Contents

Day 24

Scatterplot

- Different colors to indicate different groups
- Each dot is a case [(x, y) point]
- Temperature → explanatory variable
- Scale → response variable

Interpreting a Scatterplot

1. <u>Direction</u> of the association
2. <u>Form</u> of the associations
3. <u>Strength</u>
4. <u>Outliers</u>

Direction

Linear

- <u>One</u> ellipse major axis describes relationship well.

Positive

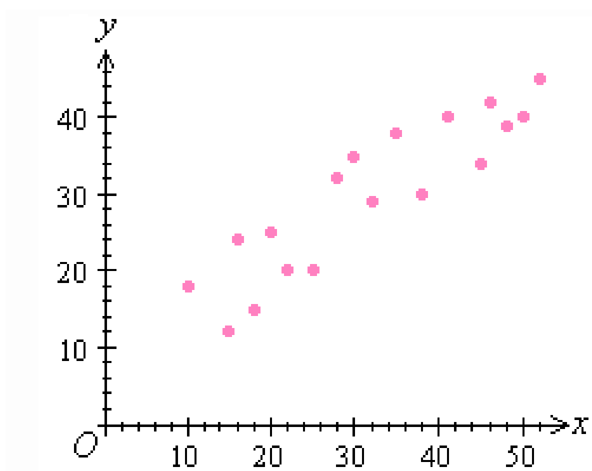

Figure 1: Positive Association

- X ↑, Y ↑
- Linear

Negative

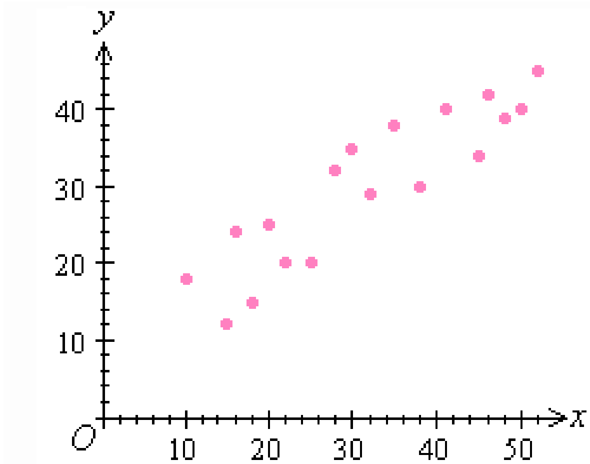

Figure 2: Negative Association

- X ↑, Y ↓
- Linear

No Association

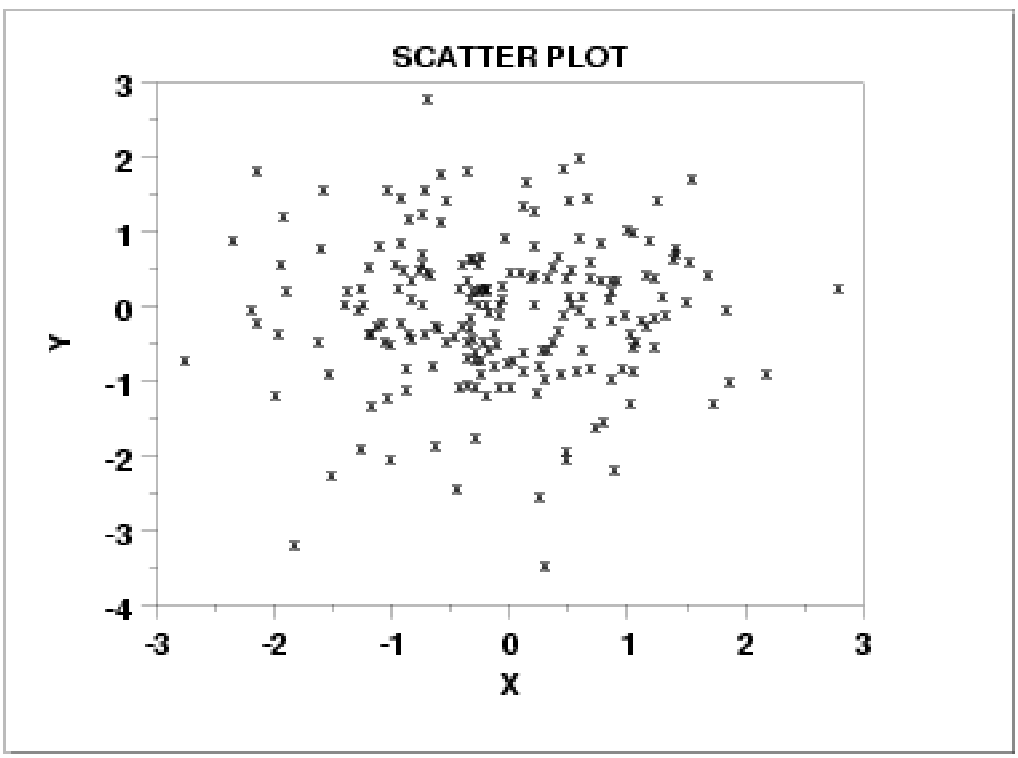

Figure 3: No Association

- Best we can do is a horizontal line

More Complex Association

- X ↑, Y ≅
- Polynomial
- Sinusoidal $(\sin(), \cos())$

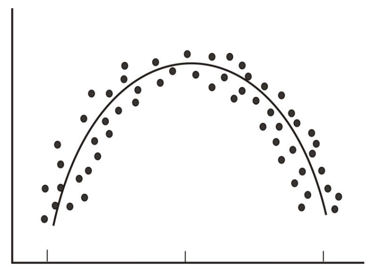

Figure 4: Complex

Nonlinear

- Exponential
- Logarithmic Power
- Need multiple ellipses to describe relationship

Strength

Only makes sense to discuss one direction & form are identified!

How closely the points follow the form you identified.

Correlation:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x} \right) \left(\frac{y - \bar{y}}{s_y} \right)$$

Figure 5: Correlation Formula

Contribution to r is t

- $x_{\mathrm{i}} > \bar{x} \rightarrow y_{\mathrm{i}} > \bar{y}$
- *fill in later from picture*

Notes about Correlation

- $-1 \leq r \leq 1$
  - $r = 1$: all points on line with positive slope
  - $r = -1$: all points on line with negative slope
- `r` is only interpretable for <u>linear</u> association!
  - Can have very strong non linear association but correlation close to 0. See more complex association figure.
- Correlation is <u>unitless</u> and <u>invariant</u> to linear transformation.
- Correlation is <u>highly</u> susceptible to outliers.

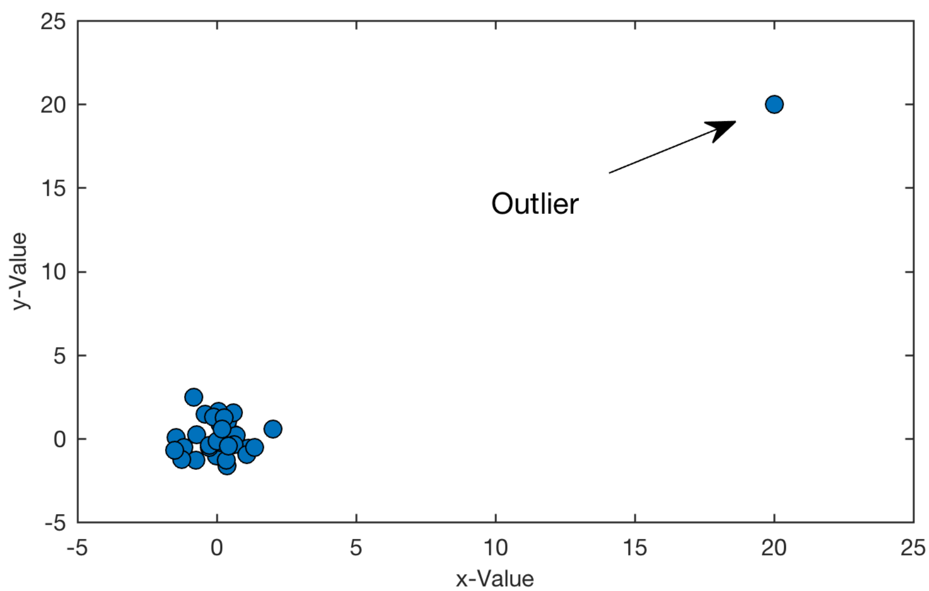

Figure 6: Outlier messing things up

- Correlation $\cong 0.85$
- Major influence on correlation

Outliers

Linear Regression

In population, x and y are related:

$$y = \beta_0 + \beta_1 + \epsilon$$

- β_0 = y-intercept (b portion in $y = mx + b$)
- β_1 = slope (m portion in $y = mx + b$)
  - both above are parameters
- $\epsilon \sim N(0, \sigma)$
  - Random variable

X is assumed fixed and is not random.

$\beta_0 + \beta_1$ is not random. - You give me x, I give you $\beta_0 + \beta_1 x$

Y is a random variable because ϵ is a random variable.

Before I observe the case:

- I know x-value
- I do not know y-value

Problem: β_0 and β_1 are parameters BUT we have sample data.

How to estimate β_0 and β_1?

Criterion:

$$SS_{(residuals)} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Figure 7: Sum of Squared Residual Formula

- $\hat{y}$ = "predicted y" = value of `y` obtained by plugging `x` into the equation of the line.

Minimize the criterion over all possible lines $\hat{y} = mx + b$

In stats: $\hat{y} = b_0 + b_1 x$. This is called the least squares regression line.

- $b_1 = r \times \frac{s_y}{s_x}$
- $b_0 = \bar{y} - b_1 \bar{x}$

Interpretation

- $y_i - \hat{y}_i$: prediction error or residual. How much above/below the least squared line the actual y-value is.
- b_1: slope is the predicted change in y for one-unit increase in x.
  - Always meaningful
- b_0: y-intercept: predicted value of y when x = 0
  - Only meaningful if $x = 0$ is a plausible data value near/in the range of observed x-values.