Correlation

- Preliminary tool before regression analysis
- Understanding the relationship between variables helps us:
  - Identify potential predictor variables
  - Understand the strength and direction of relationships
  - Detect non-linear patterns
- Correlation measures association, not causation.
  - A high correlation between two variables does not imply that one causes the other.

Sample means, variance and covariance

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Scatter plots

- Before computing numerical measures, always create scatter plots

Pattern:

- Positive trend: Points slope upward from left to right
- Negative trend: Points slope downward from left to right
- No correlation: Points scattered randomly with no clear pattern

Visual inspection can reveal:

- Linear vs non-linear relationships
- Outliers that might affect correlation measures

Pearson correlation coefficient

- Ranges from -1 to 1
- Perfect positive relationship -> r = +1
- Perfect negative relationship -> r = -1
- No linear relationship -> r = 0

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

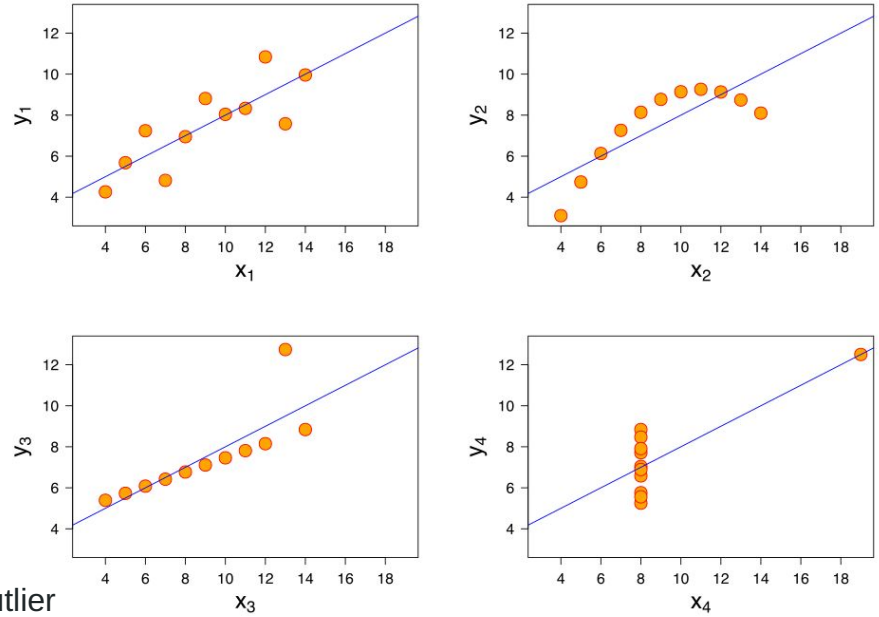
What will Pearson correlation coefficient give for the following non-linear relationship?

- Sinusoidal relationship
- Quadratic relationship

Anscombe's dataset

Each dataset contains exactly 11 data points and has:

- Identical mean of x: $\bar{x} = 9.0$
- Identical mean of y: $\bar{y} = 7.5$
- Identical correlation coefficient: r = 0.816
- Identical linear regression equation: y = 3.0 + 0.5x
- Identical R2: 0.67



Dataset Characteristics:

1. Dataset I: True linear relationship with normal scatter
2. Dataset II: Clear non-linear (quadratic) relationship
3. Dataset III: Perfect linear relationship except for one extreme outlier
4. Dataset IV: No relationship between x and y; all x values except one are identical

Statistical measures alone are insufficient. Always examine data visually through plots before and after analysis.

Spearman’s rank correlation

- Ordinal variables
- Non-linear relation
- Example

Kendall's rank correlation

- **Concordant pair:**
  For observations (xi, yi) and (xj , yj), they are concordant if:
  - (xi − xj)(yi − yj) > 0
  - Both variables change in the same direction

- **Discordant pair**
  They are discordant if:
  - (xi − xj)(yi − yj) < 0
  - Variables change in opposite directions

$$\tau = \frac{C - D}{\binom{n}{2}} = \frac{C - D}{\frac{n(n-1)}{2}}$$

Where:

- C = number of concordant pairs
- D = number of discordant pairs
- $\binom{n}{2} = \frac{n(n-1)}{2}$ = total number of pairs

- $\tau = +1$: All pairs are concordant (perfect positive association)
- $\tau = -1$: All pairs are discordant (perfect negative association)
- $\tau = 0$: Equal numbers of concordant and discordant pairs

Regression

Dependent Variable:The variable we want to predict or explain

- Denoted by y
- Contains the outcome of interest
- Examples: Sales, strength, profit, temperature

Independent Variable: Variables used to predict the dependent variable

- Denoted by x (single) or x1, x2, . . . , xp (multiple)
- Should be controllable or easily measurable
- Examples: Price, process conditions, design parameters

Regression

Different cases:

- Univariate: One dependent, one independent variable
- Multivariate: Multiple dependent and independent variables

- Linear: Relationship is a straight line
- Non-linear: Curved, exponential, logarithmic, etc.

Regression

The fundamental model for simple linear regression:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

Where:

- y_i = observed value of dependent variable for i-th observation
- x_i = value of independent variable for i-th observation
- β_0 = intercept parameter (y-value when $x = 0$)
- β_1 = slope parameter (change in y per unit change in x)
- ε_i = random error term for i-th observation
- n = total number of observations

Regression - Geometric interpretation

- The relationship between x and y is a straight line
- β_0 = y-intercept (where line crosses y-axis)
- β_1 = slope (rise over run)
- Each observation deviates from the line by ε_i

The error term accounts for model inadequacy: Linear model may be an approximation.

Regression - Ordinary least squares

- x_i values are assumed to be measured exactly
- No random error in the independent variable
- $E[\varepsilon_i] = 0$ (errors have zero mean)
- $\text{Var}(\varepsilon_i) = \sigma^2$ (constant variance - homoscedasticity)
- $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$ (independence)
- $\varepsilon_i \sim N(0, \sigma^2)$ (normality - for inference)

Regression - Ordinary least squares

We want to find estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize the sum of squared errors:

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Why squared errors?

- Treats positive and negative errors equally
- Penalizes large errors more heavily
- Mathematically tractable (differentiable)

Regression - Ordinary least squares

To minimize SSE, take partial derivatives and set equal to zero:

$$\frac{\partial \text{SSE}}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial \text{SSE}}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0$$

This gives us the normal equations:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Regression - Ordinary least squares

Solving the normal equations yields:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Alternative computational forms:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Regression - through origin

When we know a priori that the regression line passes through the origin ($\beta_0 = 0$):

Model: $y_i = \beta_1 x_i + \varepsilon_i$

Estimator:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

Regression - Fitted values and residuals

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

These represent the model's prediction for each observation.

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

Residuals are the differences between observed and predicted values.

Regression - Model quality

Coefficient of determination

R^2 measures the proportion of total variation in y explained by the regression model:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}}$$

Where:

$$\text{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(Total Sum of Squares)}$$

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{(Sum of Squared Errors)}$$

$$\text{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \quad \text{(Sum of Squares Regression)}$$

Regression - Model quality

Coefficient of determination

Fundamental relationship: $\text{SST} = \text{SSR} + \text{SSE}$
Interpretation:

- $R^2 = 0$: Model explains no variation (horizontal line at $\bar{y}$)
- $R^2 = 1$: Model explains all variation (perfect fit)
- $0 < R^2 < 1$: Proportion of variation explained
- Example: $R^2 = 0.85$ means 85% of variation in y is explained by x

Regression - Model quality

Coefficient of determination

Adjusted R^2 penalizes for the number of parameters:

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}/(n-2)}{\text{SST}/(n-1)} = 1 - (1 - R^2)\frac{n-1}{n-2}$$

Why adjust?

- R^2 always increases when adding variables
- Adjusted R^2 can decrease if added variable doesn't improve fit sufficiently

Multiple linear regression

Multiple linear regression extends simple linear regression to include several predictor variables

Multiple linear regression

The general multiple linear regression model with p predictor variables:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

For $i = 1, 2, \ldots, n$ observations.

Parameters:

- β_0: Intercept (expected value of y when all $x_j = 0$)
- β_j: Partial regression coefficient for x_j (effect of x_j holding others constant)
- ε_i: Random error term

Multiple linear regression - Matrix formulation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Minimizing $\boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ leads to:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

These are the normal equations - a system of p linear equations in p unknowns.

Multiple linear regression - Matrix formulation

If $\mathbf{X}^T\mathbf{X}$ is invertible (full rank condition):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_j \bar{x}_j$$