

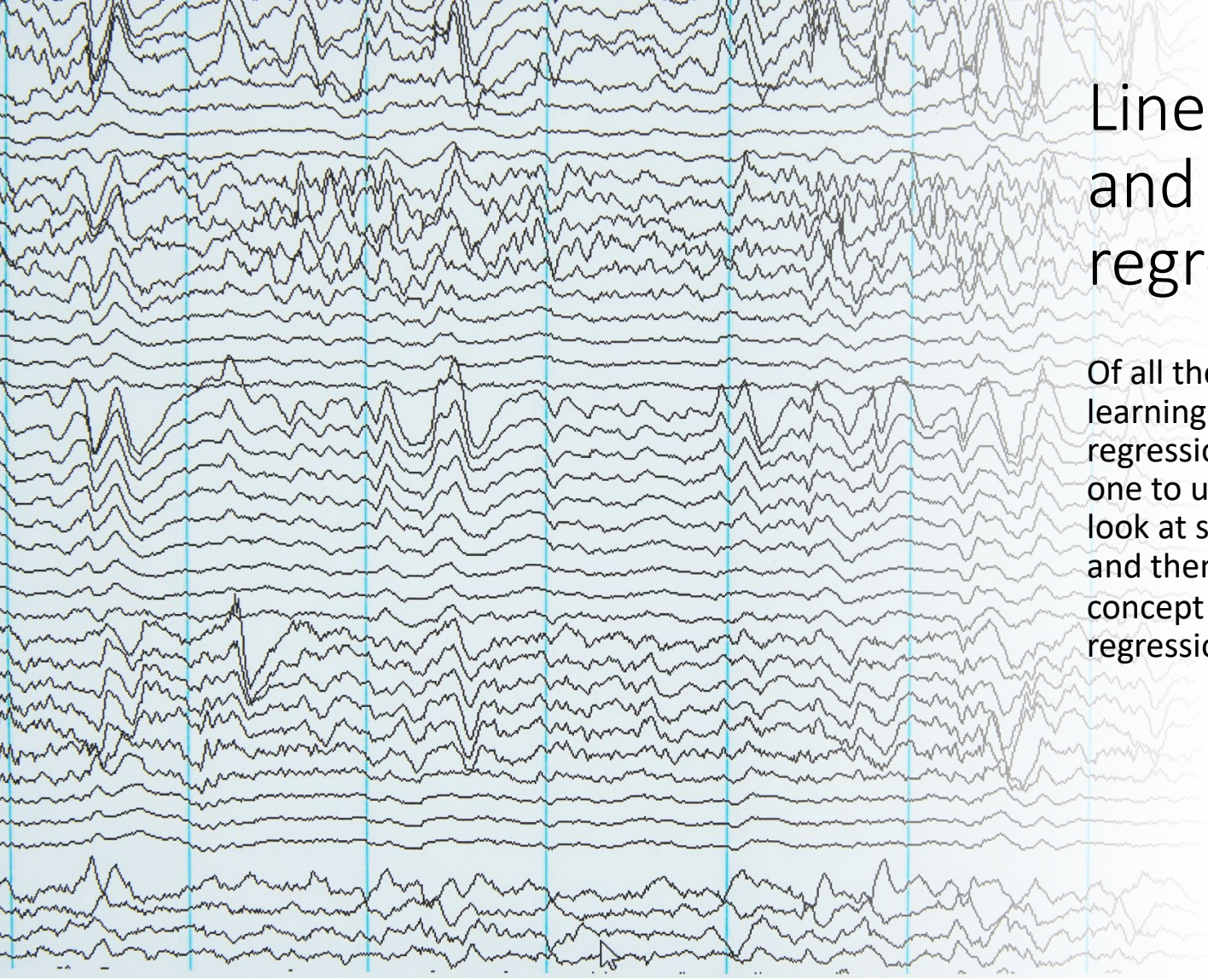


Machine Learning

Linear models

Phd. César Astudillo | Facultad de Ingeniería

Linear correlation and linear regression

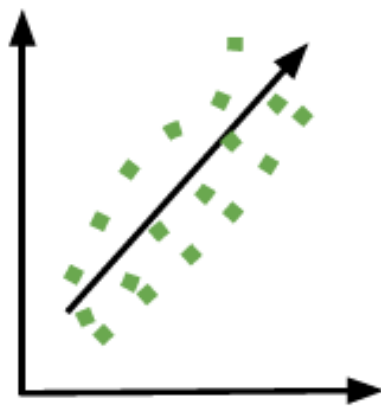


Linear correlation and linear regression

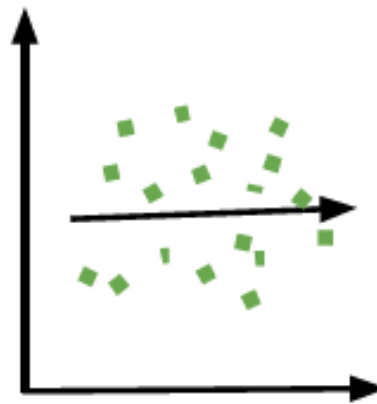
Of all the supervised machine learning techniques, the linear regression algorithm is the easiest one to understand. We will first look at simple linear regression and then we will expand the concept to multiple linear regression.

Linear Correlation: Intuition

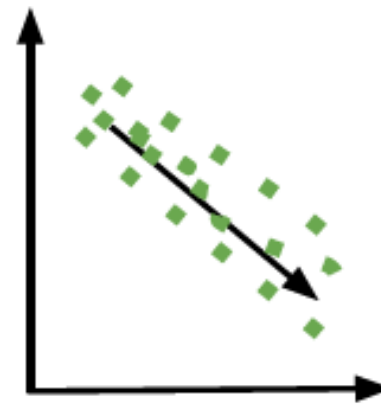
Linear Correlation



Positive
Correlation



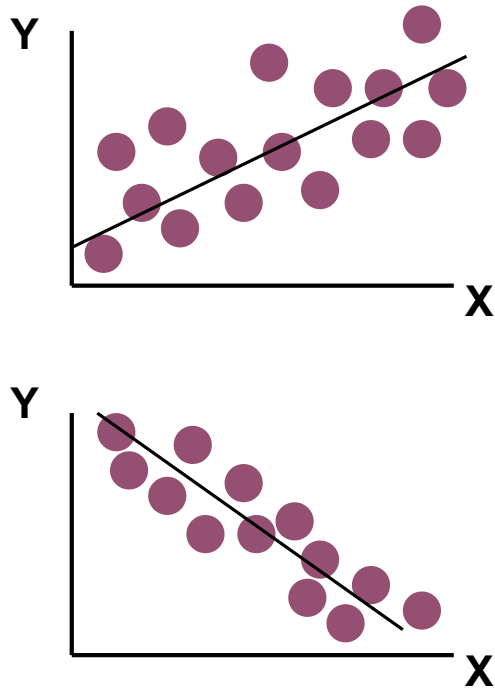
Zero
Correlation



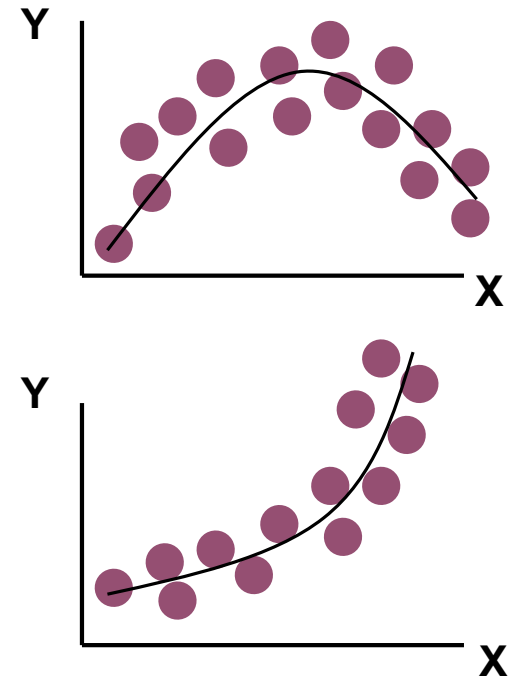
Negative
Correlation

Linear Correlation

Linear relationships

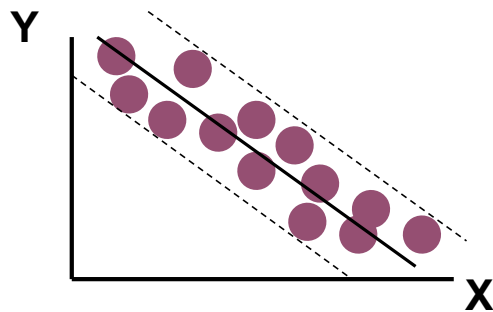
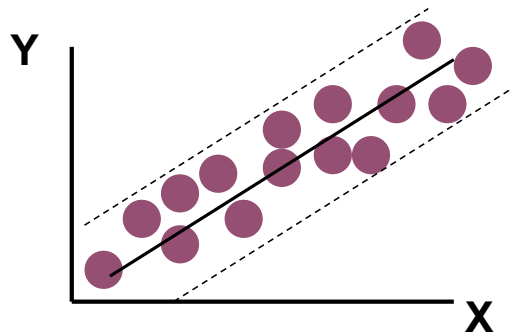


Curvilinear relationships

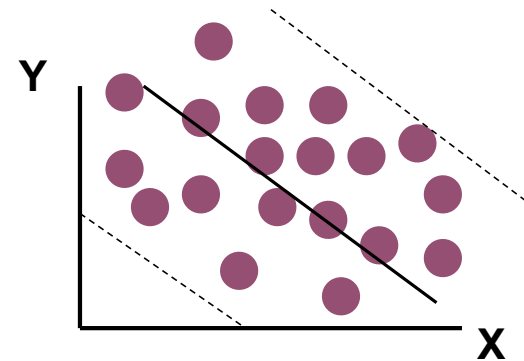
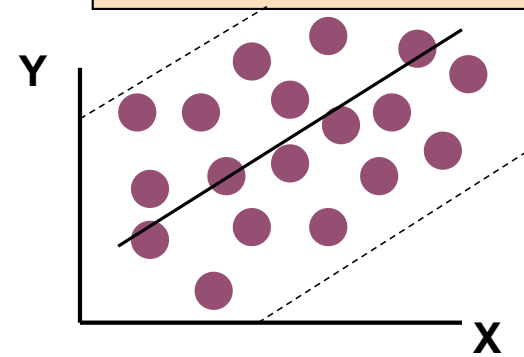


Linear Correlation

Strong relationships

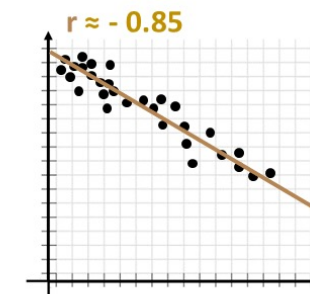
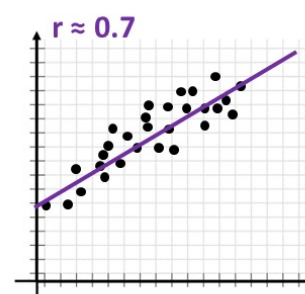
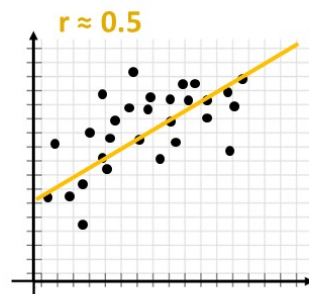
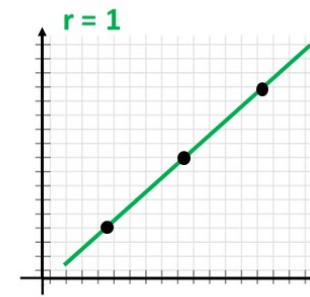
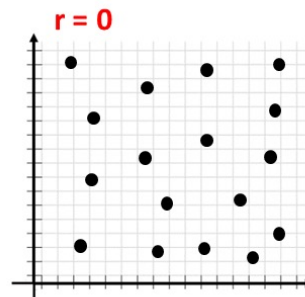
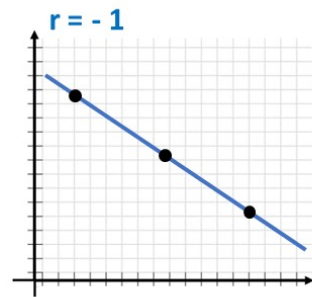


Weak relationships



Pearson's Correlation

Pearson's Correlation

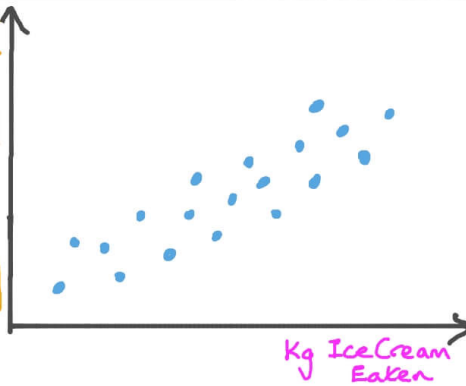


PEARSON'S CORRELATION COEFFICIENT

LOOK!
EATING
ICE CREAM
CAUSES
SUN
BURN!



Sun
Burn
Level



NO!
CORRELATION
DOES NOT
MEAN
CAUSATION!

$$r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\sqrt{\sum x^2 - n\bar{x}^2} \sqrt{\sum y^2 - n\bar{y}^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Population vs Sample

Population vs sample

A **population** is the entire group that you want to draw co

A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.

In research, a population doesn't always refer to people. It can mean a group containing elements of anything you want to study, such as objects, events, organizations, countries, species, organisms, etc.

Population Correlation Coefficient

$$P_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right) \left(\sum (y_i - \bar{y})^2\right)}}$$

Where, $\sigma_x, \sigma_y \rightarrow$ Population Standard Deviation

$\sigma_{xy} \rightarrow$ Population Covariance

$\bar{X}, \bar{Y} \rightarrow$ Population Mean

Sample Correlation Coefficient

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left(\sum (x_i - \bar{x})^2\right) \left(\sum (y_i - \bar{y})^2\right)}}$$

Where,

- $S_x, S_y \rightarrow$ Sample Standard Deviation
- $S_{xy} \rightarrow$ Sample Covariance
- $\bar{X}, \bar{Y} \rightarrow$ Sample Mean

Correlation



Measures the relative strength of the *linear* relationship between two variables



Unit-less



Ranges between -1 and 1



The closer to -1 , the stronger the negative linear relationship

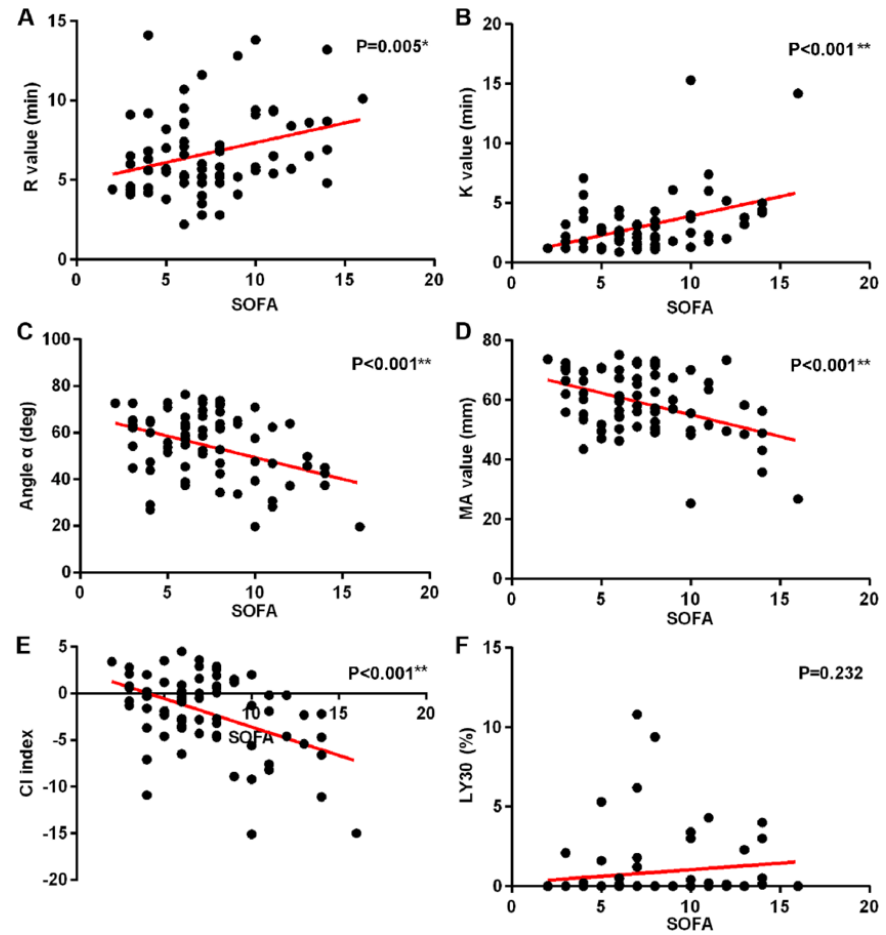


The closer to 1 , the stronger the positive linear relationship

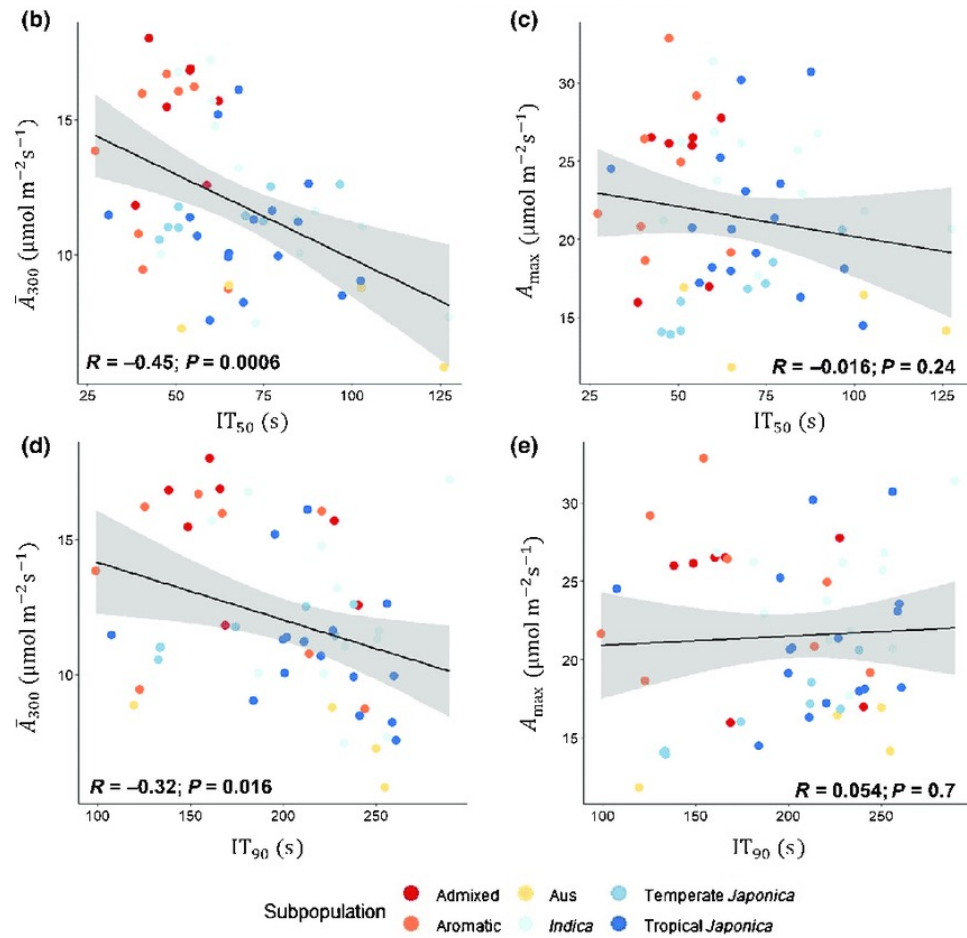


The closer to 0 , the weaker any positive linear relationship

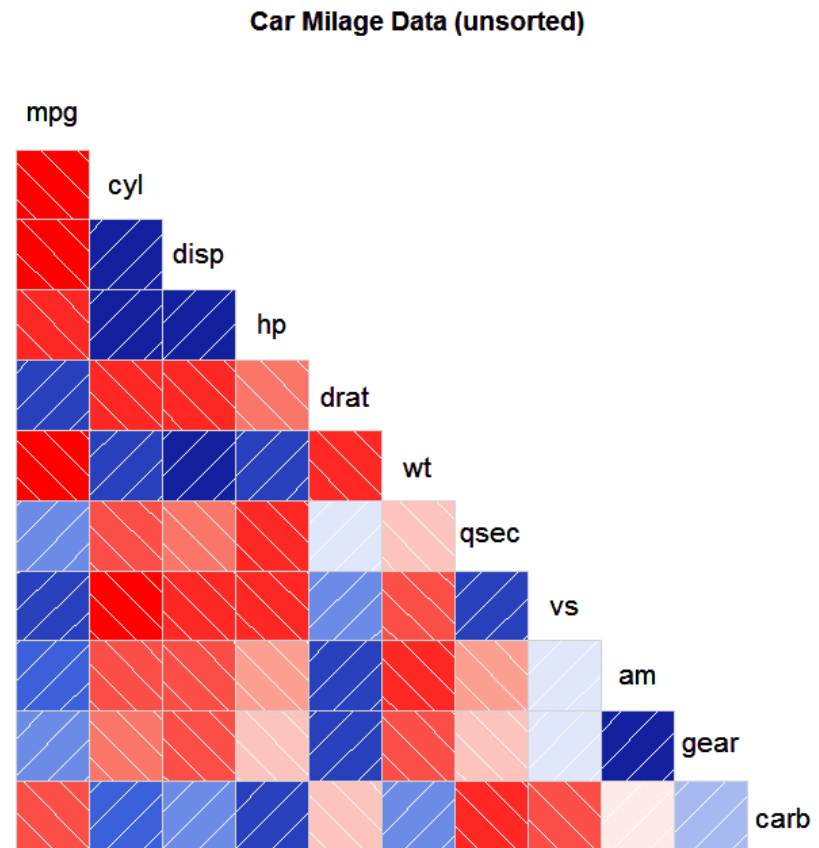
Linear correlation and scatter plots



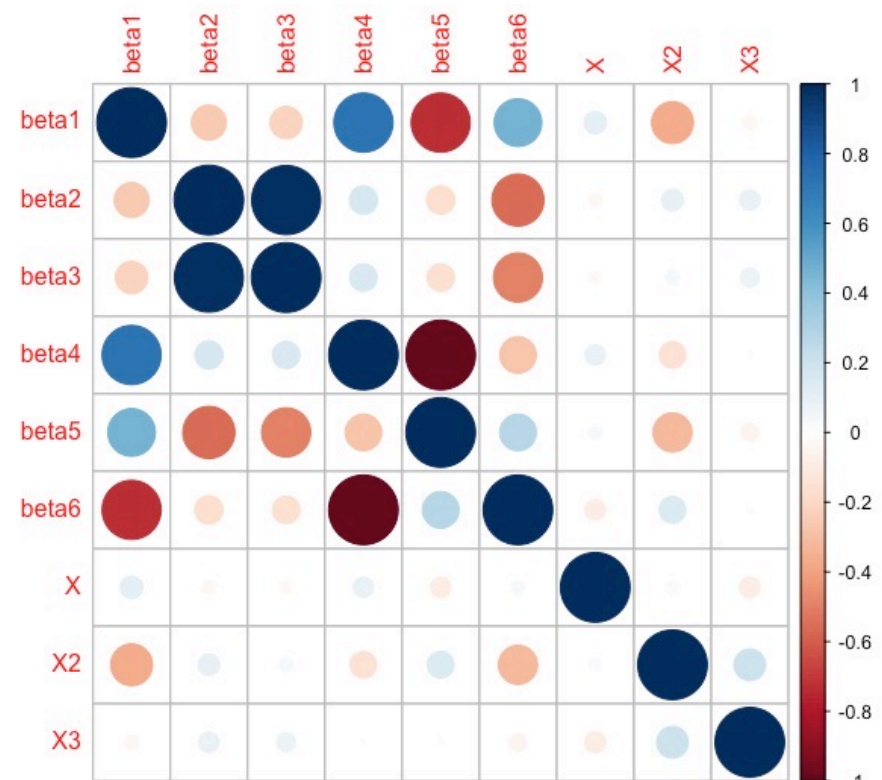
Linear correlation and scatter plots



Correlation matrix plots

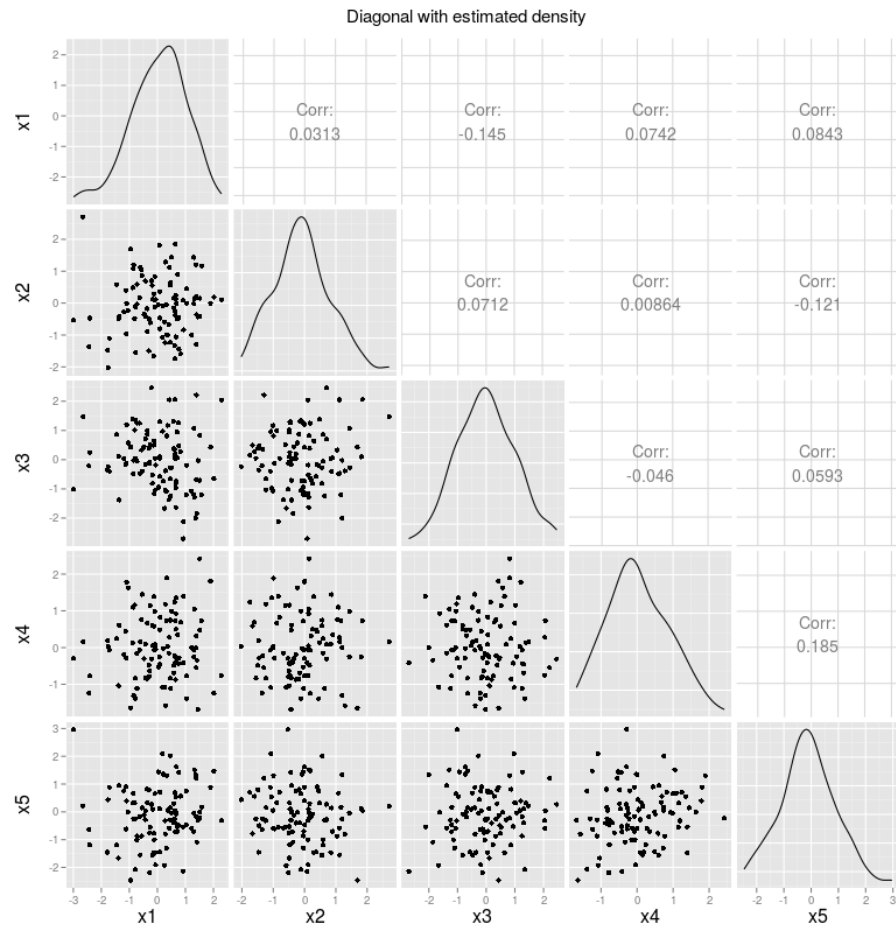


Correlation matrix plots



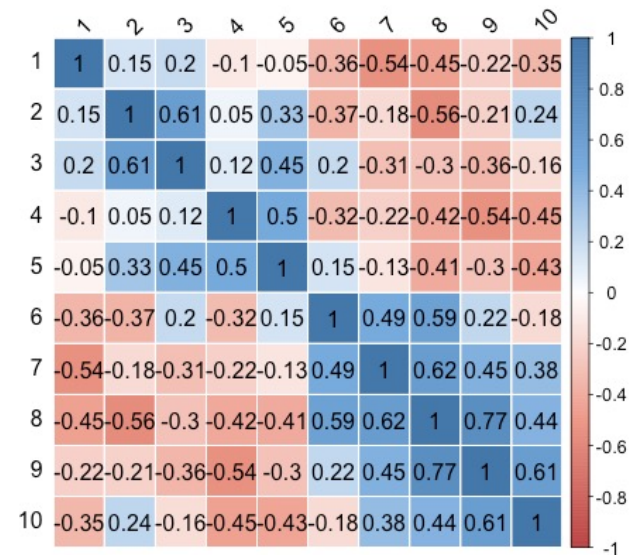
[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Correlation matrix plots



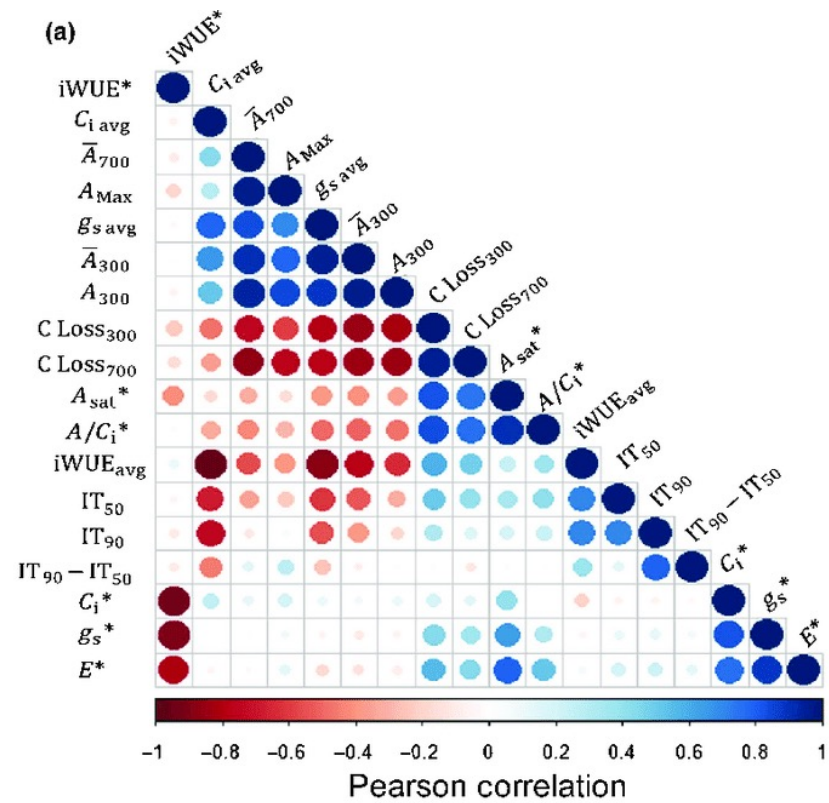
[This Photo](#) by Unknown Author is licensed under [CC BY-SA](#)

Correlation matrix plots



[This Photo](#) by Unknown Author is licensed under [CC BY](#)

Correlation Matrix Plots



Covariance vs Correlation

Definitions

- Variance: Measures the spread of a single variable around its mean.
- Correlation: Measures the linear relationship between two variables.

Formulas

- Variance Formula:
- $\text{Var}(X) = \sum (x_i - \mu)^2 / n$
- Correlation Formula (Pearson):
- $\rho(X, Y) = \text{Cov}(X, Y) / \sqrt{\text{Var}(X)\text{Var}(Y)}$

Key Differences

- Focus:
 - Variance: Spread of a single variable
 - Correlation: Relationship between two variables
- Units:
 - Variance: Squared units of the variable
 - Correlation: Unitless (normalized measure)
- Range:
 - Variance: 0 to ∞
 - Correlation: -1 to 1

Use in Machine Learning

Variance:

- Feature Selection
- Regularization (L2)
- Principal Component Analysis (PCA)

Correlation:

- Feature Selection
- Dimensionality Reduction (PCA)
- Interpretability

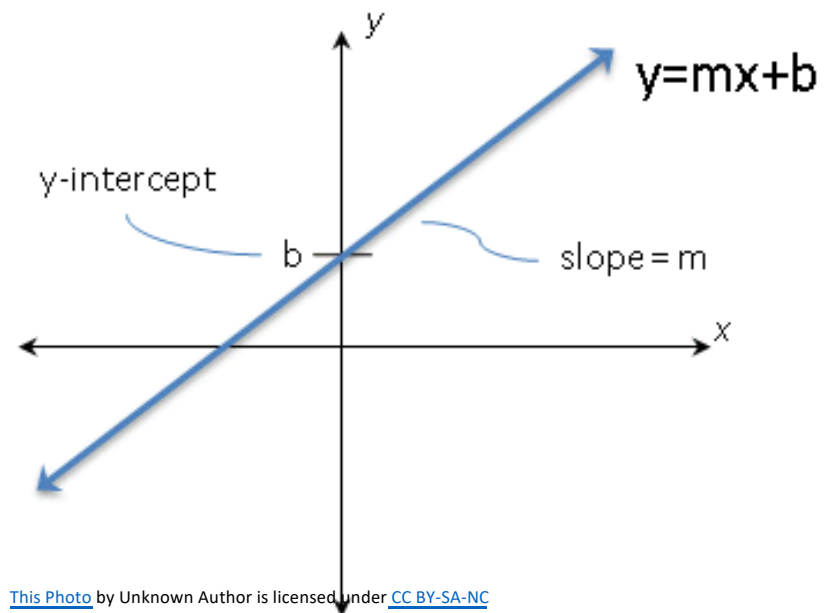
Linear Regression

What is Linear Regression?

- Linear regression is a statistical method used to model the relationship between a dependent variable (y) and one or more independent variables (x). It is often used for predicting outcomes based on patterns in data.

What is Linear Regression?

- Linear regression models the relationship between a dependent variable (y) and an independent variable (x).
- It is represented by the equation:
- $y = m * x + b$



What is m (Slope)?

- m represents the slope of the line.
- It indicates the rate of change of y with respect to x .
- Interpretation: For every unit increase in x , y changes by m units.

What is b (Intercept)?

- b represents the y -intercept of the line.
- It indicates where the line crosses the y -axis when $x = 0$.
- Interpretation: It is the value of y when there is no x .

Visualizing m and b

- A positive slope ($m > 0$) means the line rises as x increases.
- A negative slope ($m < 0$) means the line falls as x increases.
- A larger magnitude of m indicates a steeper line.
- b shifts the entire line up or down.

Example: Predicting House Prices

$$y = m * x + b$$

- y: Predicted house price
- x: Square footage of the house
- m: Increase in price per square foot
- b: Base price when square footage is zero

Linear Regression Formula

The formula for simple linear regression is:

$$y = m * x + b$$

where:

- y: Predicted value
- m: Slope (change in y per unit increase in x)
- x: Independent variable
- b: Intercept (value of y when x = 0)

How Does Prediction Work?

- Train the model on historical data by finding the best values for m and b .
- Use the learned equation to predict y for new values of x .
- Example: Predicting house prices based on square footage.

Steps for Using Linear Regression

1. Collect Data: Gather historical data with known outcomes.
2. Train the Model: Fit the line to minimize error (using methods like Ordinary Least Squares).
3. Evaluate Performance: Use metrics like R^2 , RMSE.
4. Make Predictions: Apply the model to new data points.
5. Adjust and Improve: Refine the model if needed.

Example: House Price Prediction

- Independent Variable (x): Square footage of a house
- Dependent Variable (y): Predicted price of the house
- If $m = 300$ and $b = 50,000$:

$$\text{Predicted Price} = 300 * (\text{Square Footage}) + 50,000$$

Advantages of Linear Regression

- Simple and interpretable.
- Works well for linear relationships.
- Requires minimal data preparation.
- Provides a baseline for more complex models.

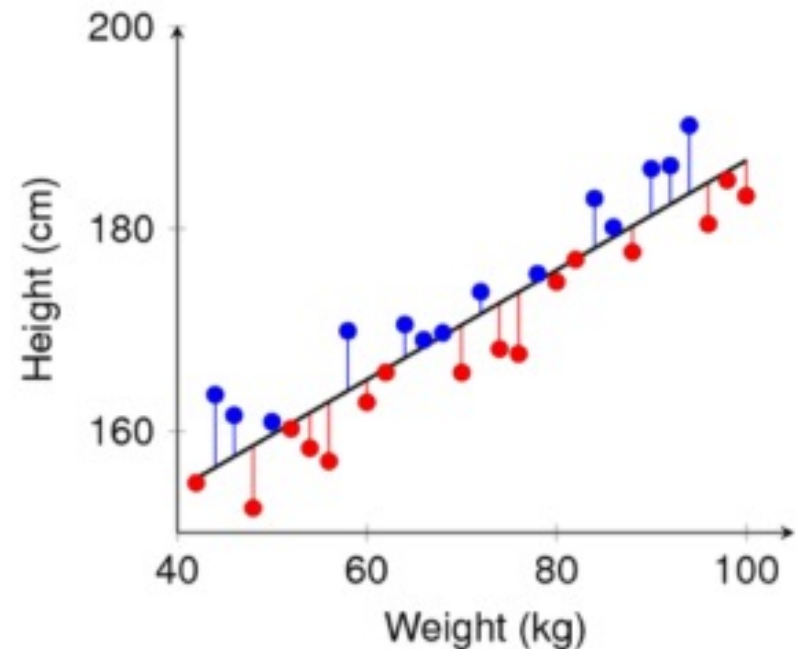
Limitations of Linear Regression

- Assumes linear relationships only.
- Sensitive to outliers.
- Poor performance with highly complex data.
- Assumes independence between features.

Residuals

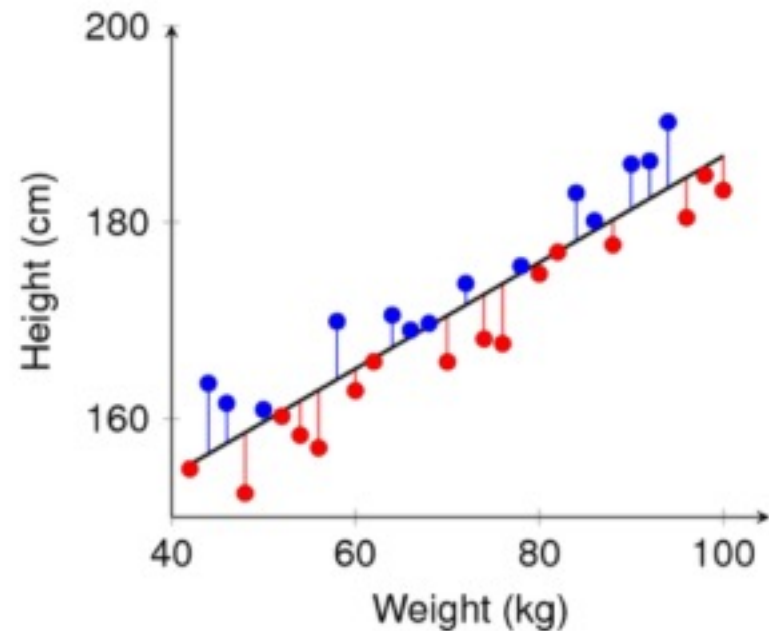
What is a Residual?

- A residual is the difference between the actual value and the predicted value by the model.
- Formula: $\text{Residual} = y_{\text{actual}} - y_{\text{predicted}}$
- Residuals help us understand how well a model fits the data.



Visualizing Residuals

- Residuals are the vertical distances between the actual data points and the regression line.
- Smaller residuals indicate a better fit.
- Residuals can be positive or negative depending on whether the predicted value is lower or higher than the actual.



Why Residuals Matter?

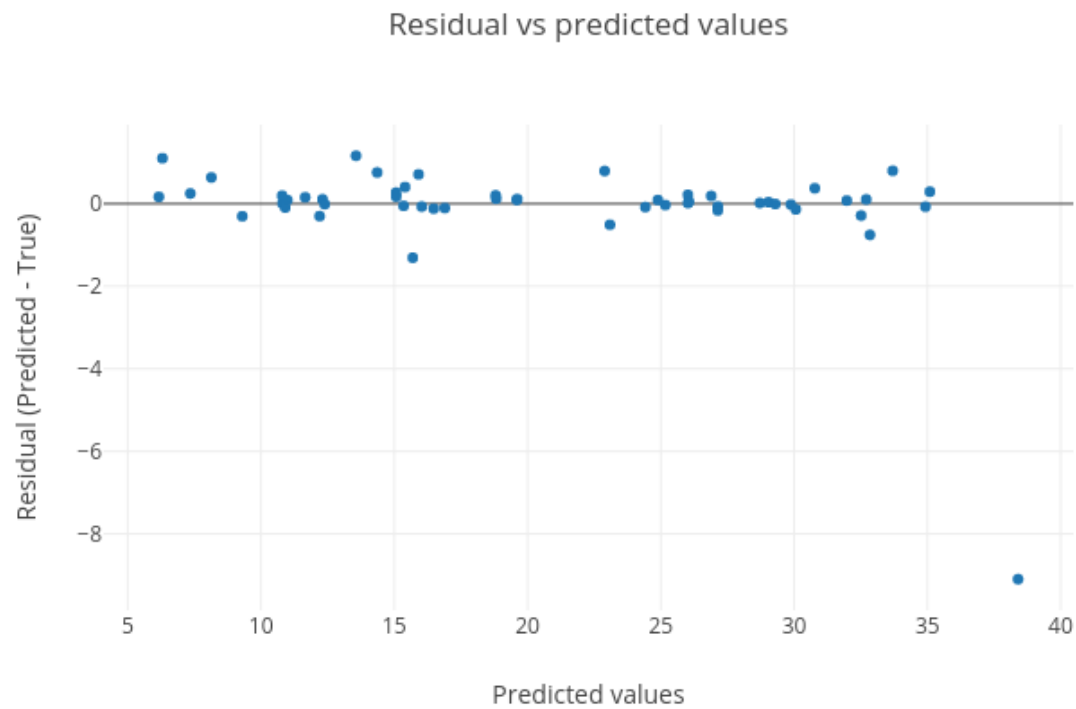
- Model Fit: Large residuals indicate poor model fit.
- Error Analysis: Consistent patterns in residuals indicate model bias.
- Assumptions Check: Helps verify assumptions of linear regression, such as linearity and homoscedasticity.

Residual Plot for Analysis

A residual plot displays residuals on the y-axis and predicted values on the x-axis.

- Good Fit: Residuals are randomly scattered around zero.
- Poor Fit: Clear patterns or non-random structure in the residual plot.
- Outliers: Large deviations from zero indicate possible outliers.

Example of residual plot

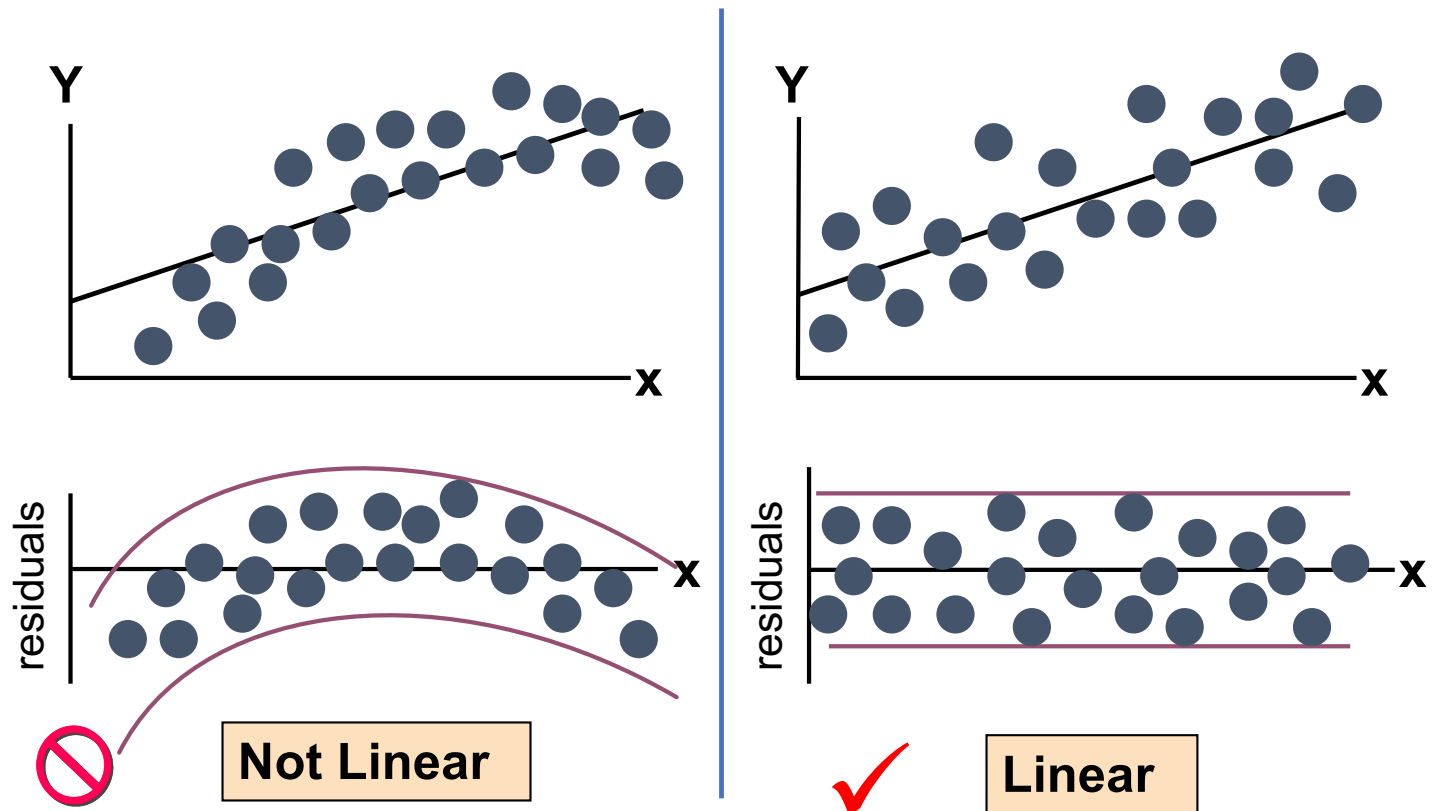


[This Photo](#) by Unknown Author is licensed under [CC BY](#)

Steps for Residual Analysis

- 1. Plot Residuals: Create a scatter plot of residuals vs. predicted values.
- 2. Check Randomness: Residuals should be randomly distributed.
- 3. Check Magnitude: Residuals should be small.
- 4. Identify Outliers: Large residuals may indicate influential points.
- 5. Test for Normality: Residuals should be approximately normally distributed.

Residual Analysis for Linearity

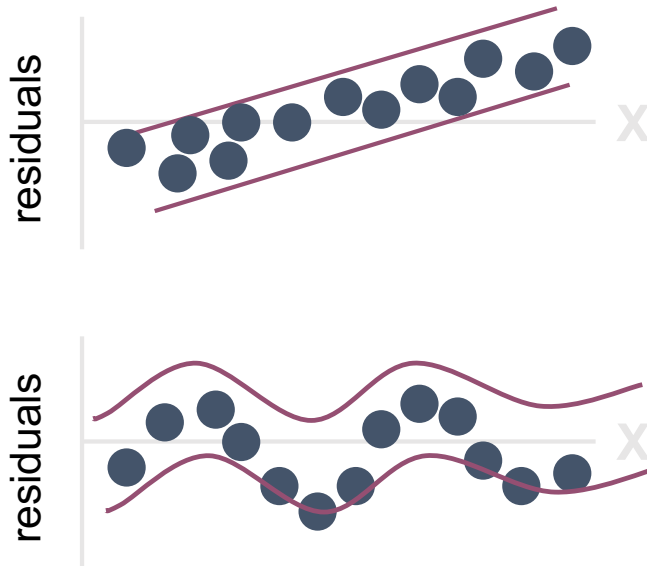


■ Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

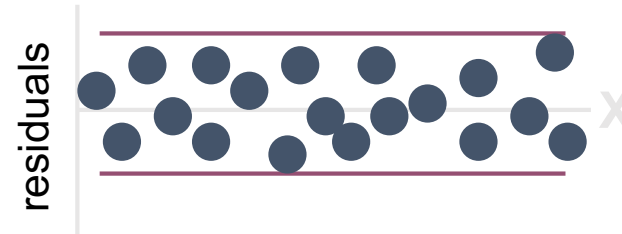
Residual Analysis for Independence



Not Independent



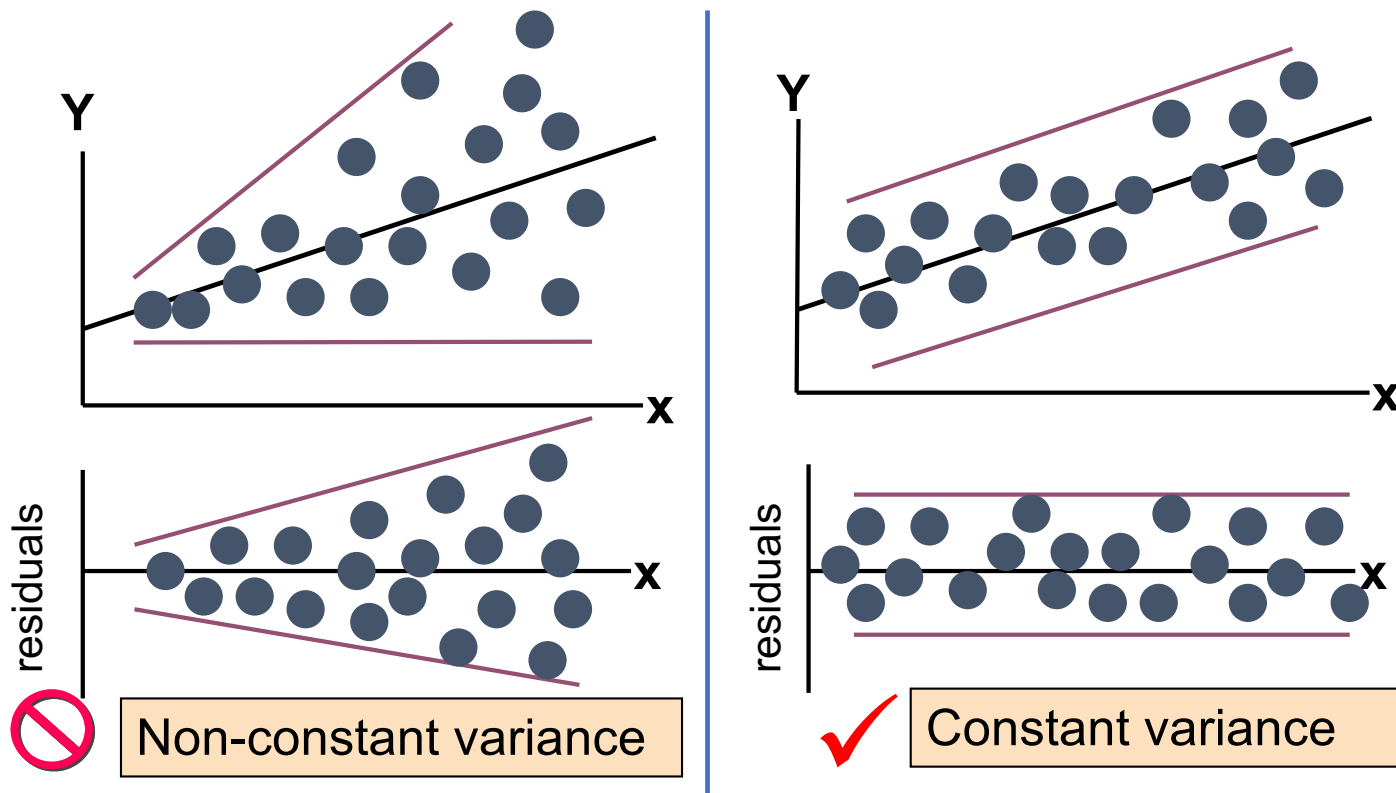
Independent



Homoscedasticity

- Homoscedasticity, or homogeneity of variances, is **an assumption of equal or similar variances in different groups being compared**. This is an important assumption of parametric statistical tests because they are sensitive to any dissimilarities. Uneven variances in samples result in biased and skewed test results.

Residual Analysis for Homoscedasticity



■ Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Multiple Linear Regression

- More than one predictor...

$$E(y) = \alpha + \beta_1 * X + \beta_2 * W + \beta_3 * Z \dots$$

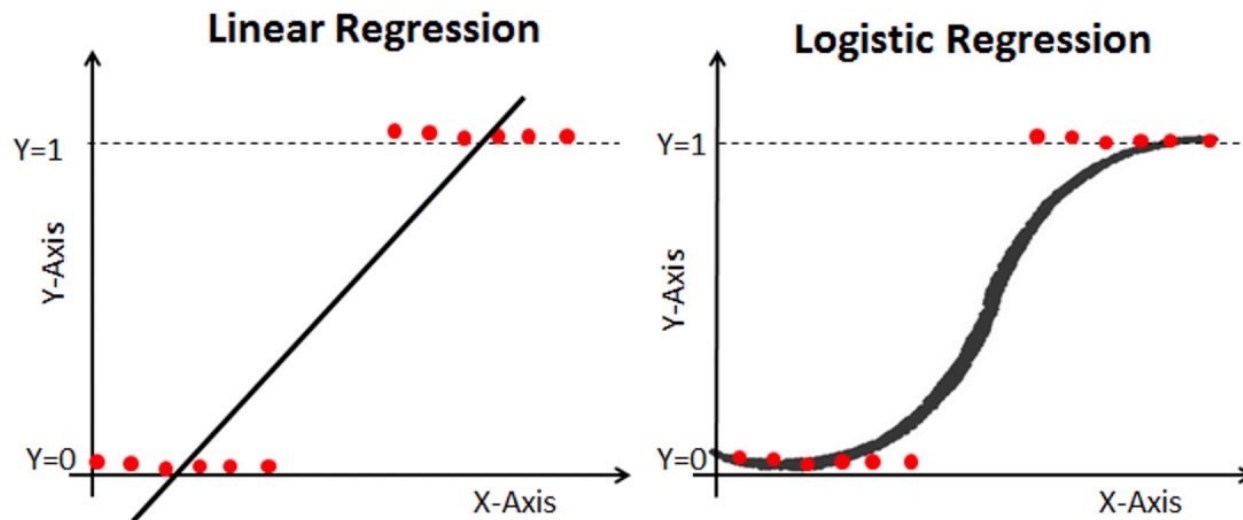
Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

Understanding Logistic Regression

What is Logistic Regression?

- Logistic Regression is a statistical method used for binary classification.
- It models the probability of an outcome based on one or more input variables.
- Unlike linear regression, it predicts probabilities using a logistic (S-shaped) curve.

Linear vs logistic regression



Source : <https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python>

Logistic Regression Equation

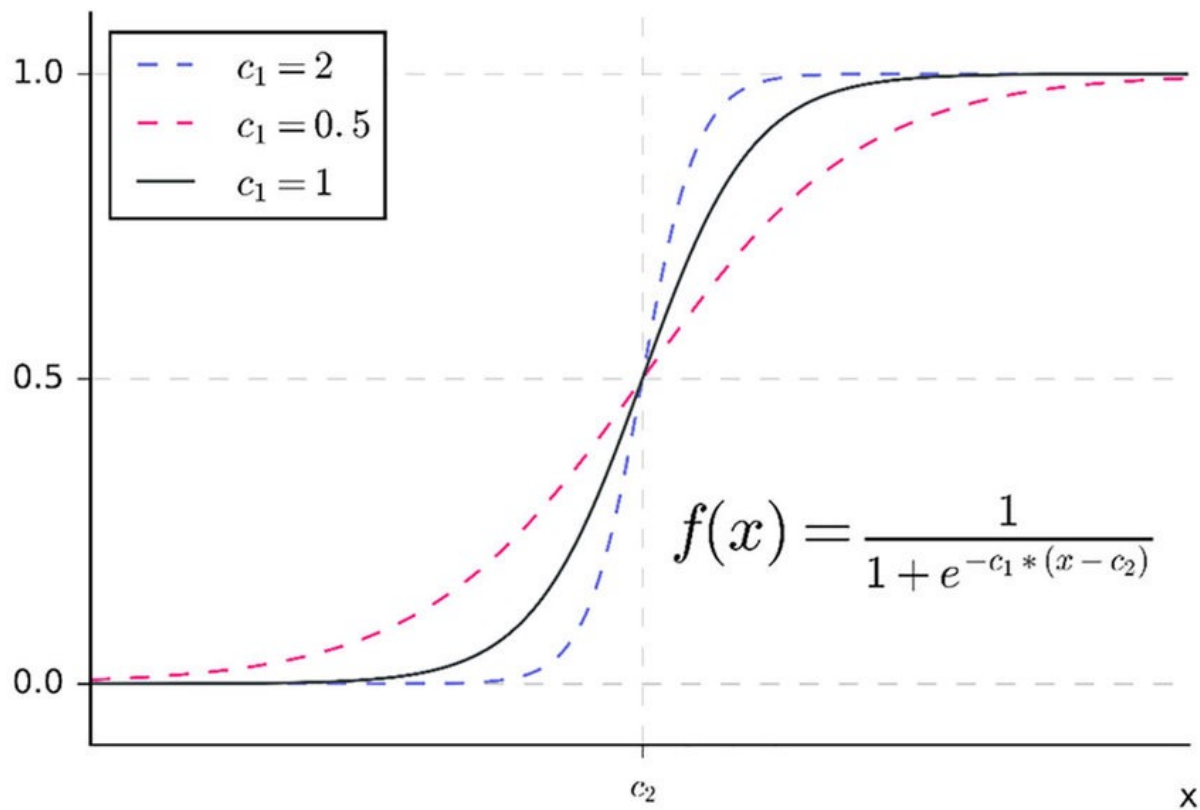
The logistic regression equation is:

$$P(y=1 | X) = 1 / (1 + e^{(-z)})$$

where $z = mX + b$

- $P(y=1 | X)$: Probability of the positive class
- e : Euler's number
- m and b : Model parameters

The sigmoid function



The Sigmoid Function

- The sigmoid function transforms the linear output into a probability between 0 and 1.
- Output close to 1 indicates the positive class, while close to 0 indicates the negative class.

Decision Boundary

- A decision boundary separates the two classes in logistic regression.
- If the predicted probability is greater than 0.5, the point is classified as positive.
- If less than 0.5, it is classified as negative.

When to Use Logistic Regression?

- Binary classification tasks (e.g., spam detection, disease prediction).
- When the relationship between variables is linear in log-odds space.
- When interpretability is important.

Limitations of Logistic Regression

- Struggles with non-linear relationships.
- Assumes independence between features.
- Sensitive to outliers.
- Not ideal for complex datasets with many features.

Shortcomings of Linear Models

1. Assumption of Linearity

- Linear models assume a linear relationship between input features and the output.
- Real-world data often exhibits non-linear patterns, leading to poor performance.
- Example: Modeling complex biological data or financial trends.

2. Sensitivity to Outliers

- Linear models are highly sensitive to outliers, which can distort the model fit.
- A single outlier can significantly affect the slope and intercept.
- Outliers can lead to misleading predictions and poor generalization.

3. Feature Independence Assumption

- Linear models assume features are independent of each other.
- Real datasets often contain correlated features (multicollinearity).
- Multicollinearity can destabilize the coefficients and reduce interpretability.

4. Limited Expressiveness

- Linear models struggle to capture complex relationships in data.
- Cannot model interactions between features effectively.
- Poor performance on datasets with complex structures like image or text data.

5. Overfitting and Underfitting

- Linear models can underfit the data when the relationship is complex.
- Overfitting can occur when the model is too flexible or when using polynomial features.
- Regularization techniques (Ridge, Lasso) can help but have limits.



Machine Learning

Linear models

Phd. César Astudillo | Facultad de Ingeniería