# Least Squares Regression and $R^2$

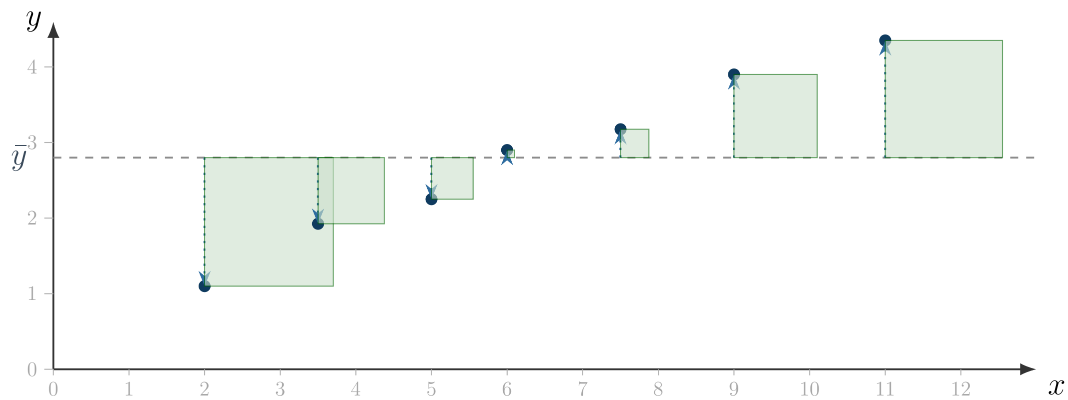
*Mr. Merrick · September 26, 2025*

## 1) Total Variance in $y$: squares to the mean (the "null model")

| Point | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| $x$ | 2.0 | 3.5 | 5.0 | 6.0 | 7.5 | 9.0 | 11.0 |
| $y$ | 1.10 | 1.925 | 2.25 | 2.90 | 3.175 | 3.90 | 4.35 |

The dashed horizontal line marks $\bar{y} = 2.800$. Each dotted arrow is a vertical deviation $(y_i - \bar{y})$. For every point, draw a **square** using that arrow as a side. The total area of all squares is

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(total variation in } y\text{)}.$$




**Record your total:** $\text{SST} = \sum (y_i - \bar{y})^2 =$ 7.7213

### Quick questions

1. The *null model* predicts every value of $y$ with $\bar{y}$. Does it take $x$ into consideration, or use $x$ to explain variation?

   It ignores $x$ entirely and predicts the same value $\bar{y}$ for every point (no relationship).

2. If we changed the units of $y$ (e.g. cm $\rightarrow$ m), how would the area of each square change?

   Areas scale by the square of the unit change since each side length (a deviation from $\bar{y}$) rescales.

3. For this dataset, does it look like there is a relationship between $y$ and $x$?

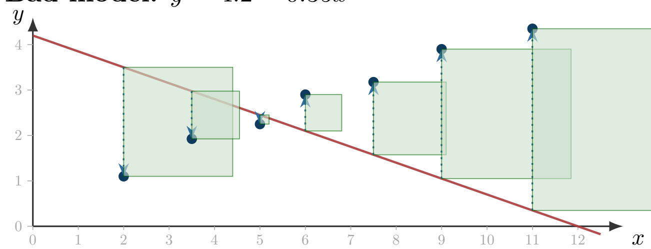
   Yes. The points rise with $x$ in an almost straight line—strong positive linear association.

## 2) Least Squares: choose a model to minimize squared residuals

A linear model predicts $\widehat{y} = a + bx$. Each residual is $e_i = y_i - \widehat{y}_i$ (Actual − Predicted — remember *AP*). We choose $(\hat{a}, \hat{b})$ that *minimizes* the total *sum of squared errors.* Draw squares for each model's residuals.

$$\mathrm{SSE}(a,b) = \sum_{i=1}^{n} \Big(y_i - \hat{y}_i\Big)^2 = \sum_{i=1}^{n} \Big(y_i - (a + bx_i)\Big)^2.$$

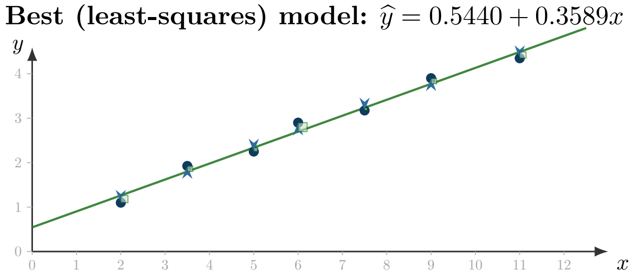
**Bad model:** $\widehat{y} = 4.2 - 0.35x$



| | $x_i$ | $y_i$ | $\widehat{y_i}$ | $e_i = y_i - \widehat{y_i}$ | $e_i^2$ |
|---|---|---|---|---|---|
| A | 2.0 | 1.10 | 3.500 | -2.400 | 5.760 |
| B | 3.5 | 1.925 | 2.975 | -1.050 | 1.103 |
| C | 5.0 | 2.25 | 2.450 | -0.200 | 0.040 |
| D | 6.0 | 2.90 | 2.100 | 0.800 | 0.640 |
| E | 7.5 | 3.175 | 1.575 | 1.600 | 2.560 |
| F | 9.0 | 3.90 | 1.050 | 2.850 | 8.123 |
| G | 11.0 | 4.35 | 0.350 | 4.000 | 16.000 |

$\mathrm{SSE}_{\mathrm{bad}} = \sum e_i^2 = 34.225$

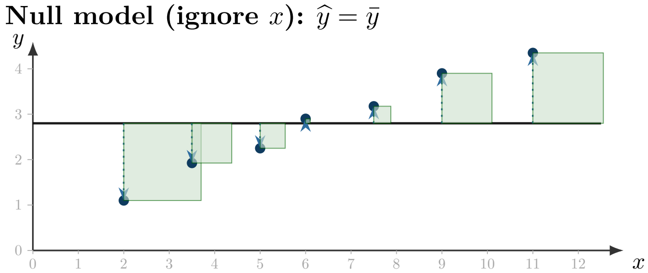
**Best (least-squares) model:** $\widehat{y} = 0.5440 + 0.3589x$



| | $x_i$ | $y_i$ | $\widehat{y_i}$ | $e_i = y_i - \widehat{y_i}$ | $e_i^2$ |
|---|---|---|---|---|---|
| A | 2.0 | 1.10 | 1.2618 | -0.1618 | 0.026179 |
| B | 3.5 | 1.925 | 1.8001 | 0.12485 | 0.015588 |
| C | 5.0 | 2.25 | 2.3385 | -0.0885 | 0.007832 |
| D | 6.0 | 2.90 | 2.6974 | 0.2026 | 0.041047 |
| E | 7.5 | 3.175 | 3.2357 | -0.06075 | 0.003691 |
| F | 9.0 | 3.90 | 3.7741 | 0.1259 | 0.015851 |
| G | 11.0 | 4.35 | 4.4919 | -0.1419 | 0.020136 |

$\mathrm{SSE}_{\mathrm{best}} = \sum e_i^2 = 0.1303$

**Null model (ignore $x$):** $\widehat{y} = \bar{y}$



| | $x_i$ | $y_i$ | $\widehat{y_i}$ | $e_i = y_i - \widehat{y_i}$ | $e_i^2$ |
|---|---|---|---|---|---|
| A | 2.0 | 1.10 | 2.800 | -1.700 | 2.890 |
| B | 3.5 | 1.925 | 2.800 | -0.875 | 0.766 |
| C | 5.0 | 2.25 | 2.800 | -0.550 | 0.303 |
| D | 6.0 | 2.90 | 2.800 | 0.100 | 0.010 |
| E | 7.5 | 3.175 | 2.800 | 0.375 | 0.141 |
| F | 9.0 | 3.90 | 2.800 | 1.100 | 1.210 |
| G | 11.0 | 4.35 | 2.800 | 1.550 | 2.403 |

$\mathrm{SSE}_{\mathrm{null}} = \sum e_i^2 = 7.7213$

### Quick questions

1. Which model has the *smallest* total square area?

   The least-squares model (middle row).

2. The bottom row ("ignore $x$") gives a baseline amount of square area. How can we tell if another model is an *improvement* compared to this baseline?

   Compare its total residual square area to the baseline's; smaller than baseline means improvement, larger means worse.

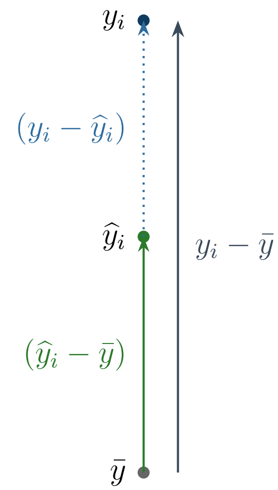
3. If a model's square area is only a little smaller than the baseline, what does that suggest about $x$? What if the model's square area is much smaller?

   Only a little smaller $\Rightarrow$ $x$ explains little of the variation in $y$. Much smaller $\Rightarrow$ $x$ explains a large share of the variation.

### 3) Decomposing squares and $R^2$

Any response $y_i$ can be decomposed into

$$y_i - \bar{y} \;=\; (y_i - \widehat{y}_i) \;+\; (\widehat{y}_i - \bar{y}).$$




Squaring and summing over points leads to the **sum-of-squares identity**

$$\underbrace{\text{SST}}_{\sum(y_i-\bar{y})^2} \;=\; \underbrace{\text{SSR}}_{\sum(\widehat{y}_i-\bar{y})^2} \;+\; \underbrace{\text{SSE}}_{\sum(y_i-\widehat{y}_i)^2}.$$

The coefficient of determination is

$$R^2 \;=\; \frac{\text{SSR}}{\text{SST}} \;=\; 1 - \frac{\text{SSE}}{\text{SST}},$$

the *proportion of total square area explained* by using $x$.

**Shade/identify the squares:**

- On the *null model* panel, your squares show SST.
- On the *best model* panel, your squares show SSE.
- The *explained* squares correspond to SSR = SST − SSE.

| | SST | SSE (best) | SSR = SST − SSE | $R^2 = \frac{\text{SSR}}{\text{SST}}$ |
|---|---|---|---|---|
| **Values** | 7.7213 | 0.1303 | 7.5909 | 0.9831 |

### Practice

1. Explain why the explained squares (SSR) must be *nonnegative*.

   They are sums of squares $(\widehat{y}_i - \bar{y})^2$, and squares are never negative; geometrically, area cannot be negative.

2. If a different line (not least squares) is used, which quantity necessarily increases, SSE or SST? Why?

   SSE increases (or stays the same) because the least-squares line minimizes the sum of squared residuals. SST depends only on $y$ and $\bar{y}$ and is unaffected by the choice of line.

3. In this dataset, $R^2$ is very close to 1. What does that tell you about the usefulness of $x$ for predicting $y$?

   Nearly all of the total variation in $y$ is explained by the linear relationship with $x$; $x$ is highly predictive here.

4. What is the lowest possible value of $R^2$ and what does it mean in context? What is the largest value of $R^2$ and what does it mean in context?

   $R^2_{\min} = 0$: using $x$ gives no improvement over predicting everyone with $\bar{y}$ (no explained variation). $R^2_{\max} = 1$: a perfect linear fit—every residual is 0, so the model explains *all* the variation.

5. $\star$ Prove SST = SSR + SSE.

   **Goal.** Show $\sum(y_i - \bar{y})^2 = \sum(\widehat{y}_i - \bar{y})^2 + \sum(y_i - \widehat{y}_i)^2$.
   **Step 1: Pointwise decomposition.** For each $i$,

$$y_i - \bar{y} \;=\; (y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y}).$$

   Squaring gives

$$(y_i - \bar{y})^2 = (y_i - \widehat{y}_i)^2 + (\widehat{y}_i - \bar{y})^2 + 2(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}).$$

   Summing over $i$,

$$\underbrace{\sum(y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum(y_i - \widehat{y}_i)^2}_{\text{SSE}} + \underbrace{\sum(\widehat{y}_i - \bar{y})^2}_{\text{SSR}} + 2\sum(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}).$$

   Thus it suffices to show the cross term is 0.

   **Step 2: Normal equations $\Rightarrow$ orthogonality.** Write residuals $e_i = y_i - \widehat{y}_i$. For least squares with an intercept, the normal equations give

$$\sum_{i=1}^{n} e_i = 0 \quad \text{and} \quad \sum_{i=1}^{n} e_i x_i = 0.$$

   Because $\widehat{y}_i = \hat{a} + \hat{b}\, x_i$, we have

$$\sum e_i\, \widehat{y}_i = \hat{a} \sum e_i + \hat{b} \sum e_i x_i = 0 + 0 = 0.$$

   Now expand the cross term:

$$\sum e_i(\widehat{y}_i - \bar{y}) = \underbrace{\sum e_i \widehat{y}_i}_{=0} - \bar{y} \underbrace{\sum e_i}_{=0} = 0.$$

   Hence $2\sum(y_i - \widehat{y}_i)(\widehat{y}_i - \bar{y}) = 0$.

   **Conclusion.** The cross term vanishes, so

$$\text{SST} = \text{SSR} + \text{SSE}.$$

   **Geometric intuition (optional).** Let $\mathbf{y}$ be the data vector, $\mathbf{1}$ the all-ones vector, and $X = [\mathbf{1}, \mathbf{x}]$. Then $\widehat{\mathbf{y}}$ is the orthogonal projection of $\mathbf{y}$ onto the column space of $X$. Decompose around the mean: $\mathbf{y} - \bar{y}\,\mathbf{1} = (\widehat{\mathbf{y}} - \bar{y}\,\mathbf{1}) + (\mathbf{y} - \widehat{\mathbf{y}})$, where the two addends are orthogonal. By the Pythagorean theorem,

$$\|\mathbf{y} - \bar{y}\,\mathbf{1}\|^2 = \|\widehat{\mathbf{y}} - \bar{y}\,\mathbf{1}\|^2 + \|\mathbf{y} - \widehat{\mathbf{y}}\|^2,$$

   which is exactly SST = SSR + SSE. *Note:* The intercept is essential—without it, the identity holds with $\bar{y}$ replaced by 0 (about the origin).