Week 11

Monday, 15 January 2024 9:16 AM

SCATTER PLOTS → between two numerical variables

FORM → Linear / parabolic / Exponential

DIRECTION → POSITIVE OR NEGATIVE

Correlation value → STRENGTH → STRONG / MODERATE / WEAK

Question 1 : Correlations

Order the correlation magnitudes corresponding to the following scatterplots from highest (strongest) to lowest (weakest).

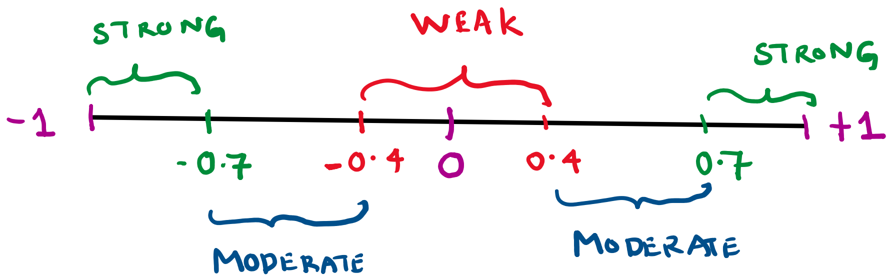
(a) strong association but Weak Correlation

(b) Strong Positive Linear relationship

(c) Weak positive Linear relationship

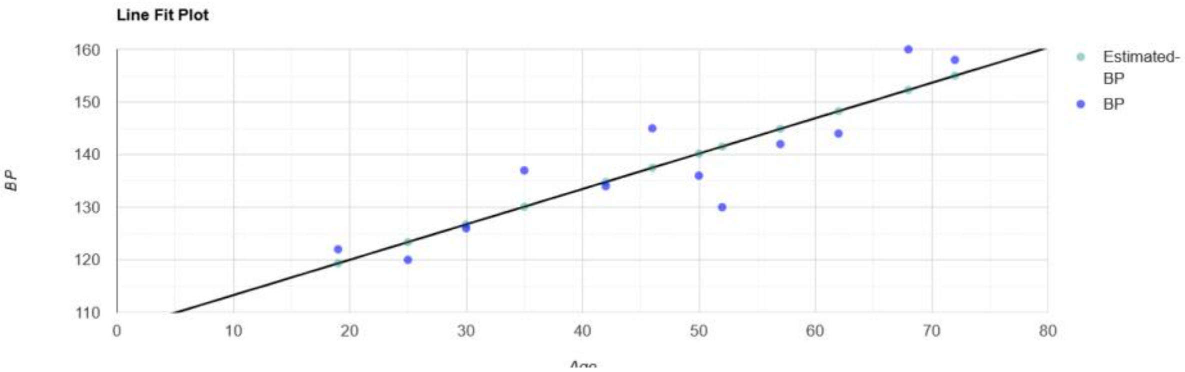
(d) strong Negative Linear relationship

CORRELATION VALUE , R



Age	19	25	30	35	42	46	50	52	57	62	68	72
BP	122	120	126	137	134	145	136	130	142	144	160	158

Using statskingdom.com



R-Squared (R^2) equals **0.7942**.

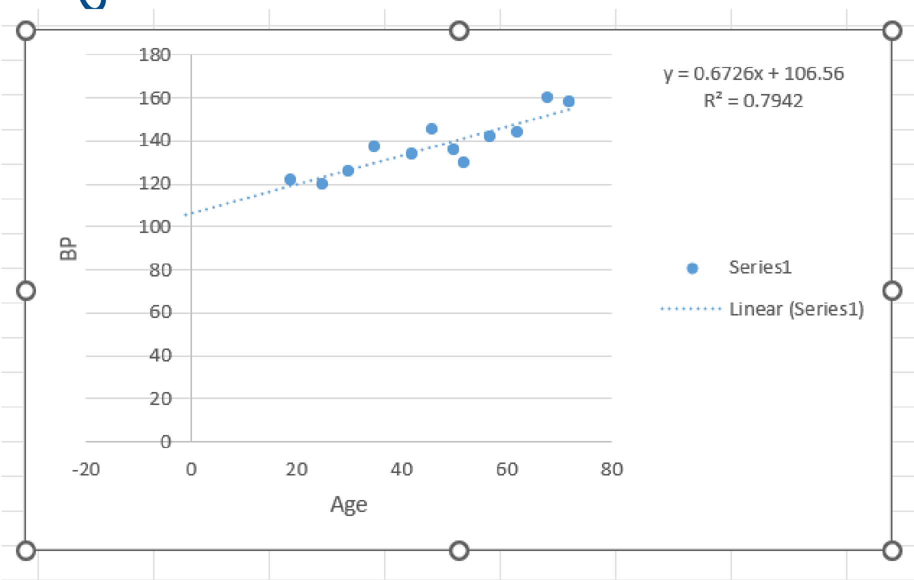
Correlation (R) equals **0.8912**.

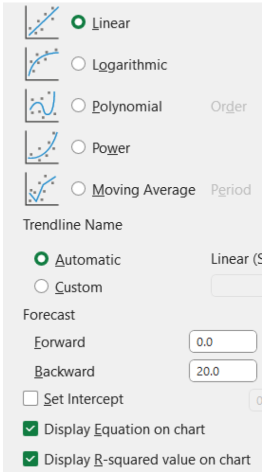
Regression line equation

$$\hat{Y} = 106.5577 + 0.6726X$$

$$Y = b_0 + b_1 x$$

Using Microsoft Excel,





$$R^2 = 0.7942$$

$$R = \sqrt{0.7942} = 0.8912$$

- The correlation coefficient (r) gives us a numerical measurement of the strength of the linear relationship between the explanatory and response variables.

(X-value) ↓ Horizontal Axis ↓ Independent Variable

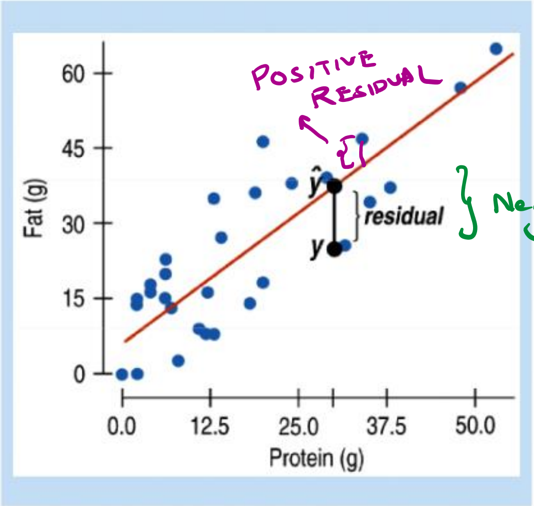
(Y-value) ↓ Vertical Axis ↓ Dependent Variable

RESIDUALS :

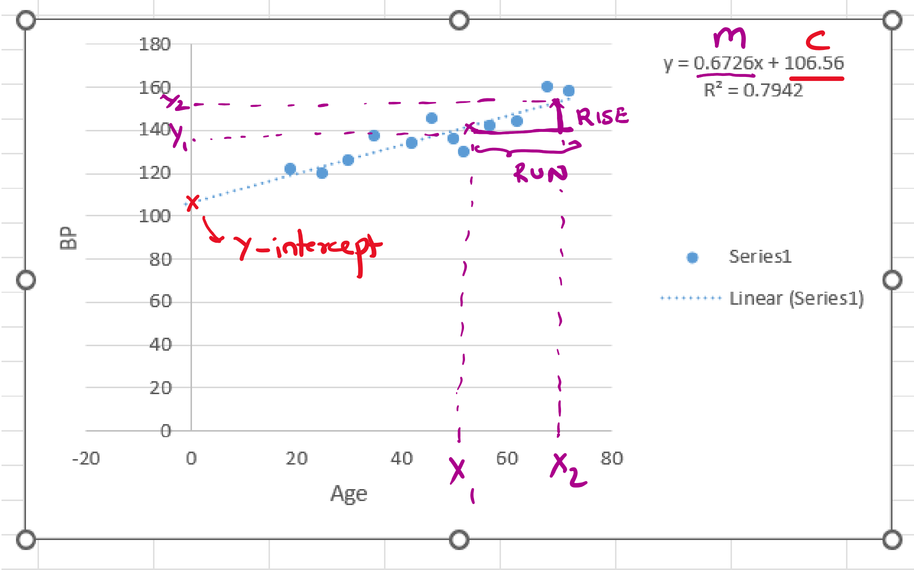
$$residual = observed - predicted = y - \hat{y}$$

- A negative residual means the predicted value (the line) is above the observation

- A positive residual means the predicted value (line) lies below the observation



LINEAR REGRESSION EQUATION → LINEAR MODEL



$$Y = mx + c$$

$$Y = b_0 + b_1 X$$

SLOPE → m or b_1

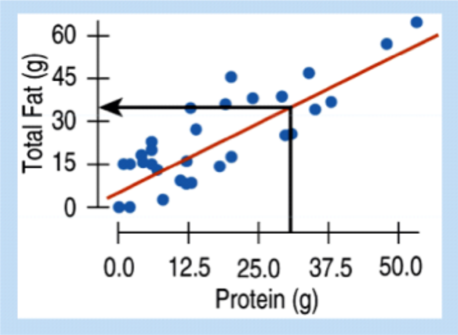
Y-intercept → c or b_0

b_0 or $C = 106.56$

$$\text{Slope} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{\text{RISE}}{\text{RUN}} = 0.6726 = m \text{ or } b_1$$

Fat Versus Protein Example

- The regression line for the Burger King data fits the data well:
  - The equation is
    $\hat{y} = b_0 + b_1 x$
    $\widehat{fat} = 6.8 + 0.97\, protein.$
  - The *predicted fat* content for a BK Broiler chicken sandwich (30gm protein) is 6.8 + 0.97(30) = 35.9 grams of fat. (Note SPSS will compute the linear model for you.



$b_1 \rightarrow$ Slope $= 0.97$

$b_0 \rightarrow$ Y-intercept $= 6.8$

Correlation, $R = 0.83$

$R^2 = (0.83)^2 = 0.69$

69% of the variation in fat is explained by the Linear Model

OR

31% of the variability in Fat is left in the Residuals

R-value $\longrightarrow$ strength of the linear model

R^2-value $\longrightarrow$ How much of the variation (in Y) is explained by the Linear Model.

SLOPE (m or b_1):

For each unit of x-value, Y-value increases by b_1

For 1g of protein, Fat increases by 0.97 g.

Y-intercept (C or b_0)

When $x = 0$, $y =$ C-value or b_0

When there is no protein, Fat content is 6.8 g

Question 3 : Interpreting the linear regression equation

Using the calculated regression equation $\hat{y} = 6.8 + 0.97x$ for the Burger King fat and protein data, answer the following:

1. Interpret the slope in the context of the variables fat and protein.
2. Is it appropriate to interpret the intercept? Explain.
3. Use the equation to predict the fat content for a menu item that has 40 grams of protein
4. Calculate the residual for a chicken sandwich that has 31 grams of protein and 22 grams of fat.
5. Should the regression equation be used to predict the fat content for a menu item with 100 grams of protein? Explain.

1)

For 1g of protein, Fat increases by 0.97g.

2) Y-intercept is meaningful

When there is no protein, Fat content is 6.8 g

3)

$\hat{y} = 6.8 + 0.97x$

$x = 40g$

$\hat{y} = 6.8 + (0.97 \times 40)$

$\hat{y} = 6.8 + 38.8 = 45.6g$

Reliable as interpolation

5) $x = 100g$

Not reliable as we have to do extrapolation.

$\hat{y} = 6.8 + (0.97 \times 100)$

$= 103.8g$

observed

4) protein, $x = 31g$ Fat, $y = 22g$

$$\hat{y} = 6.8 + 0.97x$$

$y = 6.8 + (0.97 \times 31)$

$y = 36.87$

Predicted Fat, $y = 36.87$

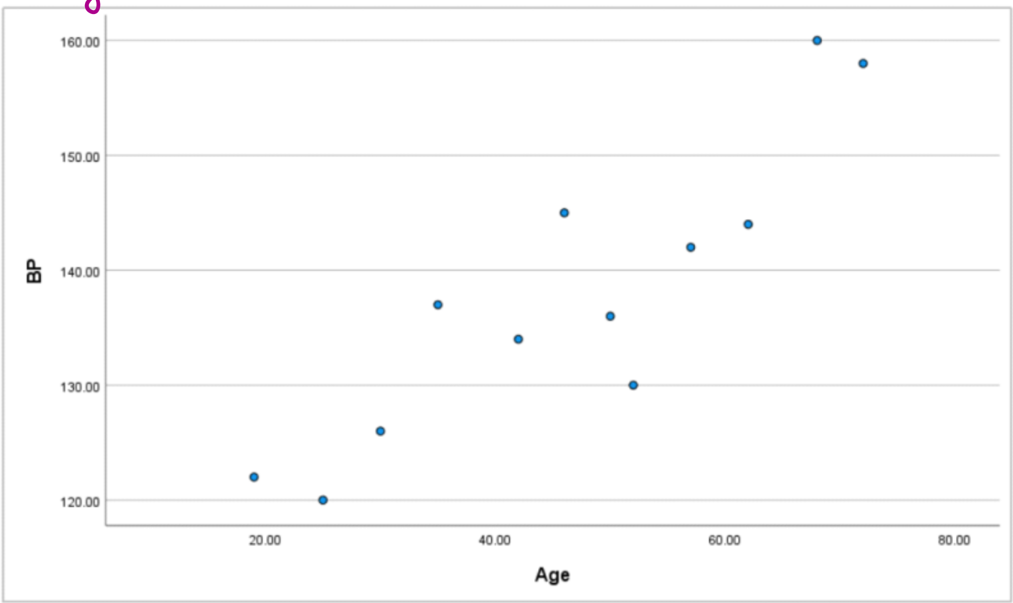
$residual = observed - predicted = y - \hat{y}$

$= 22 - 36.87 = -14.87$

Negative residual of 14.87

Age	19	25	30	35	42	46	50	52	57	62	68	72
BP	122	120	126	137	134	145	136	130	142	144	160	158

Using SPSS



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.891[a]	.794	.774	6.07546

a. Predictors: (Constant), Age

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
Y-intercept 1	(Constant) b_0	106.558	5.331		19.988	<.001
slope	Age b_1	.673	.108	.891	6.212	<.001

a. Dependent Variable: BP

Y-intercept, b_0 = 106.558

slope, b_1 = 0.673

$y = b_0 + b_1 x$

$y = 106.558 + 0.673x$

H_0: There is no Linear relationship

H_A: There is a Linear relationship

OR

H_0: slope is equal to zero, $b_1 = 0$

H_A: slope is not equal to zero, $b_1 \neq 0$

Test-statistic = 6.212

p-value = < 0.001

p-value is Low, Reject Null Hypothesis

∴ There is a Linear relationship