Scatterplots, model residuals and predictions

Original skill preparation; versioned source scope

Scatterplots, Model Residuals and Predictions

- Describe the pattern in a scatterplot and read a **line of best fit** 最佳拟合直线.
- Interpret the slope and intercept of a fitted model in context.
- Find and interpret a **residual** 残差.
- Tell **interpolation** 内插 from **extrapolation** 外推, and judge when a prediction can be trusted.

Reading a scatterplot

- Direction: positive (up to the right) or negative.
- Form: linear, curved (such as exponential), or no clear pattern.
- Strength: how closely the points cluster around the pattern.
- An association between two variables does not show that one causes the other.

The fitted line

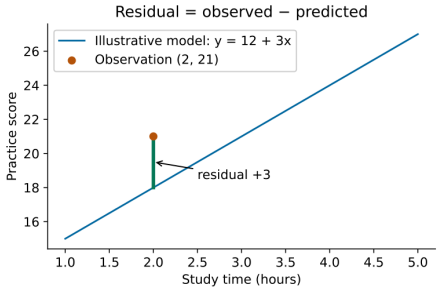
- In $\hat{y} = a + bx$, the slope b is the predicted change in y for each one-unit increase in x. Say "predicted", since real points vary.
- The intercept a is the predicted y at $x = 0$. If $x = 0$ is outside the data, it is only an extrapolation.
- To predict, substitute x into the equation, or read from the line, not from the nearest point.

Residuals

- A positive residual means the point lies above the line; negative means below.
- On a residual plot, a random scatter around zero suggests the linear model fits; a curve suggests another form.

$$\text{residual} = \text{observed } y - \text{predicted } y$$

Figure: Residuals



- The line y = 12 + 3x with the observation (2, 21) three units above it
- Match each part of the diagram to the text.

Linear or exponential models

- If equal steps in x give roughly equal differences in y, a linear model fits.
- If they give roughly equal ratios, an exponential model fits: $y = ab^x$.
- Look at the shape of the scatterplot before choosing.

Worked example: a delivery model

- Slope: each extra kilometre adds a predicted 1.5 minutes.
- Intercept: 8 minutes at 0 km, but 0 km is outside the data, so this is an extrapolation (it might be loading time).
- At 6 km: $T = 8 + 1.5(6) = 17$ minutes. The observed time was 15, so the residual is $15 - 17 = -2$ minutes: faster than predicted.

Worked example: choosing a model

- Differences: 12, 36, 108, not constant. Ratios: all 3.
- Model: $y = 6(3)^x$.
- Prediction at $x = 4$: $6 \times 81 = 486$, but only if the growth continues, since $x = 4$ is outside the data.

Common errors

- Getting the residual the wrong way round (predicted minus observed).
- Saying the slope "causes" the change.
- Trusting a prediction far outside the observed range.

Common errors (2)

- Reading a prediction from a nearby point instead of from the line.
- Fitting a line to clearly curved data.

Key terms

- **line of best fit**: A line that models the trend in a scatterplot.
- **residual**: The observed value minus the value the model predicts.
- **interpolation**: Predicting inside the range of the data.
- **extrapolation**: Predicting outside the range of the data.

Check yourself

- A model is $\hat{y} = 40 - 2x$. At $x = 5$ the observed value is 33. Find the residual.
- The data run from $x = 1$ to $x = 12$. Is a prediction at $x = 30$ **interpolation** or extrapolation?
- What does a negative residual tell you about a point?