Linear Regression Model

1. Meaning of slope

In a regression line $\hat{y} = 2.5x + 10$, the slope 2.5 means:

A. y increases by 10 for every 1 increase in x
B. x increases by 2.5 for every 1 increase in y
C. y increases by 2.5 for every 1 increase in x
D. x increases by 10 for every 2.5 increase in y

---

2. Correlation and regression

If the correlation between x and y is close to 0, the regression line will:

A. Always be steep
B. Be nearly horizontal
C. Have weak predictive power
D. Pass through (0,0)

- A **linear regression model** describes the relationship between two quantitative variables using a line.
- The model is written as $\hat{y} = a + bx$, where **b** is the slope and **a** is the y-intercept.
- The **slope** represents the predicted change in the response variable for a one-unit increase in the explanatory variable.
- The model is used for **prediction**, not proof of causation.
- Valid use requires a **linear pattern, no strong outliers, and constant variability**.

- Linear regression — 线性回归
- Regression line — 回归直线
- Slope — 斜率
- Intercept — 截距
- Correlation (r) — 相关系数
- Residual — 残差
- Predicted value — 预测值
- Observed value — 观测值
- Outlier — 离群值
- Extrapolation — 外推（外插预测）

A teacher collects data on the number of hours students study per week (x) and their test scores (y). A computer output gives the following regression results:

$$\hat{y} = 12.5x + 48$$

Correlation: $r = 0.82$

(a) Interpret the slope

Interpret the slope in context.

(b) Interpret the intercept

Explain what the intercept means in this situation.

(c) Strength and direction

Describe the relationship between study time and test score.

(d) Prediction

Predict the test score for a student who studies 6 hours per week.

(e) Reasonableness

Is it reasonable to use this model to predict a student who studies 15 hours per week? Explain.

(f) Residual interpretation

A student who studies 4 hours scores 70, but the model predicts 60.

Compute and interpret the residual.

(a) Slope

Each additional 1 hour of study increases predicted score by **12.5 points**.

(b) Intercept

A student who studies **0 hours is predicted to score 48**.

(c) Relationship

Strong positive linear relationship between study time and score.

(d) Prediction

$$\hat{y} = 12.5(6) + 48 = 123$$

Answer: 123

(e) Reasonableness

Not reasonable — 15 hours is likely outside data range (extrapolation).

(f) Residual

Residual = observed − predicted = $70 - 60 = 10$
Student scored 10 points above prediction.

Stretch and Challenge

📊 1. Competing Models

A data set is modelled using two regression equations:

- Model A: $\hat{y} = 3.2x + 10, r^2 = 0.64$
- Model B: $\hat{y} = 5.1\sqrt{x} + 6, r^2 = 0.71$

Which model is better? Justify beyond just r^2.

📉 2. Transformation Decision

A scatterplot of x vs y shows a strong exponential increase.

Explain:

- whether a linear model is appropriate
- what transformation would improve the model
- what shape the transformed graph should take

📊 3. Outlier Impact

A data set has a strong positive linear relationship. One extreme point is added far from the cluster.

Explain:

- effect on slope
- effect on correlation
- effect on r^2

📈 4. Residual Pattern Analysis

A residual plot shows a clear U-shape pattern.

Explain:

- what this says about the model
- whether linear regression is appropriate
- what alternative model should be considered

1. Competing Models

Model B better: higher r^2, but also check residuals → choose **best pattern fit, not just** r^2.

---

2. Transformation Decision

Not linear → use **log or ln transform of y** → makes relationship linear.

---

3. Outlier Impact

Slope: may change a lot

Correlation: decreases

r^2: decreases

---

4. Residual U-shape

Model is **not appropriate (non-linear pattern)** → use **quadratic model**.

Test for Slope

- Slope (β_1) — 斜率（总体斜率）
- Null hypothesis (H_0) — 原假设
- Alternative hypothesis (H_1) — 备择假设
- t-statistic — t检验统计量
- p-value — p值
- Significance level (α) — 显著性水平
- Degrees of freedom — 自由度
- Standard error — 标准误
- Reject H_0 — 拒绝原假设
- Linear relationship — 线性关系

1. Tests whether there is a **linear relationship in the population.**
2. Based on the **population slope** β_1, not sample slope.
3. Null hypothesis usually: $H_0 : \beta_1 = 0$ (no linear relationship).
4. Alternative hypothesis: $H_1 : \beta_1 \neq 0$ (or one-sided).
5. Uses a **t-test for slope**.
6. Test statistic = (estimated slope − 0) / standard error.
7. Assumes **LINE conditions** (Linearity, Independence, Normality, Equal variance).
8. Small p-value ⇒ **strong evidence against H_0.**
9. If $p < \alpha$, we **reject H_0** and conclude a linear relationship exists.
10. Conclusion must be written in **context (real-world interpretation)**.

A researcher studies the relationship between hours of exercise per week (x) and resting heart rate (y). A sample of 25 people is collected, and the regression output is:

$$\hat{y} = -1.8x + 78$$

Standard error of slope = 0.6
Test statistic $t = -3.00$
p-value = 0.006
df = 23

(a) State hypotheses

Write the null and alternative hypotheses.

(b) Conditions

State the conditions needed for this test.

(c) Decision

At α = 0.05, what is your conclusion?

(d) Interpretation

Interpret the result in context.

(e) Meaning of slope

Interpret the slope in context.

(a) Hypotheses

$H_0 : \beta_1 = 0$ (no linear relationship)
$H_1 : \beta_1 \neq 0$

(b) Conditions (LINE)

Linearity, Independence, Normal residuals, Equal variance.

(c) Decision

p = 0.006 < 0.05 ⇒ **Reject** H_0

(d) Conclusion (context)

There is **statistically significant evidence of a linear relationship** between exercise and resting heart rate.

(e) Slope meaning

Each additional 1 hour of exercise per week is associated with a **decrease of 1.8 bpm** in resting heart rate.

True or False

- H0: $\beta 1=0$ means there is no linear relationship in the population

True or False

- A small p-value means we fail to reject the null hypothesis. —

True or False

- The t-test for slope is used to test whether the sample slope is significant.

False – testing population slope is significant