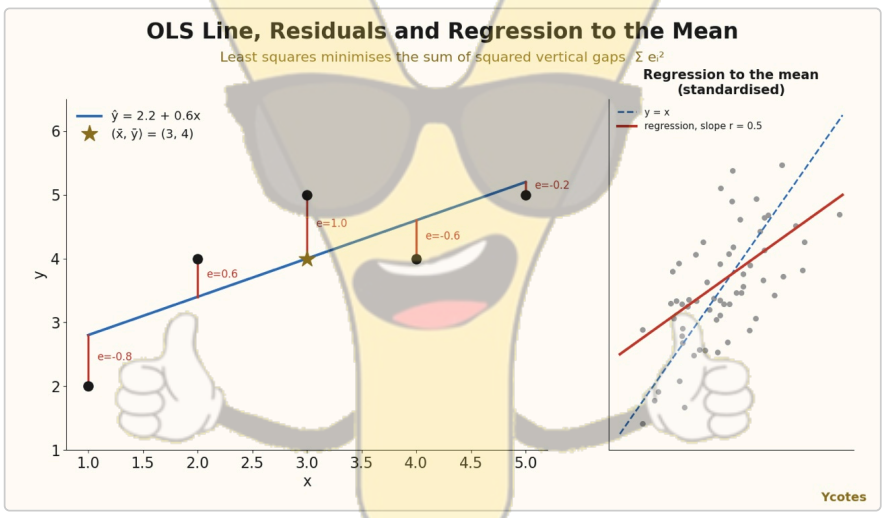
UNIT–4 : Regression (PYQ Solution)

PART–A Answers (≈70 Words Each)

Q1. What is a residual? What does regression to the mean mean? (Expected)

A residual is the vertical gap between the observed and fitted value: $e_i = y_i - \hat{y}_i$. Least squares chooses the line that minimises $\sum e_i^2$; residuals sum to zero and are used for diagnostics.

Regression to the mean: because $|r| < 1$, an extreme value of x is on average followed by a less extreme y (measured in SD units, $\hat{z}_y = r z_x$). A very tall father tends to have a tall, but on average shorter, son.



PART–B Answers (≈250 Words Each)

Q1. In a partially destroyed record of a correlation analysis only the following is legible: variance of x = 9, regression equations 8x - 10y + 66 = 0 and 40x - 18y = 214. Find (i) the means of x and y (ii) the correlation coefficient (iii) the standard deviation of y. (2020, old scheme)

(i) Means: both regression lines pass through $(\bar{x}, \bar{y})$. Solve $8x - 10y = -66$ and $40x - 18y = 214$. Multiply the first by 5: $40x - 50y = -330$. Subtract: $32y = 544 \Rightarrow \bar{y} = 17$, then $8\bar{x} = 10(17) - 66 = 104 \Rightarrow \bar{x} = 13$.

(ii) Correlation: assume $8x - 10y + 66 = 0$ is y on x: $y = 0.8x + 6.6$, so $b_{yx} = 0.8$. Then $40x - 18y = 214$ is x on y: $x = 0.45y + 5.35$, so $b_{xy} = 0.45$.

$$r^2 = b_{yx}\, b_{xy} = 0.8 \times 0.45 = 0.36 \Rightarrow r = +0.6$$

(sign of r is the sign of the slopes, positive). The other choice gives $b_{xy} = 1.25,\ b_{yx} = 2.22$, product 2.78 > 1, which is impossible, so the assumption is right.

(iii) SD of y: $\sigma_x = \sqrt{9} = 3$.

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \Rightarrow 0.8 = 0.6 \times \frac{\sigma_y}{3} \Rightarrow \sigma_y = 4$$

Answer: $\bar{x} = 13,\ \bar{y} = 17,\ r = 0.6,\ \sigma_y = 4.$

Q2. Fit a simple linear regression by least squares to x: 1 to 5, y: 2, 4, 5, 4, 5; find R^2 and test the slope. (Expected)

$n = 5$, $\bar{x} = 3$, $\bar{y} = 4$, $S_{xx} = \sum(x - \bar{x})^2 = 10$, $S_{xy} = \sum(x - \bar{x})(y - \bar{y}) = 6$, $S_{yy} = 6$.

OLS estimates: $b_1 = S_{xy}/S_{xx} = 0.6$, $b_0 = \bar{y} - b_1\bar{x} = 4 - 1.8 = 2.2$. So $\hat{y} = 2.2 + 0.6x$.

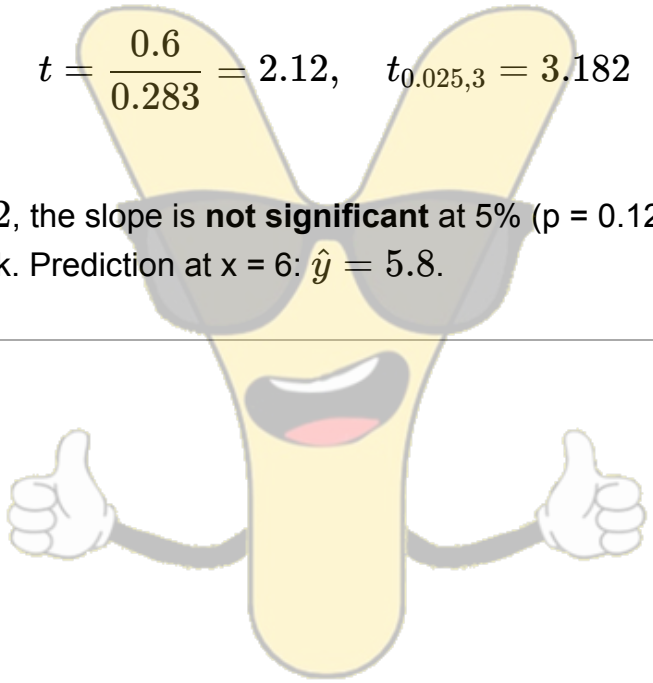
x	y	$\hat{y}$	e
1	2	2.8	-0.8
2	4	3.4	0.6
3	5	4.0	1.0
4	4	4.6	-0.6
5	5	5.2	-0.2

$SSE = 0.64 + 0.36 + 1 + 0.36 + 0.04 = 2.4$, $SSR = 6 - 2.4 = 3.6$, so $R^2 = 3.6/6 = 0.60$ (r = 0.775).

Test of slope $H_0: \beta_1 = 0$: $s^2 = SSE/(n-2) = 0.8$, $s = 0.894$, $SE(b_1) = s/\sqrt{S_{xx}} = 0.283$.

$$t = \frac{0.6}{0.283} = 2.12, \quad t_{0.025,3} = 3.182$$

Since $2.12 < 3.182$, the slope is **not significant** at 5% (p = 0.124); with only 5 points the evidence is weak. Prediction at x = 6: $\hat{y} = 5.8$.



PART–C Answers (≈500 Words Each)

Q1. If theta is the angle between the two regression lines, show that tan(theta) = ((1 - r^2)/r) sigma_x sigma_y/(sigma_x^2 + sigma_y^2). Explain the significance when r = 0 and r = +-1. (2019, 2020, old scheme)

Regression lines:

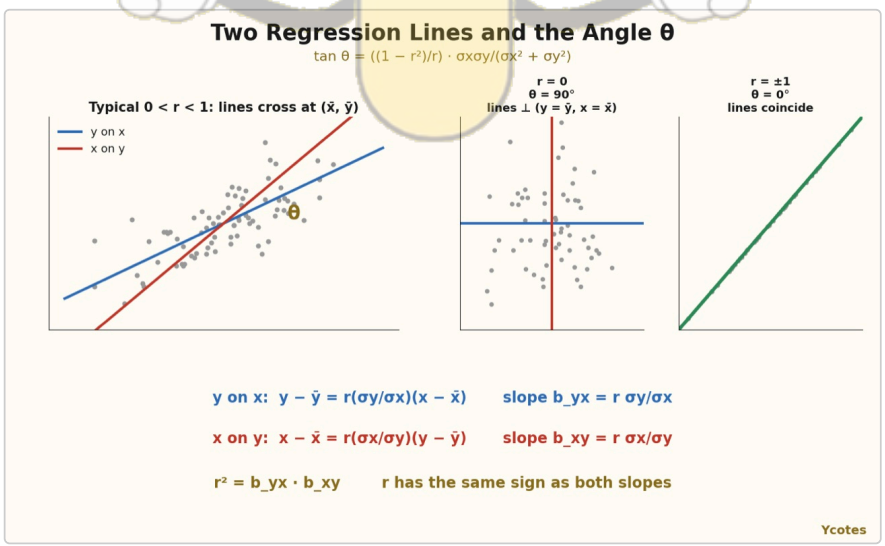
- y on x: $y - \bar{y} = r\dfrac{\sigma_y}{\sigma_x}(x - \bar{x})$, slope $m_1 = r\dfrac{\sigma_y}{\sigma_x}$
- x on y: $x - \bar{x} = r\dfrac{\sigma_x}{\sigma_y}(y - \bar{y})$, i.e. $y - \bar{y} = \dfrac{\sigma_y}{r\sigma_x}(x - \bar{x})$, slope $m_2 = \dfrac{\sigma_y}{r\sigma_x}$

Angle between two lines: $\tan\theta = \left|\dfrac{m_2 - m_1}{1 + m_1 m_2}\right|$.

$$m_2 - m_1 = \frac{\sigma_y}{\sigma_x}\left(\frac{1}{r} - r\right) = \frac{\sigma_y}{\sigma_x}\cdot\frac{1 - r^2}{r}, \qquad 1 + m_1 m_2 = 1 + \frac{\sigma_y^2}{\sigma_x^2} = \frac{\sigma_x^2 + \sigma_y^2}{\sigma_x^2}$$

$$\tan\theta = \frac{\sigma_y}{\sigma_x}\cdot\frac{1 - r^2}{r}\cdot\frac{\sigma_x^2}{\sigma_x^2 + \sigma_y^2} = \boxed{\frac{1 - r^2}{r}\cdot\frac{\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2}}$$

(For r negative the modulus is taken, so θ is the acute angle.)



Significance:

- **r = 0:** $\tan\theta \to \infty$, so $\theta = 90^\circ$. The lines become $y = \bar{y}$ and $x = \bar{x}$, perpendicular. Variables are uncorrelated, and x gives no help in predicting y.
- **r = +1 or -1:** $\tan\theta = 0$, so $\theta = 0^\circ$. The two lines coincide: perfect linear relationship, all points on one line (positive slope for r = +1, negative for r = -1).
- As $|r|$ increases from 0 to 1, the angle shrinks from 90 degrees to 0, so the **smaller the angle, the stronger the correlation**.

Q2. Fit a second degree parabola to x: 1 to 9 and y: 2, 6, 7, 8, 10, 11, 11, 10, 9. (2019, old scheme)

Fit $y = a + bx + cx^2$. Normal equations:

$$\sum y = na + b\sum x + c\sum x^2$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

From the data: $n = 9,\ \sum x = 45,\ \sum y = 74,\ \sum x^2 = 285,\ \sum x^3 = 2025,\ \sum x^4 = 15333,\ \sum xy = 421,\ \sum x^2y = 2771.$

$$74 = 9a + 45b + 285c$$

$$421 = 45a + 285b + 2025c$$

$$2771 = 285a + 2025b + 15333c$$

Solving the three equations: $a = -0.9286,\ b = 3.5232,\ c = -0.2673.$

$$\boxed{y = -0.929 + 3.523x - 0.267x^2}$$

Check: at x = 5 the fitted value is $-0.929 + 17.616 - 6.683 = 10.0$ (actual 10). The negative c shows the curve rises and then falls, with its peak near $x =$

$-b/2c = 6.6$.

Q3. Fit a second degree parabola to x: 0, 1, 2, 3, 4 and y: 1, 1.8, 1.3, 2.5, 6.3. (2020, old scheme)

$n = 5,\ \sum x = 10,\ \sum y = 12.9,\ \sum x^2 = 30,\ \sum x^3 = 100,\ \sum x^4 = 354,\ \sum xy = 37.1,\ \sum x^2 y = 130.3$.

$$12.9 = 5a + 10b + 30c$$

$$37.1 = 10a + 30b + 100c$$

$$130.3 = 30a + 100b + 354c$$

Eliminating: from the first two, $37.1 - 2(12.9) = 11.3 = 10b + 40c$. From the first and third: $130.3 - 6(12.9) = 52.9 = 40b + 174c$. Multiply the first result by 4: $45.2 = 40b + 160c$; subtract: $7.7 = 14c \Rightarrow c = 0.55$. Then $10b = 11.3 - 22 = -10.7 \Rightarrow b = -1.07$, and $5a = 12.9 + 10.7 - 16.5 = 7.1 \Rightarrow a = 1.42$.

$$\boxed{y = 1.42 - 1.07x + 0.55x^2}$$

Check: at x = 4: $1.42 - 4.28 + 8.8 = 5.94$ (actual 6.3). A parabola is a **non-linear (polynomial) regression**, linear in the coefficients a, b, c.

Q4. Explain multiple regression with adjustment, diagnostics, interaction terms, non-linear transforms and qualitative predictors. (Expected)

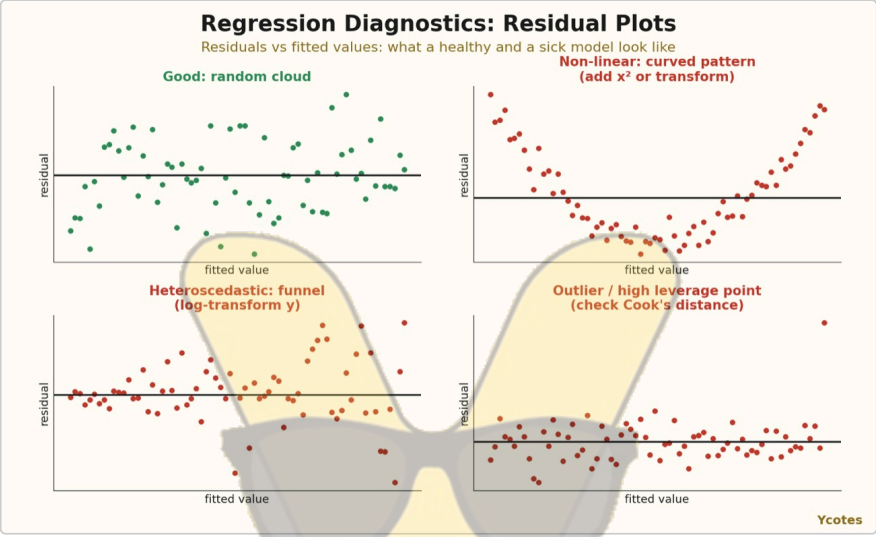
Multiple regression: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \varepsilon$, fitted by OLS $\hat{\beta} = (X^T X)^{-1} X^T y$. Each β_j is the change in y per unit change in x_j **holding the other predictors fixed**; this is **adjustment** for confounders.

Example (8 points): $\hat{y} = 1.5 + 1.5x_1 + 1.0x_2$, $R^2 = 0.996$. Adjusted $R^2 = 1 - (1 - R^2)\dfrac{n-1}{n-k-1} = 0.994$; overall $F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)} = 626$.

Diagnostics (check with residual plots):

1. **Linearity:** curved residual pattern means add x^2 or transform.

2. **Constant variance:** funnel shape (heteroscedasticity) means transform y (log) or use weights.
3. **Normality of residuals:** Q-Q plot.
4. **Independence:** Durbin-Watson for time-ordered data.
5. **Multicollinearity:** $VIF_j = \dfrac{1}{1 - R_j^2}$; VIF > 5 to 10 is a problem (in the example $r_{x_1 x_2} = 0.905$, $VIF = 5.5$).
6. **Outliers and influence:** leverage, Cook's distance.



Interaction terms: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. The effect of x_1 is $\beta_1 + \beta_3 x_2$, so it depends on x_2 (for example, effect of study hours differs by gender).

Non-linear transforms: add x^2 (polynomial), or use $\log x$, $\sqrt{x}$, $\log y$ to straighten curves or stabilise variance; the model stays linear in the coefficients.

Qualitative predictors: a category with k levels is coded with $k - 1$ **dummy variables** (0/1); the omitted level is the baseline. If $D = 1$ for "female", β_D is the average difference in y between female and male at the same x.
