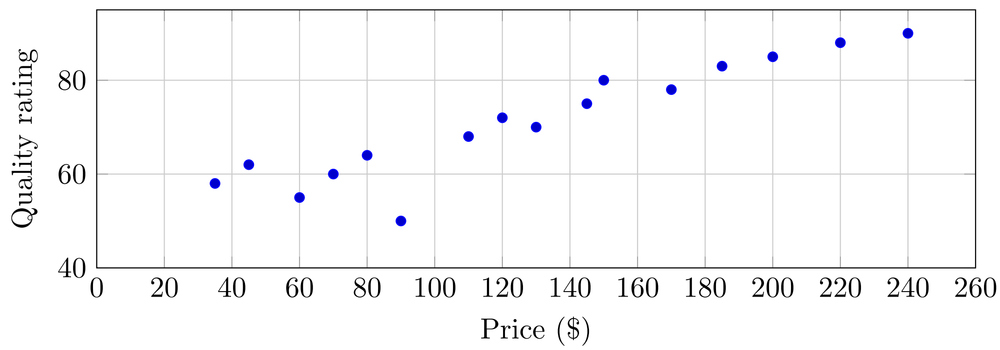
Unit 2: Extra Practice

Mr. Merrick · October 21, 2025

Problem 1. Describing a relationship

A study recorded the price (in dollars) and expert quality rating (0–100) for $n = 16$ Bluetooth speakers. A scatterplot (below) shows the data.



(a) Describe the direction, form, and strength of the association, and identify any notable features.

Solution. The association between price and quality rating is *positive*, *roughly linear*, and *moderately strong*. The scatter appears smaller at higher prices. A possible low-rating point is near ($90, 50).

(b) Based solely on the scatter plot, would a least-squares regression of rating on price be reasonable? Justify using the graph.

Solution. Yes. The trend is approximately linear with roughly constant spread and no strong curvature or influential outliers.

(c) In context, explain what a point near $(\$90, 50)$ suggests to a shopper.

Solution. A $90 speaker with a rating near 50 appears to underperform for its price compared with similarly priced models.

Problem 2. Interpreting slope and intercept

The least-squares line for predicting rating y from price x (in dollars) for a different set of speakers is

$$\widehat{y} = 41.3 + 0.21x, \qquad s = 5.6, \quad r^2 = 68\%.$$

These data came from speakers priced between about \$30 and \$250.

(a) Interpret the slope and the intercept in context.

Solution. Slope 0.21: each additional \$1 is associated with an average 0.21-point increase in the predicted rating (about 2.1 points per \$10). Intercept 41.3 is the prediction at \$0; it is not meaningful in context but anchors the line.

(b) Estimate the rating for a \$150 speaker and interpret $s = 5.6$.

Solution. $\hat{y} = 41.3 + 0.21(150) = 72.8$. The typical prediction error (typical/average size of a residual) is about 5.6 *rating points*.

(c) Is a \$500 prediction advisable? Explain.

Solution. No—\$500 is far beyond the data range, so extrapolation would be unreliable.

Problem 3. From output to equation

A computer regresses weekly study hours y on weekly work hours x for $n = 12$ students (working between 5 and 25 hours per week) and reports:

Predictor	Coef	SE Coef	t	p
Constant	18.2	2.9	6.28	< 0.001
Work	-0.41	0.15	-2.73	0.021
$S = 3.7$	$R^2 = 42\%$		$R^2_{\text{adj}} = 36\%$	

(a) Write the least-squares equation and interpret the slope.

Solution. $\widehat{y} = 18.2 - 0.41x$. Each additional hour of work per week is associated with about 0.41 fewer study hours, on average.

(b) If a student works 20 hours, what is the predicted study time? Comment on practical reasonableness.

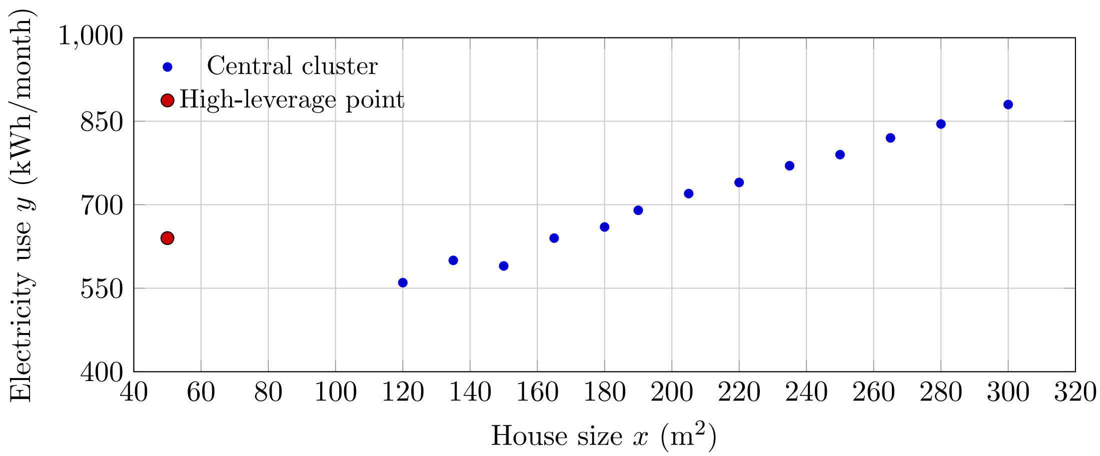
Solution. $\widehat{y} = 18.2 - 0.41(20) = 10.0$ hours; this is plausible and within the data range (interpolation).

(c) Compute and interpret the residual for a student who worked $x = 10$ hours and studied $y = 12$ hours. Then sketch or describe the residual plot pattern you would expect.

Solution. Predicted $\widehat{y} = 18.2 - 0.41(10) = 14.1$. Residual $e = 12 - 14.1 = -2.1$ hours (studied about two hours less than predicted). A suitable residual plot would show points scattered around 0 with no curvature and roughly constant spread.

Problem 4. Influential point vs. outlier

The scatterplot shows monthly electricity use (y, in kWh) versus house size (x, in m^2). Most houses fall between 100–300 m^2.



(a) Explain why the leftmost point is likely *influential* for the regression line.

Solution. Its x-value is far from the mean ($\Rightarrow$ high leverage). High-leverage points can strongly affect the slope and intercept because the least-squares line balances horizontal spread as well as vertical distances.

(b) If that point were removed, what would you expect to happen to the slope and to R^2? Explain your reasoning using the figure.

Solution. In *this configuration*, removing the leftmost point would typically make the slope steeper (less pull toward the left) and would likely increase R^2 because the remaining points align more closely with a straight line.

Problem 5. AP-style free response

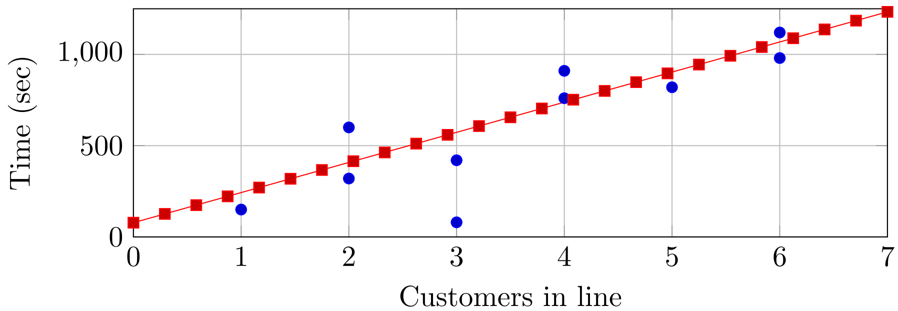
A manager samples $n = 10$ checkout lines. Let x be the number of customers ahead of a shopper and y the total checkout time (sec). The regression output is:

Predictor	Coef	SE Coef	t	p
Constant	78	96	0.81	0.44
Customers in line	165	28	5.89	< 0.001
$S = 190$	$R^2 = 78\%$		$R^2_{\text{adj}} = 75\%$	

(a) Write the least-squares equation. Interpret the slope in context.

Solution. $\widehat{y} = 78 + 165x$. Each additional customer ahead adds about 165 seconds to the predicted checkout time.

(b) Circle on the sketch the most likely outlier and explain why:



Solution. The point at $(3, 80)$ is a vertical outlier—much lower than predicted—and increases the scatter.

(c) Interpret $R^2 = 78\%$.

Solution. About 78% of the variation in checkout times is explained by the linear relationship with the number of customers ahead.

Problem 6. Using residual to recover an observed value

For wolves, the fitted line for weight (kg) on length (m) is $\widehat{y} = -16.46 + 35.02x$. A wolf of length 1.40 m has residual -9.67 kg.

(a) What is this wolf's actual weight?

Solution. $\widehat{y} = -16.46 + 35.02(1.40) = 32.56$ kg. Actual $y = \widehat{y} + e = 32.56 - 9.67 = 22.89$ kg.

(b) Interpret the residual.

Solution. The wolf weighed about 9.7 kg *less* than predicted for its length.

Problem 7. Multiple parts, mixed skills

Biologists measured mass y (g) and length x (mm) for 11 frogs and obtained the regression line

$$\widehat{y} = -546 + 6.086x, \qquad r^2 \approx 0.819.$$

(a) Interpret the slope in context.

Solution. For each additional millimeter of length, predicted mass increases by about 6.086 grams on average.

(b) Interpret r^2 in context.

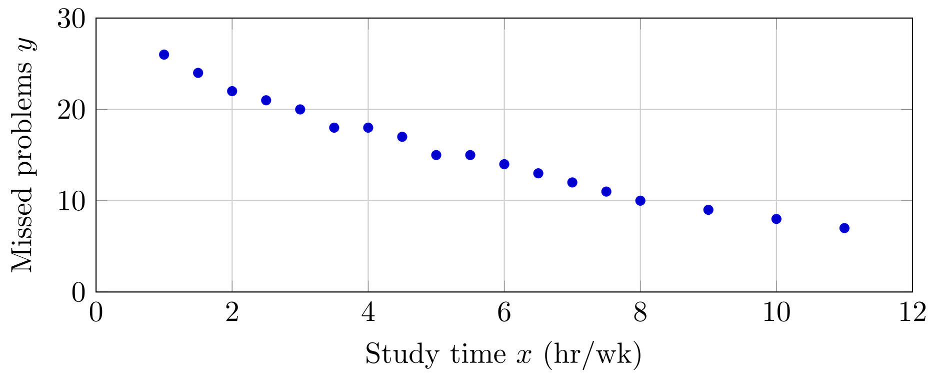
Solution. About 81.9% of the variability in frog mass is explained by the linear relationship with length.

(c) On a residual plot, which frog would have the larger magnitude residual: one with $(x = 130, y = 220)$ or one with $(x = 170, y = 530)$? Show work.

Solution. At $x = 130$: $\widehat{y} = -546 + 6.086(130) = 246.2$, so $e = 220 - 246.2 = -26.2$. At $x = 170$: $\widehat{y} = 492.6$, so $e = 530 - 492.6 = 37.4$. The second has the larger $|e|$.

Problem 8. Correlation: sign, magnitude, and meaning

The plot shows a relationship between study time (x, hours/week) and number of missed homework problems (y) for $n = 18$ students.



(a) Based on the plot, state the *direction*, *form*, and *strength* of the association and give a rough estimate of the *sign* of r.

Solution. Direction: negative; form: roughly linear; strength: strong (little scatter). Therefore r is negative and close to -1 (plausibly between -0.9 and -0.98).

(b) A computer output (not shown) reports $R^2 = 0.92$ and a *negative* slope. Compute r and interpret R^2 in context.

Solution. $r = \text{sign}(b_1)\sqrt{R^2} = -\sqrt{0.92} \approx -0.959$. $R^2 = 0.92$ means about 92% of the variation in missed problems among these students is explained by the linear relationship with study time.

Problem 9. Changing units: what changes, what doesn't

For $n = 25$ headphones, the least-squares line for predicting quality rating y (0–100 points) from price x (US dollars) is

$$\hat{y} = 12.0 + 0.45x, \qquad R^2 = 64\%.$$

Answer the following about unit changes.

(a) If price is recorded in *cents* ($x_c = 100x$), write the new regression equation $\hat{y}$ in terms of x_c. What happens to r and R^2?

Solution. $\hat{y} = 12.0 + 0.45(x_c/100) = 12.0 + 0.0045\,x_c$. Linear rescaling of x does *not* change r or R^2; both stay the same (r unchanged in sign/magnitude; $R^2 = 64\%$).

(b) Suppose ratings are converted to a *5-star* scale with $y_\star = y/20$. Write the regression of $y_\star$ on dollars x. What happens to r and R^2?

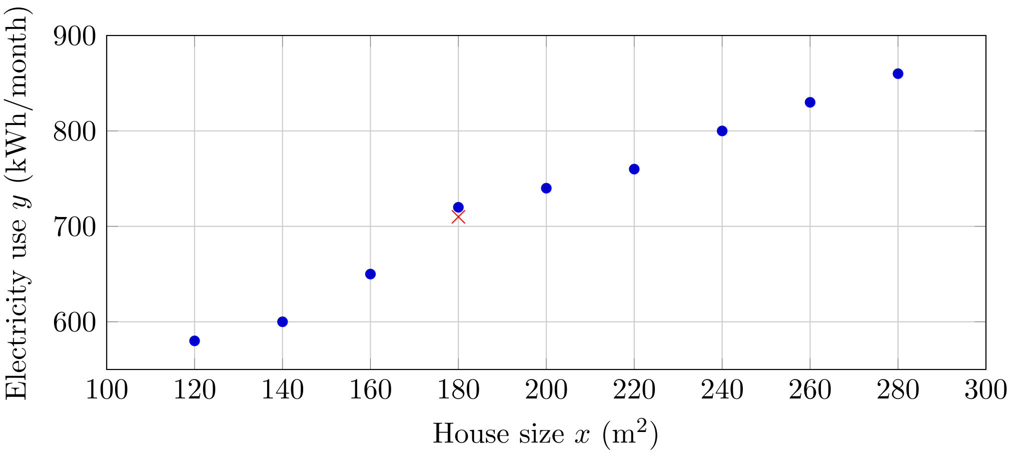
Solution. Divide both intercept and slope by 20: $\widehat{y_\star} = 12.0/20 + (0.45/20)x = 0.60 + 0.0225x$. Linear rescaling of y also leaves r and R^2 unchanged.

(c) Briefly explain why r and R^2 are invariant to these linear unit changes.

Solution. r standardizes both variables (center/scale), so multiplying or adding constants cancels out. R^2 depends only on r in simple linear regression ($R^2 = r^2$), so it is also invariant.

Problem 10. "Line through the means" & residual properties

For the homes below (electricity vs. size), the sample means are $\bar{x} = 180$ m^2 and $\bar{y} = 710$ kWh. The point $(\bar{x}, \bar{y})$ is marked with a red cross.



(a) Must the least-squares regression line pass through the cross at $(\bar{x}, \bar{y})$? Explain.

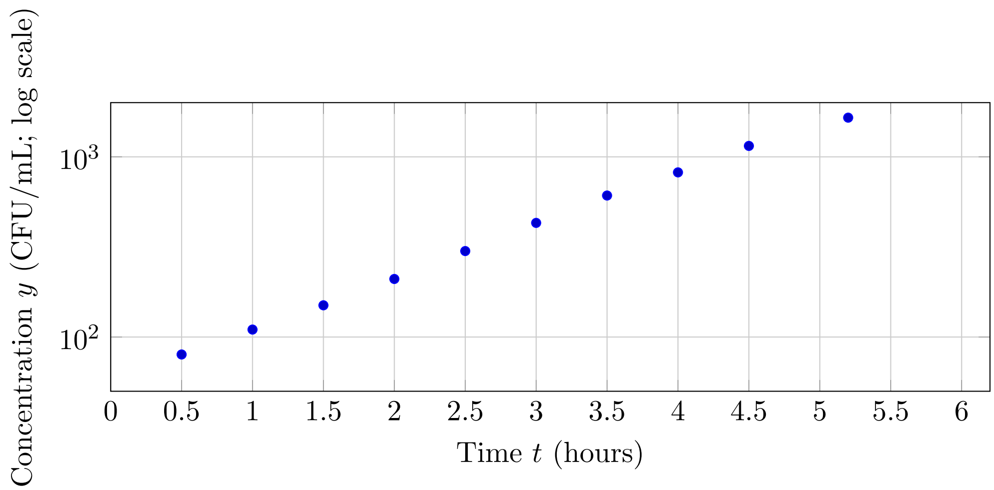
Solution. Yes. In simple linear regression, the LSRL always passes through $(\bar{x}, \bar{y})$ because the line ensures $\bar{y} = \hat{y}$ when residuals sum to zero.

(b) For any fitted least-squares line on this dataset, what is the sum and the mean of the residuals? Briefly justify.

Solution. The sum of residuals is 0 and the mean residual is 0. Least-squares with an intercept forces the residuals to balance out by construction.

Problem 11. Log re-expression and multiplicative interpretation

The plot shows bacterial concentration y (CFU/mL) versus incubation time t (hours) for $n = 10$ trials.



A linear model was fit to $\log_{10} y$ versus t, giving

$$\log_{10} \hat{y} = 2.10 + 0.18\,t, \qquad R^2 = 94\%.$$

(a) Interpret the slope 0.18 in *multiplicative* terms for y.

Solution. Each additional hour multiplies the predicted concentration by $10^{0.18} \approx 1.51$ (about a 51% increase) on average.

(b) Predict the concentration at $t = 3$ hours on the original scale and comment on model fit using R^2.

Solution. $\log_{10} \hat{y} = 2.10 + 0.18(3) = 2.64 \;\Rightarrow\; \hat{y} = 10^{2.64} \approx 437$ CFU/mL. With $R^2 = 94\%$, the log-linear model explains most of the variability in $\log_{10} y$, indicating a good fit.

(c) Briefly explain why the log transformation was appropriate based on the plot.

Solution. On the raw scale the growth is curved and multiplicative; on the log scale the points are approximately linear with roughly constant spread, matching linear-model assumptions.