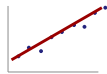


Exploring Two-Variable Data

AP Statistics ■ Unit 2

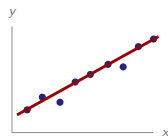
AP Statistics



Exploring Two-Variable Data

What we will cover

- 1 Related variables
- 2 Two categorical variables
- 3 Scatterplots and correlation
- 4 The regression line
- 5 Residuals and fit
- 6 Departures from linearity



1

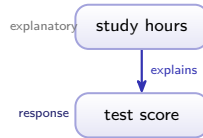
Related

Related variables

When two variables change together, they show an **association** 关联.

- **explanatory** 解释变量 variable predicts
- **response** 响应变量 variable is the outcome

An apparent pattern may be real – or a fluke of this one sample.



2 Categorical

Two categorical variables

A **two-way table** 双向表 gives three relative frequencies:

- **joint** 联合相对频率: $\text{cell} \div \text{grand total}$
- **marginal** 边际相对频率: $\text{margin} \div \text{total}$
- **conditional** 条件相对频率: $\text{cell} \div \text{row/col}$

Different **conditional distributions** 条件分布 $\Rightarrow$ association – but never **causation** 因果关系.

	pet	none
boy	30	20
girl	25	25

$P(\text{pet} \mid \text{boy}) = 30/50 = 0.60.$

3 Scatter

Exploring Two-Variable Data
└ Scatter

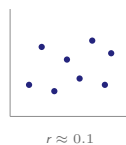
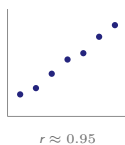
Scatterplots and correlation

Describe a **scatterplot** 散点图 by
direction 方向, **form** 形态,
strength 强度, outliers.

The **correlation** 相关系数 r measures
linear strength:

$$-1 \leq r \leq 1$$

r near ± 1 is tight; near 0 is a cloud. r is **not resistant**, and never implies cause.



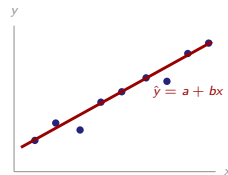
The regression line

The least-squares regression

line 最小二乘回归线:

$$\hat{y} = a + bx$$

- **slope** 斜率 b : predicted change in y per unit x
- **intercept** 纵截距 a : predicted y at $x = 0$



Use it to **predict** 预测, but never **extrapolate** 外推 far beyond the data.

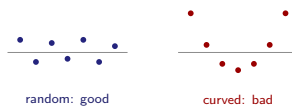
5 Residuals

Residuals and fit

A **residual** 残差 is the miss: $y - \hat{y}$.

- random **residual plot** 残差图 $\Rightarrow$ a line fits
- a curve $\Rightarrow$ line is wrong

r^2 决定系数 is the fraction of y 's variation the line explains; $b = r \frac{s_y}{s_x}$.



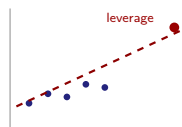
6 Departures

Departures from linearity

An **influential point** 影响点 swings the line when removed:

- a y-outlier sits far vertically
- a **high-leverage** 高杠杆 point sits at extreme x

If the pattern curves, a **transformation** 变换 (e.g. $\log y$) can straighten it for a line.



7 Recap

Worked example – using a regression line

$\hat{y} = 3 + 2x$ with $r = 0.85$. Predict y at $x = 10$; the actual value is 25. Find the residual and interpret r^2 .

- $\hat{y} = 3 + 2(10) = 23$
- Residual = actual – predicted
= $25 - 23 = +2$ (the line **under**-predicted).
- $r^2 = 0.85^2 = 0.7225 \Rightarrow$ **about 72% of the variation in y is explained by x .**
- Slope: each 1-unit rise in x predicts a 2-unit rise in y **on average**.

Residual = **actual** – **pre-dicted** (positive means above the line). Interpret in **context** with “on average” – and never claim causation from r alone.

Unit 2 – key takeaways

1 Association

explanatory vs response; conditionals show it

2 Correlation

$-1 \leq r \leq 1$; linear only; not resistant; not cause

3 Regression

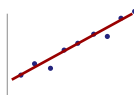
$\hat{y} = a + bx$; read slope + intercept; don't extrapolate

4 Fit

residual plot checks linearity; r^2 explained; leverage

Check yourself

- 1 What range must the correlation r fall in?
- 2 In $\hat{y} = a + bx$, what does the slope b mean?
- 3 What does a curved residual plot tell you?
- 4 Does a high r prove causation?



Answers 1. $-1 \leq r \leq 1$ 2. predicted change in y per unit x 3. a line is the wrong model 4. no