Regression

→ Prediction is a key task of statistics.

let, there be heights of 928 people, $\bar{x} = 69.1$ in. Now, predict the height of son whose father is 72 in. This addition information helps us to make a better prediction. Regression does just that.

Correlation coefficient :-

The scatter plot is useful in visualizing two quantitative variables.

Things we can get of scatter plot:

① it may have a direction (sloping up/down)
② form (a scatter that clusters around a line is linear)
③ strength (how closely do points follow form)

→ If the form is linear, then a good measure of strength is correlation coefficient (r)

(x_i, y_i) → data

$$r = \frac{1}{n} \sum_{i=1}^{n} \frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}$$

s_x, s_y → std dev

Correlation measures linear association :-

If we have pairs of data and the scatter follows a linear form, then we can summarize the data by $\bar{x}, s_x, \bar{y}, s_y, r$

→ When we plot these pairs x & y, then we use
x → explanatory variable or predictor
y → response variable.

⇒ r is always b/w -1 & 1.
$r = 1$ mean a perfect positive linear relationship

⇒ r is not affected by changing the center or the scale of either variable.

⇒ Correlation coefficient is only useful for measuring linear association.

⇒ Also, correlation does not mean causation.

Regression line :-

If the scatterplot shows a linear association, then this relationship can be summarized by a line.

line equation, $\hat{y}_i = a + b\hat{x}_i$

The idea is to chose the line that minimizes the sum of squared distances between the observed y_i & $\hat{y}_i$.

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (a+bx_i))^2$$

Method of least squares.

$b = r\frac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$. This line $\hat{y} = a+bx$ is regression line.

↳ the expected mean value of y when $x=0$
the value at which regression line crosses y-axis.

Regression to the mean :-

The main use of regression is to predict y from x.
Given x, $\hat{y} = a+bx$

→ The prediction for y at $x=\bar{x}$ and $y=\bar{y}$,

$b = r\frac{s_y}{s_x}$ x

→ but if x is 1 standard deviation above $\bar{x}$, then $\hat{y}$ is only $r s_y$ above $\bar{y}$

Regression effect → the bottom group on the first test will on average show some improvement on the second test & the top group will on average fall back.

(or)

extremely low/high variables move close to average when measured second time.

Predicting y from x & x from y :-

If we are given x, we use the regression line $\hat{y} = a + bx$ to predict y

→ for this we need $\bar{x}, \bar{y}, s_x, s_y$ and r.

Q) if avg midterm score = 49.5 ; S = 10.2 ; r = 0.67
avg final score = 69.1 ; S = 11.8
If a student get 41 in midterm, predict final score.

A) 41 is 8.5 below average
⇒ 0.83 std below average $\left[z = \frac{41 - 49.5}{10.2} = -0.833 \right]$

Standardizing it,
looking at ~~score~~ formula of slope of regression line, we predic final score would be

~~$b = r \frac{s_y}{s_x}$ $\bar{x} + 0.83$~~ $r \times 0.83 \times S_{final}$

⇒

⇒ $\hat{y} = a + bx$
$= 69.1 - (0.67 \times 0.83 \times 11.8) = \underline{62.5}$

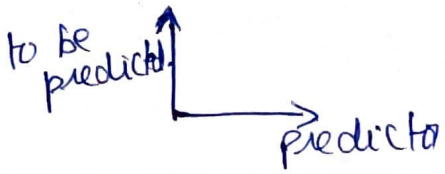
⇒ If student get 58, the final score would be ; 8.5 + average.

~~58 (0.67 × 0.83 × 11.8)~~ $\frac{58 - 49.5}{10.2} = +0.83$

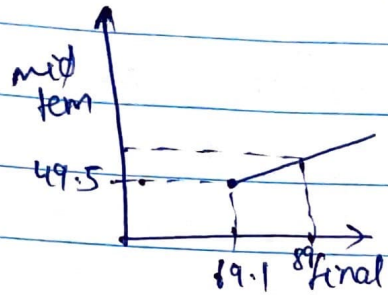
⇒ 5 | 69.1

⇒ ~~58~~ + (0.67 × 0.83 × 11.8) 51 + (0.67 × 11.8)

⇒ ~~64.32~~ 74.56



⇒ If student get 89 in final, find mid-term.



89 is above average

$$z = \frac{89 - 69.1}{11.8} = 1.68 \text{ std dev above avg}$$

$$\Rightarrow 49.5 + (1.68 \times 10.2)$$

$$\Rightarrow \underline{60.98}$$

Normal approximation given x :

Regression requires that the scatter plot is football shaped.

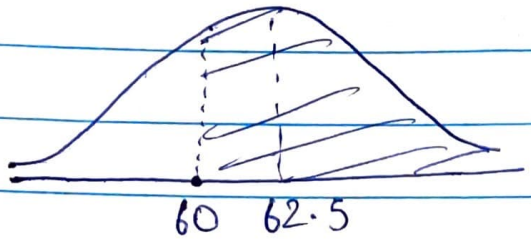
For a given value of x, we can predict y value simply looking a point falling on y, but y values near the pairs at x value follow a normal curve

→ To standardize, subtract off the predicted value of $\hat{y}$, then divide by $(\sqrt{1-r^2}\, s_y)$

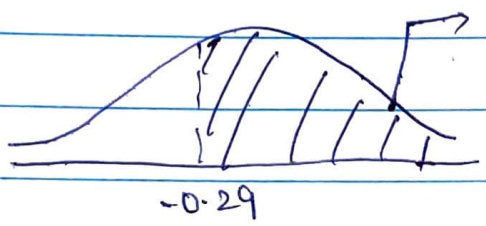
Q Among student who scored around 41 in midterm, what % scored above 60 in final?

A 62.5 in ~~for~~ final

⇒ Normal curve is centered at 62.5



$$\frac{60 - 62.5}{\sqrt{1-(0.67)^2} \times Sf} = -0.29$$



Normal approximation use in linear regression :-

Tells us more about y-values. From regression, we know the average (predicted value for specific x), and the normal approximation tells us more about the actual values for a specific x and how they look like, as they follow normal curve.

Residual plots :-

For each observation, we have an observed 'y' value and we have a predicted 'y' value. Difference b/w them is residual. It is used for checking if regression is appropriate.

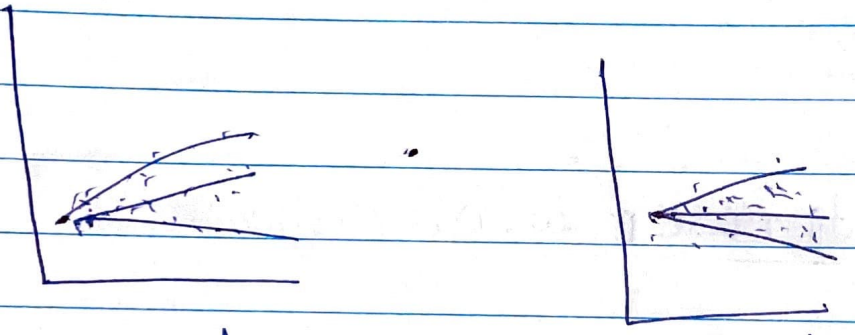
$$e_i = y_i - \hat{y}_i$$

It should show an unstructured horizontal line.

→ A curved pattern suggests scatter is not linear. Linear Regression should not be applied but it may still be possible to analyze these data with regression after transforming data.

Heteroscedastic :-

Variability changes with x values.



normal

residual.

If a plot looks hetero, take log of both variables.

→ Points with very large residual is outlier.