X is zero. Get a and b from your data and the line is fully pinned down.

Slope = rise per step

If **b = 0.5**, then every 1-unit increase in X raises the prediction by half a point. If **b = −1.43**, every 1-unit increase in X *lowers* the prediction by 1.43 points — the minus sign means the line slopes **down**. **Read the sign for direction, read the size for steepness.**

CHECK

In the equation $\hat{Y} = a + bX$, which letter is the slope — the change in predicted Y for each 1-unit increase in X?

b. The slope b is the change in $\hat{Y}$ per one-unit step in X — the rate of the line. a is the intercept, the predicted Y when X = 0 (where the line starts on the left).

4. Making a prediction

This is the moment regression earns its keep. Plug an X into the equation and do the arithmetic — that's the whole skill. For our data, predicting **wellbeing** (M) from **social media hours** (H), the line comes out to **$\hat{Y} = 28.80 - 1.43X$**. So what wellbeing would we predict for a student on social media **6 hours** a day?

$$\begin{aligned}\hat{Y} &= 28.80 - 1.43 \times 6 \\ &= 28.80 - 8.58 \\ &= 20.2\end{aligned}$$

So we'd predict a wellbeing score of about **20** for that student. Notice we did *not* need that exact person in our data — we read the prediction straight off the line. That's the power: one equation predicts Y for *any* X you hand it.

Read the slope first

The **slope of −1.43** is the headline: **each extra hour of daily social media predicts about 1.4 fewer wellbeing points.** The minus sign means the line slopes **down** — more hours, lower predicted wellbeing. Whenever you meet a regression, find the slope and translate it into a plain-English sentence like that one.

CHECK **A study-hours line is $\hat{Y} = 2.0 + 0.3X$, predicting GPA from hours studied per day. What GPA does it predict for a student who studies 4 hours?**

3.2. Plug in: $\hat{Y} = 2.0 + 0.3 \times 4 = 2.0 + 1.2 = 3.2$. The intercept (2.0) is the starting point; the slope (0.3) adds 0.3 GPA points for each hour studied, so four hours adds 1.2.

5. Residuals — the line's leftover error

No line threads through every dot perfectly. For each student, there's the wellbeing the line **predicted** ($\hat{Y}$) and the wellbeing they **actually** had (Y). The gap between them is the **residual**:

```
residual = actual Y − predicted Ŷ
```

A residual of **0** means the dot landed exactly on the line — a perfect call. A **positive** residual means the person scored *higher* than the line predicted (dot above the line); a **negative** residual means they scored *lower* (dot below). Those vertical gaps are precisely what least-squares makes as small as possible — squaring them first so a big miss in either direction counts heavily and positives can't cancel negatives.

Why "least squares" cares about residuals

The best-fit line is the one with the **smallest total squared residual**. Square each gap, add them all up, and no other straight line gives a smaller sum. That single idea is the whole engine — the residuals are the errors, and the line is chosen to crush them.

CHECK **The line predicts a student's wellbeing as $\hat{Y} = 20$, but the student's actual score was 17. What's the residual?**

−3. residual = actual − predicted = 17 − 20 = −3. It's negative, so the student scored *lower* than the line predicted — their dot sits below the line.

6. r^2 — how good is the line?

r^2 (*r*-squared, the *coefficient of determination*) tells you what **fraction of the variation in Y** the line actually explains. The recipe is

delightfully simple: **square the correlation** and you've got it. It's the bridge straight back to Unit 6.

Our wellbeing correlation was $r = -.46$, so $r^2 = (-.46)^2 = .22$ — about **22%** of the differences in students' wellbeing are explained by their social media use. The other 78%? Sleep, money, friends, classes, genetics — everything else regression didn't measure. So the line is real and useful, but social media is one slice of a much bigger pie.

And when we predict **loneliness** from social media (our headline $r = +.52$), $r^2 = (.52)^2 = .27$: about **27%** of the variance in loneliness is shared with social media use. A few anchors worth memorizing:

r^2	As a percent	What it means
0	0%	The line explains nothing; X is useless for predicting Y.
.22	22%	Our wellbeing line — modest but real.
.50	50%	The line explains half the variation — quite strong.
1.0	100%	Every dot is exactly on the line; perfect prediction.

r^2 is always between 0 and 1 (0% to 100%) — it can never be negative and never exceed 1, because squaring a correlation (itself between −1 and +1) lands you in that range. Bigger r^2 means a tighter, more trustworthy line.

CHECK

A correlation between sleep and exam scores is $r = .40$. What is r^2, and what does it mean?

$r^2 = .16$. Square the correlation: $(.40)^2 = .16$, so about 16% of the variation in exam scores is explained by sleep. r^2 is never the same as r unless r is 0 or 1.

Worked example — fit, predict, interpret

Let's walk a clean line end to end. Suppose a tutoring study fits **$\hat{Y}$ = 2.0 + 0.5X**, predicting an exam grade improvement (Y, in points) from the number of tutoring sessions attended (X). Predict the improvement at **X = 10** sessions, then say what each piece means in plain English. Predict each step before you reveal it.

Worked example — $\hat{Y}$ = 2.0 + 0.5X, find Y at X = 10

Plug in, simplify, then translate the slope and intercept into sentences.

STEP 1 — WRITE THE EQUATION WITH THE X FILLED IN

$\hat{Y}$ = 2.0 + 0.5 × `10`. The intercept (2.0) stays; the slope (0.5) gets multiplied by your X.

STEP 2 — DO THE MULTIPLICATION

0.5 × 10 = `5.0`. That's how much the slope contributes at X = 10.

STEP 3 — ADD THE INTERCEPT

$\hat{Y}$ = 2.0 + 5.0 = `7.0`. We predict about a **7-point** improvement for a student who attends 10 sessions.

STEP 4 — INTERPRET THE SLOPE (B = 0.5)

Each **additional tutoring session** predicts about `0.5` more points of improvement. Ten sessions about 5 extra points on top of the baseline.

STEP 5 — INTERPRET THE INTERCEPT (A = 2.0)

A student who attends **zero** sessions (X = 0) is predicted to improve by `2.0` points anyway — the baseline the line starts from. (Make sure X = 0 is realistic before you trust this.)

Now the same thing on **our** dataset. Open the **SJNY Student Life** dataset and predict **wellbeing** (M) from **social media hours** (H) using three functions — type them in any empty cell:

```
=SLOPE(M2:M201, H2:H201)          b = −1.43   (the slope)
=INTERCEPT(M2:M201, H2:H201)      a = 28.80   (the intercept)
=RSQ(M2:M201, H2:H201)            r² = .22    (22% explained)
```

Order matters

It's **(y-range, x-range)** — the thing you're **predicting** comes FIRST. Here that's wellbeing (M), then social media (H). Flip them and your slope and intercept come out wrong.

Stat in the Wild — the danger of extrapolation

A regression line is only trustworthy **inside the range of data you actually measured.** Step outside it and you get nonsense. Picture a toddler growing a few inches a year. Draw that line and stretch it forward 30 years — congratulations, you've predicted a **50-foot-tall adult.** The line was real *for toddlers*; it falls apart everywhere else. Anytime you hear "*if this trend continues...*" run far into the future, that's the joke — the trend almost never continues.

Stay inside the range

Our wellbeing line was built from students with roughly **0 to 8** hours of social media. Predicting wellbeing for someone at **20 hours a day** is **extrapolation** — outside the data, off the map. And remember: a great-fitting line still isn't proof that X *causes* Y. Regression predicts; it never proves cause.

The fine print — when the line is trustworthy

Regression inherits correlation's assumptions and adds one of its own. **Linearity** — the line is only a fair summary if the cloud is actually line-shaped. **Independent observations** — one row per case. **Watch for outliers** — a single extreme point can grab the line and tilt it, especially at the edges of the X range. And the new one: **even scatter along the line** — the residuals should be about as puffy at low X as at high X (the ten-dollar word is *homoscedasticity*; the plain-English version is "even noise"). If the cloud fans out as X grows, the line still fits, but its predictions get much less trustworthy at the wide end. The scatterplot-first habit from Unit 6 screens every one of these.

In Google Sheets

You don't fit lines by hand — Sheets does it. Open the **SJNY Student Life** dataset. Four functions cover everything, and the order is always **(y, x)** — predicted variable first:

```
=SLOPE(M2:M201, H2:H201)                the slope b
=INTERCEPT(M2:M201, H2:H201)            the intercept a
=RSQ(M2:M201, H2:H201)                  r² (variation explained)
```

To let Sheets do the prediction for you — no arithmetic at all — use `=FORECAST` (newer Sheets call it `=FORECAST.LINEAR`) or its cousin `=TREND`:

```
=FORECAST(6, M2:M201, H2:H201)          20.2   (predicted wellbeing at X = 6)
=FORECAST.LINEAR(6, M2:M201, H2:H201)   20.2   (same thing, newer name)
=TREND(M2:M201, H2:H201, 6)             20.2   (predicts Y for the new X = 6)
```

Note that `=FORECAST` wants **(new-X, y-range, x-range)** — the X you're predicting *for* comes first, then the same (y, x) ranges. To **graph the line**: highlight your two columns, then **Insert ▸ Chart**, pick a **scatter chart**, and in the Chart editor turn on the **Trendline**

(Customize ▸ Series) — and check the box to show the equation and r^2 right on the chart. There's your line of best fit, drawn for you.

Try it yourself — Linear Regression

Paste a predictor and an outcome, hit Run, and the calculator builds $\hat{Y} = a + bX$ — then predicts Y for any X you choose. Hit **Load class data** to see a real result first.

Watch out for

Order in the functions is **(y, x)** — predicted variable first. The intercept can be meaningless: "predicted wellbeing at 0 hours" might be fine, but "predicted GPA at 0 study hours" may sit outside the data. Don't extrapolate — predict inside the range, not miles beyond it. And r^2 is a *percent of variation explained*, not the slope: $r^2 = .22$ means 22% explained; the slope says steepness and direction. A good-fitting line still isn't proof that X causes Y.

Key terms

Regression · **Line of best fit** · **Least-squares line** · **Slope (b)** · **Intercept (a)** · **Ŷ (predicted y)** · **Prediction** · **Residual** · **r^2 / coefficient of determination** · **Extrapolation**

PRACTICE — NEVER GRADED

Check yourself

Ten quick true/false items with instant explanations. (Different from your Canvas quiz — real practice.)

Sigma says: The slope is the whole story of a regression — it's how much $\hat{Y}$ moves for each 1-step jump in X. Read its sign for direction and its size for steepness, then plug in a number to predict. And r^2 is just r squared: the share of the variation the line explains.