# Simple linear regression: the least-squares line, residuals and R²

Ten problems on fitting a straight line by least squares: the slope and intercept from sums of squares, residuals that sum to zero, the split of the total sum of squares and R² as the squared correlation, the slope from a correlation and two standard deviations, the standard error of the slope with a t test and confidence interval, what changes with the units, one point of high leverage, the normal equations from calculus, intervals for a mean response and for a new observation, and a negative R² on test data, with worked solutions and the mistakes of inverting the slope to predict x, dividing by n instead of n − 2, reading r as the fraction of variance explained, extrapolating far past the data and trusting training R² as a test score.

**BEFORE YOU START**

Simple linear regression fits the line $\hat{y} = b_0 + b_1 x$ that makes the squared vertical distances from the data to the line as small as possible. It is the one-feature case of the least-squares problems elsewhere on this site, but in statistics it comes with its own vocabulary: the slope as a ratio of sums of squares, the residuals, $R^2$ as the share of the variation in $y$ that the line explains, and a standard error that says how far the slope could move with new data. These ten problems compute all of them by hand on small data sets, connect the slope to the correlation, and end with the number every machine-learning library reports as a score, $R^2$ on held-out data, which can be negative.

- **Data and sums of squares.** Pairs $(x_i, y_i)$ for $i = 1, \dots, n$, with means $\bar{x}$ and $\bar{y}$. $S_{xx} = \sum (x_i - \bar{x})^2$, $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$ and $S_{yy} = \sum (y_i - \bar{y})^2$.
- **The least-squares line.** $b_1 = S_{xy} / S_{xx}$ and $b_0 = \bar{y} - b_1 \bar{x}$. The line passes through $(\bar{x}, \bar{y})$. With the correlation $r$ and the standard deviations $s_x$, $s_y$, the slope is also $b_1 = r\, s_y / s_x$.
- **Residuals.** $e_i = y_i - \hat{y}_i$, with $\hat{y}_i = b_0 + b_1 x_i$. For a line with an intercept, $\sum e_i = 0$ and $\sum x_i e_i = 0$.
- **Sums of squares.** $\mathrm{SST} = S_{yy}$, $\mathrm{SSE} = \sum e_i^2$ and $\mathrm{SSR} = \sum (\hat{y}_i - \bar{y})^2 = b_1^2 S_{xx}$, with $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$.
- $R^2$. $R^2 = \mathrm{SSR}/\mathrm{SST} = 1 - \mathrm{SSE}/\mathrm{SST}$, and for simple regression $R^2 = r^2$. On data the line was not fitted to, $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$ can be negative.
- **Inference.** $s = \sqrt{\mathrm{SSE}/(n-2)}$ estimates the noise standard deviation; two degrees of freedom go to $b_0$ and $b_1$. $\mathrm{SE}(b_1) = s/\sqrt{S_{xx}}$, and $t = b_1 / \mathrm{SE}(b_1)$ has a $t$ distribution with $n - 2$ degrees of freedom when the true slope is 0. Two-sided 5% cutoffs: $t_3 = 3.182$, $t_4 = 2.776$, $t_8 = 2.306$.

- **Intervals at a new $x_0$.** For the mean response, SE $= s\sqrt{1/n + (x_0 - \bar{x})^2/S_{xx}}$. For a single new observation, add 1 under the root: $s\sqrt{1 + 1/n + (x_0 - \bar{x})^2/S_{xx}}$.

Builds on: Variance, covariance and correlation, Hypothesis tests and p-values: z and t tests, power and multiple comparisons

## Problems

**1.** Five students report hours studied and test scores: difficulty ·

| $x$ (hours) | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $y$ (score) | 52 | 55 | 61 | 64 | 68 |

Find the least-squares line.

**2.** For the line of Problem 1, find the five residuals. Check that $\sum e_i = 0$ and $\sum x_i e_i = 0$, and predict the score for 6 hours. difficulty ·

**3.** For the data of Problem 1, find SST, SSR and SSE, then $R^2$. Check that $R^2 = r^2$. difficulty ··

**4.** Adult heights have mean 170 cm and sd 10 cm, weights mean 70 kg and sd 12 kg, with correlation $r = 0.5$. Find the line predicting weight from height and the prediction at 180 cm. Then find the line predicting height from weight, and the predicted height for 82 kg. difficulty ··

**5.** For the data of Problem 1, find $s$, the standard error of $b_1$, the $t$ statistic for $H_0 : \beta_1 = 0$ and a 95% confidence interval for the slope. difficulty ··

**6.** In Problem 1, the hours are re-recorded in minutes. What are the new slope, intercept and $R^2$? What are the slope and intercept if instead the hours are centered, $x' = x - 3$? difficulty ··

**7.** Add a sixth student to the data of Problem 1, with 10 hours and a score of 60. Find the new line and $R^2$, and explain what the one point did. difficulty ···

**8.** Minimize $\mathrm{SSE}(b_0, b_1) = \sum (y_i - b_0 - b_1 x_i)^2$ with calculus. Derive the two normal equations, and show they give $b_1 = S_{xy}/S_{xx}$ and $b_0 = \bar{y} - b_1 \bar{x}$. difficulty ···

**9.** For the data of Problem 1 at $x_0 = 6$ hours, find a 95% confidence interval for the mean score and a 95% prediction interval for one new student's score. difficulty ···

**10.** The line of Problem 1 is tested on two new students, with $(6, 66)$ and $(7, 70)$. Find $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$ on these two points, where SST uses their own mean. difficulty ···

## Answers

1. $\hat{y} = 47.7 + 4.1x$
2. Residuals $0.2, -0.9, 1, -0.1, -0.2$, summing to $0$ with $\sum x_i e_i = 0$; at 6 hours, $\hat{y} = 72.3$
3. SST $= 170$, SSR $= 168.1$, SSE $= 1.9$, $R^2 \approx 0.989$; $r = 41/\sqrt{10 \cdot 170} \approx 0.9944$ and $r^2 \approx 0.989$
4. Weight $= -32 + 0.6 \cdot$ height, 76 kg at 180 cm; height $= 170 + 0.417(\text{weight} - 70)$, 175 cm at 82 kg
5. $s \approx 0.796$, $\text{SE}(b_1) \approx 0.252$, $t \approx 16.3$; 95% interval $4.1 \pm 3.182 \cdot 0.2517 \approx (3.30, 4.90)$
6. Minutes: $b_1 \approx 0.0683$, $b_0 = 47.7$, $R^2 \approx 0.989$; centered: $b_1 = 4.1$, $b_0 = 60$
7. $\hat{y} \approx 56.64 + 0.807x$, $R^2 \approx 0.195$, down from $0.989$
8. $b_1 = \dfrac{\sum x_i(y_i - \bar{y})}{\sum x_i(x_i - \bar{x})} = \dfrac{S_{xy}}{S_{xx}}$, $b_0 = \bar{y} - b_1\bar{x}$
9. Mean score: $72.3 \pm 2.66 \approx (69.6, 75.0)$; one new student: $72.3 \pm 3.67 \approx (68.6, 76.0)$
10. $R^2 \approx -9.08$

# Worked solutions

## Problem 1

> Five students report hours studied and test scores:
>
> | $x$ (hours) | 1 | 2 | 3 | 4 | 5 |
> |---|---|---|---|---|---|
> | $y$ (score) | 52 | 55 | 61 | 64 | 68 |
>
> Find the least-squares line.

| | |
|---|---|
| $\bar{x} = 3, \bar{y} = 300/5 = 60$ | The two means. |
| $x - \bar{x} = -2, -1, 0, 1, 2$; $y - \bar{y} = -8, -5, 1, 4, 8$ | Center both variables. |
| $S_{xy} = 16 + 5 + 0 + 4 + 16 = 41$; $S_{xx} = 10$ | Products of the centered values, and squares of the centered $x$. |
| $b_1 = 41/10 = 4.1$; $b_0 = 60 - 4.1 \cdot 3 = 47.7$ | Slope, then the intercept from the point $(\bar{x}, \bar{y})$. |
| $\hat{y} = 47.7 + 4.1x$ | Each extra hour goes with 4.1 more points, on average, in these data. |

## Problem 2

> For the line of Problem 1, find the five residuals. Check that $\sum e_i = 0$ and $\sum x_i e_i = 0$, and predict the score for 6 hours.

| | |
|---|---|
| $\hat{y} = 51.8, 55.9, 60, 64.1, 68.2$ | $47.7 + 4.1x$ at $x = 1, \dots, 5$. |
| $e = 0.2, -0.9, 1, -0.1, -0.2$ | Observed minus fitted. |
| $\sum e_i = 0.2 - 0.9 + 1 - 0.1 - 0.2 = 0$ | The intercept makes the residuals balance. |
| $\sum x_i e_i = 0.2 - 1.8 + 3 - 0.4 - 1 = 0$ | The slope makes the residuals uncorrelated with $x$. |

| | |
|---|---|
| **Residuals** $0.2, -0.9, 1, -0.1, -0.2$, **summing to** $0$ **with** $\sum x_i e_i = 0$; **at** $6$ **hours,** $\hat{y} = 72.3$ | The two zero sums are the normal equations (Problem 8), so they hold for every least-squares line with an intercept. |

## Problem 3

For the data of Problem 1, find SST, SSR and SSE, then $R^2$. Check that $R^2 = r^2$.

| | |
|---|---|
| $\text{SST} = 64 + 25 + 1 + 16 + 64 = 170$ | The squared centered scores from Problem 1. |
| $\text{SSE} = 0.04 + 0.81 + 1 + 0.01 + 0.04 = 1.9$ | The squared residuals from Problem 2. |
| $\text{SSR} = b_1^2 S_{xx} = 16.81 \cdot 10 = 168.1$ | And $168.1 + 1.9 = 170$. |
| $R^2 = 168.1/170 \approx 0.9888$ | The share of the variation in $y$ explained by the line. |
| $\text{SST} = 170$, $\text{SSR} = 168.1$, $\text{SSE} = 1.9$, $R^2 \approx 0.989$; $r = 41/\sqrt{10 \cdot 170} \approx 0.9944$ **and** $r^2 \approx 0.989$ | $r^2 = S_{xy}^2/(S_{xx}S_{yy}) = b_1^2 S_{xx}/S_{yy} =$ SSR/SST. |

## Problem 4

Adult heights have mean 170 cm and sd 10 cm, weights mean 70 kg and sd 12 kg, with correlation $r = 0.5$. Find the line predicting weight from height and the prediction at 180 cm. Then find the line predicting height from weight, and the predicted height for 82 kg.

| | |
|---|---|
| $b_1 = r\, s_y/s_x = 0.5 \cdot 12/10 = 0.6$ kg per cm | The slope from the correlation. |
| $b_0 = 70 - 0.6 \cdot 170 = -32$ | The line passes through the means. |
| At 180 cm: $-32 + 0.6 \cdot 180 = 76$ kg | One sd above the mean in height predicts only 0.5 sd above in weight: regression toward the mean. |

| | |
|---|---|
| Height on weight: slope $r\,s_x/s_y = 0.5 \cdot 10/12 \approx 0.417$ cm per kg | Swap the roles of $x$ and $y$; the correlation is the same. |
| **Weight** $= -32 + 0.6 \cdot$ height, 76 **kg at** 180 **cm; height** $= 170 + 0.417(\text{weight} - 70)$, 175 **cm at** 82 **kg** | 82 kg is one sd above the mean, so the prediction is half an sd, 5 cm, above. Inverting the first line gives a slope of $1/0.6 \approx 1.67$ instead (Mistake 1). |

## Problem 5

For the data of Problem 1, find $s$, the standard error of $b_1$, the $t$ statistic for $H_0\colon \beta_1 = 0$ and a 95% confidence interval for the slope.

| | |
|---|---|
| $s^2 = \text{SSE}/(n-2) = 1.9/3 \approx 0.6333$ | Two degrees of freedom are used by $b_0$ and $b_1$. |
| $s \approx 0.7958$ | The typical size of a residual. |
| $\text{SE}(b_1) = 0.7958/\sqrt{10} \approx 0.2517$ | $s/\sqrt{S_{xx}}$. |
| $t = 4.1/0.2517 \approx 16.29$, with 3 degrees of freedom | Far past the cutoff 3.182; two-sided $p \approx 0.0005$. |
| $s \approx 0.796$, $\text{SE}(b_1) \approx 0.252$, $t \approx 16.3$; 95% **interval** $4.1 \pm 3.182 \cdot 0.2517 \approx (3.30, 4.90)$ | With only 5 points the $t$ cutoff is 3.182, not 1.96. Dividing by $n$ instead of $n-2$ gives an interval that is too narrow (Mistake 2). |

## Problem 6

In Problem 1, the hours are re-recorded in minutes. What are the new slope, intercept and $R^2$? What are the slope and intercept if instead the hours are centered, $x' = x - 3$?

| | |
|---|---|
| $x_{\text{min}} = 60x$, so $S_{xy}$ and $S_{xx}$ scale by 60 and 3600 | Every centered $x$ is multiplied by 60. |

| | |
|---|---|
| $b_1 = 4.1/60 \approx 0.0683$ points per minute | The slope is per unit of $x$, so it shrinks by the factor 60. |
| $b_0 = 60 - 0.0683 \cdot 180 = 47.7$; $R^2$ unchanged | The fitted values are the same, so the residuals and $R^2$ are too. |
| Centered: slope 4.1, intercept $\bar{y} = 60$ | With $\bar{x}' = 0$, $b_0 = \bar{y} - b_1 \cdot 0$. |
| **Minutes:** $b_1 \approx 0.0683$, $b_0 = 47.7$, $R^2 \approx 0.989$; **centered:** $b_1 = 4.1$, $b_0 = 60$ | A slope's size depends on the units, so it is not by itself a measure of how strong the relationship is. $R^2$ and $r$ do not depend on units. |

### Problem 7

> Add a sixth student to the data of Problem 1, with 10 hours and a score of 60. Find the new line and $R^2$, and explain what the one point did.

| | |
|---|---|
| $\bar{x} = 25/6 \approx 4.167$, $\bar{y} = 360/6 = 60$ | The mean score does not move, because the new score is 60. |
| $S_{xy} = \sum x_i(y_i - 60) = 41$ | The new point has $y - \bar{y} = 0$, so it adds nothing; and $\sum(y_i - \bar{y}) = 0$ lets $\bar{x}$ drop out. |
| $S_{xx} = 155 - 6\bar{x}^2 \approx 50.83$ | $\sum x_i^2 = 55 + 100 = 155$; the new point is far from the others in $x$. |
| $b_1 = 41/50.83 \approx 0.807$; $b_0 = 60 - 0.807 \cdot 4.167 \approx 56.64$ | Slope, then intercept. |
| $\hat{y} \approx 56.64 + 0.807x$, $R^2 \approx 0.195$, **down from** 0.989 | SSR $= 0.807^2 \cdot 50.83 \approx 33.07$ of SST $= 170$. One point with an extreme $x$, a point of high leverage, cut the slope by a factor of 5. Plot the data before trusting a line. |

## Problem 8

Minimize $\text{SSE}(b_0, b_1) = \sum (y_i - b_0 - b_1 x_i)^2$ with calculus. Derive the two normal equations, and show they give $b_1 = S_{xy}/S_{xx}$ and $b_0 = \bar{y} - b_1 \bar{x}$.

| | |
|---|---|
| $\partial\,\text{SSE}/\partial b_0 = -2\sum (y_i - b_0 - b_1 x_i) = 0$ | That is, $\sum e_i = 0$. |
| $\partial\,\text{SSE}/\partial b_1 = -2\sum x_i (y_i - b_0 - b_1 x_i) = 0$ | That is, $\sum x_i e_i = 0$. |
| First equation over $n$: $\bar{y} - b_0 - b_1 \bar{x} = 0$ | So $b_0 = \bar{y} - b_1 \bar{x}$, and the line goes through $(\bar{x}, \bar{y})$. |
| Substitute: $\sum x_i \big((y_i - \bar{y}) - b_1 (x_i - \bar{x})\big) = 0$ | $y_i - b_0 - b_1 x_i = (y_i - \bar{y}) - b_1 (x_i - \bar{x})$. |
| $b_1 = \dfrac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \dfrac{S_{xy}}{S_{xx}}$, $b_0 = \bar{y} - b_1 \bar{x}$ | Replacing $x_i$ by $x_i - \bar{x}$ in front changes neither sum, because $\sum (y_i - \bar{y}) = 0$ and $\sum (x_i - \bar{x}) = 0$. The Hessian has determinant $4nS_{xx} > 0$, so this is a minimum. |

## Problem 9

For the data of Problem 1 at $x_0 = 6$ hours, find a 95% confidence interval for the mean score and a 95% prediction interval for one new student's score.

| | |
|---|---|
| $\hat{y}_0 = 72.3$ and $(x_0 - \bar{x})^2 / S_{xx} = 9/10 = 0.9$ | Problem 2, and the distance of $x_0$ from the center of the data. |
| Mean: $\text{SE} = 0.7958\sqrt{0.2 + 0.9} \approx 0.8347$ | $1/n = 0.2$. |
| New student: $\text{SE} = 0.7958\sqrt{1 + 0.2 + 0.9} \approx 1.1533$ | One more $s^2$ for the student's own scatter around the line. |
| Margins $3.182 \cdot 0.8347 \approx 2.66$ and $3.182 \cdot 1.1533 \approx 3.67$ | The $t$ cutoff with 3 degrees of freedom. |

| | |
|---|---|
| **Mean score:** $72.3 \pm 2.66 \approx (69.6, 75.0)$; **one new student:** $72.3 \pm 3.67 \approx (68.6, 76.0)$ | The prediction interval is always wider. More data shrinks the first toward 0, but the second never shrinks below about $\pm 1.96\sigma$. |

### Problem 10

> The line of Problem 1 is tested on two new students, with $(6, 66)$ and $(7, 70)$. Find $R^2 = 1 - \text{SSE}/\text{SST}$ on these two points, where SST uses their own mean.

| | |
|---|---|
| Predictions 72.3 and 76.4 | $47.7 + 4.1 \cdot 6$ and $47.7 + 4.1 \cdot 7$. |
| Errors $-6.3$ and $-6.4$, so $\text{SSE} = 39.69 + 40.96 = 80.65$ | Both students score well below the line. |
| Test mean 68, so $\text{SST} = 4 + 4 = 8$ | The spread of the test scores around their own mean. |
| $R^2 = 1 - 80.65/8$ | SSE is ten times SST. |
| $R^2 \approx -9.08$ | A negative $R^2$ means the line predicts worse than the constant 68, the test set's own mean. On training data with an intercept $R^2 \geq 0$ is guaranteed; on new data it is not (Mistake 5). |

## Where this goes wrong

### 1. Inverting the slope to predict x from y

Problem 4 has a line for weight from height, and solving it for height looks like the line for height from weight.

| | |
|---|---|
| $\text{Weight} = -32 + 0.6 \cdot \text{height}$ | Right so far. |

| | |
|---|---|
| “Solve for height: height $=$ (weight $+$ 32)$/0.6$.” | That line minimizes the vertical distances to weight, not the distances in height. Least squares in the other direction is a different problem, with slope $r\, s_x/s_y$. |
| At 82 kg: height $= 114/0.6 = 190$ cm ✗ | The right prediction is 175 cm. The inverted slope $1/0.6 \approx 1.67$ cm per kg is four times 0.417, since the two slopes multiply to $r^2 = 0.25$, not to 1. |

### 2. Dividing SSE by n instead of n − 2

Problem 5 needs an estimate of the noise variance, and an average of squared residuals looks like dividing by $n$.

| | |
|---|---|
| SSE $= 1.9$, $n = 5$ | Right so far. |
| “$s^2 = 1.9/5 = 0.38$.” | The residuals are forced to satisfy two equations, $\sum e_i = 0$ and $\sum x_i e_i = 0$, so only $n - 2$ of them are free, and SSE$/n$ underestimates $\sigma^2$ on average. |
| SE$(b_1) \approx 0.195$, interval $\approx (3.48, 4.72)$ ✗ | With $n - 2 = 3$: SE$(b_1) \approx 0.252$ and interval $(3.30, 4.90)$. The wrong interval is about 23% too narrow, and it covers the true slope only about 91% of the time instead of 95%. |

### 3. Reading r as the fraction of variance explained

The correlation of Problem 4 is 0.5, and a half sounds like a share.

| | |
|---|---|
| $r = 0.5$ between height and weight | Right so far. |
| “Height explains 50% of the variation in weight.” | The fraction explained is $R^2 = r^2$, not $r$. |

| | |
|---|---|
| 50% of the variance in weight is explained ✗ | It is $r^2 = 25\%$. A correlation of 0.5 leaves three quarters of the variance unexplained, and even $r = 0.7$ explains less than half. |

### 4. Extrapolating far past the data

The line of Problem 1 has a clean $R^2 \approx 0.989$, so it seems safe to use anywhere.

| | |
|---|---|
| $\hat{y} = 47.7 + 4.1x$, fitted for $1 \le x \le 5$ | Right so far. |
| "A student who studies 40 hours scores $47.7 + 4.1 \cdot 40$." | The data say nothing about the relationship outside 1 to 5 hours, and the prediction interval of Problem 9 already widens with $(x_0 - \bar{x})^2$. |
| $\hat{y} = 211.7$ at 40 hours ✗ | On a 100-point test. A high $R^2$ says the line fits the range it was fitted on; it says nothing about where the line stops being a line. |

### 5. Trusting training R² as a test score

The line of Problem 1 has $R^2 \approx 0.989$ on its own data, and that number is reported as how good the model is.

| | |
|---|---|
| $R^2 \approx 0.989$ on the five training points | Right so far. |
| "So the model will explain about 99% of the variation in new scores." | Training $R^2$ measures the fit to the points that chose the line. With few points and a flexible model it is optimistic, and it says nothing about new data from a different group. |
| Expected test $R^2 \approx 0.99$ ✗ | On the two test students of Problem 10, $R^2 \approx -9.08$. Report $R^2$ on held-out data, and remember it is not bounded below by 0. |

**More at eulerpad.app.** This is one of 82 free practice sets with worked solutions at eulerpad.app/practice, alongside guides to the topics behind them.

**EulerPad**, the app, reads your handwritten math on iPad, iPhone and Android and checks it line by line. It's free to start.

This set is also Chapter 22 of *Probability & Statistics for Machine Learning*, a 250-problem workbook with worked solutions: eulerpad.app/workbooks

eulerpad.app/practice/simple-linear-regression-and-r-squared