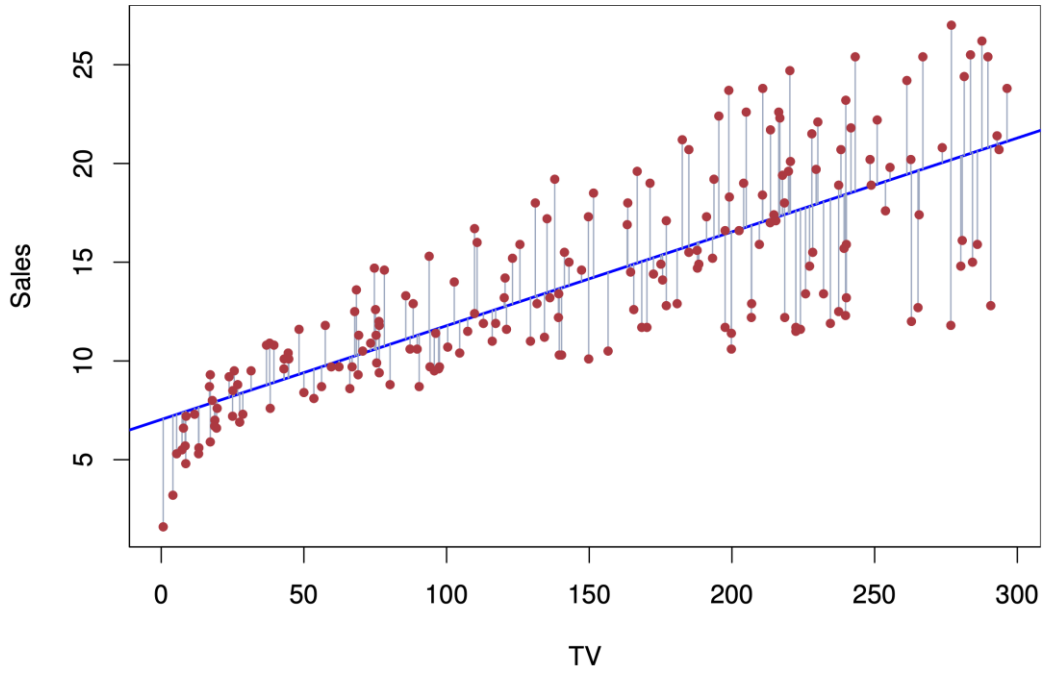
Linear Regression

Linear Regression

- predicting a numerical variable based on another numerical variable.
- This type of statistical analysis is used when both variables are quantitative (unlike mean comparison tests where the independent variable is categorical).
- Before applying linear regression, it is essential to check whether a linear relationship exists between the variables. To do this, we start with the Pearson correlation test, or use Spearman correlation if certain assumptions are not met.



Linear Regression

The Role of Correlation:

The Pearson correlation coefficient (r) not only helps check for a linear relationship but also quantifies its strength and direction. The value of r ranges from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship).

A value close to 0 suggests no linear relationship, which would make linear regression an inappropriate model.

Why Check for Linearity?

Linear regression fits a straight line (the "line of best fit") to the data. If the true underlying relationship is curved (e.g., exponential, parabolic), a straight line will give poor and misleading predictions. A scatter plot is the best tool for this initial check.

Linear Regression

The Role of Correlation:

The Pearson correlation coefficient (r) not only helps check for a linea relationship but also quantifies its strength and direction. The va ges from -1 (perfect negative linear relationship) to +1 (perfect ear relationship).

A value close to 0 suggests no linear relati would make linear regression an inappropriate model.

Why Check for Linearity?

Linear regressi ght line (the "line of best fit") to the data. If the true underlying p is curved (e.g., exponential, parabolic), a straight line will give poor and misleading predictions. A scatter plot is the best tool for this initial check.

Correlation does not imply Causation!!!

Spearman's Rank Correlation (ρ or r_s)

Spearman's correlation is a non-parametric alternative used when the assumptions for Pearson are violated. It's more robust and less restrictive.

How it works:

It converts the data into ranks (1st, 2nd, 3rd, etc.) and then calculates the correlation on those ranks.

When to use it:

When the relationship is monotonic but not linear (consistently increasing or decreasing, but at a changing rate).

When your data contains outliers.

When the data is ordinal (e.g., survey ranks: 1=Strongly Disagree, 5=Strongly Agree) or not normally distributed.

Pearson vs. Spearman

Feature	**Pearson Correlation**	**Spearman Correlation**
Data Type	Requires **interval or ratio** data (continuous numerical data).	Works well with **ordinal data** (ranked data) and continuous data.
Relationship Measured	Measures the strength and direction of a **linear** relationship.	Measures the strength and direction of a **monotonic** relationship (consistently increasing or decreasing, but not necessarily linear).
Assumptions	Has stricter assumptions: - Linear relationship - Variables are normally distributed - Homoscedasticity (constant variance of residuals)	A **non-parametric** test. Has fewer assumptions: - No assumption of normal distribution - No assumption of linearity
Sensitivity to Outliers	**Highly sensitive.** A single outlier can significantly distort the correlation coefficient.	**Robust (resistant).** Uses data ranks, so outliers have a much smaller impact.
Interpretation	Coefficient (r) indicates how well a straight line describes the relationship.	Coefficient (ρ or r_s) indicates how well a monotonic function (any consistent curve) describes the relationship.
Power	More powerful (better at detecting an effect) if all its strict assumptions are met.	Often more powerful when assumptions of Pearson correlation are violated (e.g., non-normal data, outliers).

Assumptions of Pearson VS Spearman Correlation:

Assumption	**Pearson Correlation**	**Spearman Correlation**
Variable Type	Both variables must be **continuous** (interval or ratio scale)	Variables can be **ordinal, interval, or ratio**
Relationship Type	Assumes a **linear relationship** between variables	Assumes a **monotonic relationship** (consistently increasing or decreasing)
Normality	Both variables should be **normally distributed**	**No normality assumption** required
Outliers	Highly sensitive to **outliers**	**Robust to outliers** (uses ranks instead of raw values)
Homoscedasticity	Assumes **constant variance** along the relationship	**No homoscedasticity assumption** required
Sample Size	Requires adequate sample size (typically ≥ 30)	Can work with smaller samples, but more reliable with larger n
Data Level	Requires precise numerical measurements	Works with ranks, so less sensitive to exact numerical values
Measurement Scale	Interval or ratio scale required	Ordinal, interval, or ratio scale acceptable

Autocorrelation

Autocorrelation occurs when the value of a variable at one time point is influenced by its own previous values. In longitudinal or time-series data, this means that measurements taken at different times from the same individual are not independent.

Example: if a patient has a blood pressure of 160 mmHg in the first month and it drops by 2 units each month, the value in the eighth month (146 mmHg) is clearly affected by the initial value. Now compare this with another patient who started at 140 mmHg and also dropped 2 units monthly — the overall trend is similar, but the dependency on the starting point creates autocorrelation.

This violates one of the key assumptions of classical linear regression — the independence of observations — and can lead to incorrect inferences if not addressed.

Solution: When autocorrelation is present, especially in repeated measures or time-series data, it's more appropriate to use models that account for these dependencies, such as Mixed Effects Models or Generalized Estimating Equations (GEE).

Key Assumptions of Linear Regression:

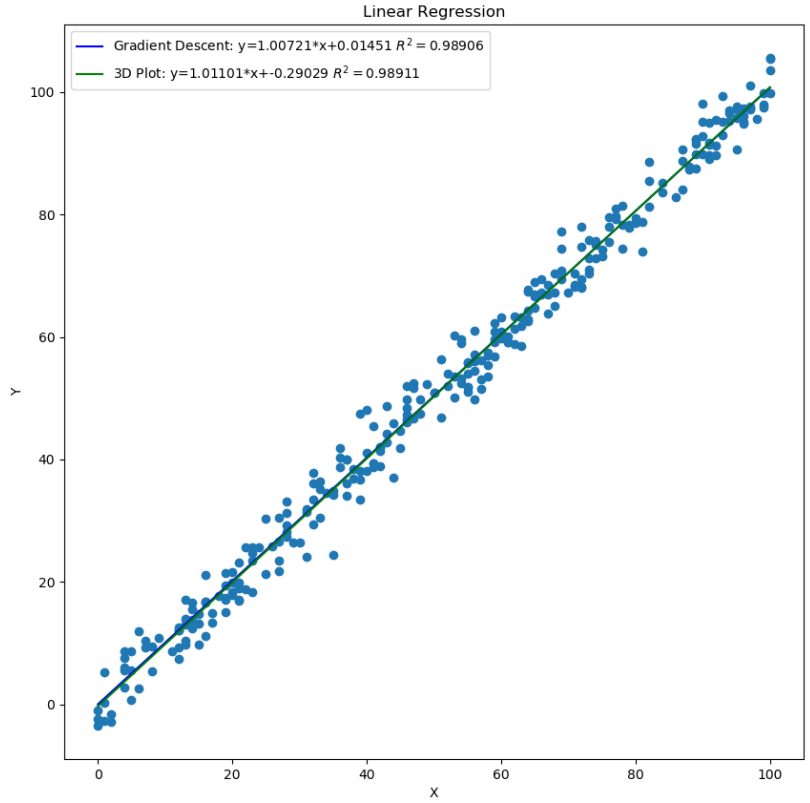
While your screenshot mentions checking for a linear relationship, linear regression has a few other important assumptions that should be verified after fitting the model to ensure the results are valid:

- Normality: The residuals (the differences between the observed values and the values predicted by the line) should be approximately normally distributed.

- Homoscedasticity: The variance of the residuals should be constant across all levels of the independent variable. In simpler terms, the data should be equally spread out around the regression line along its entire length.

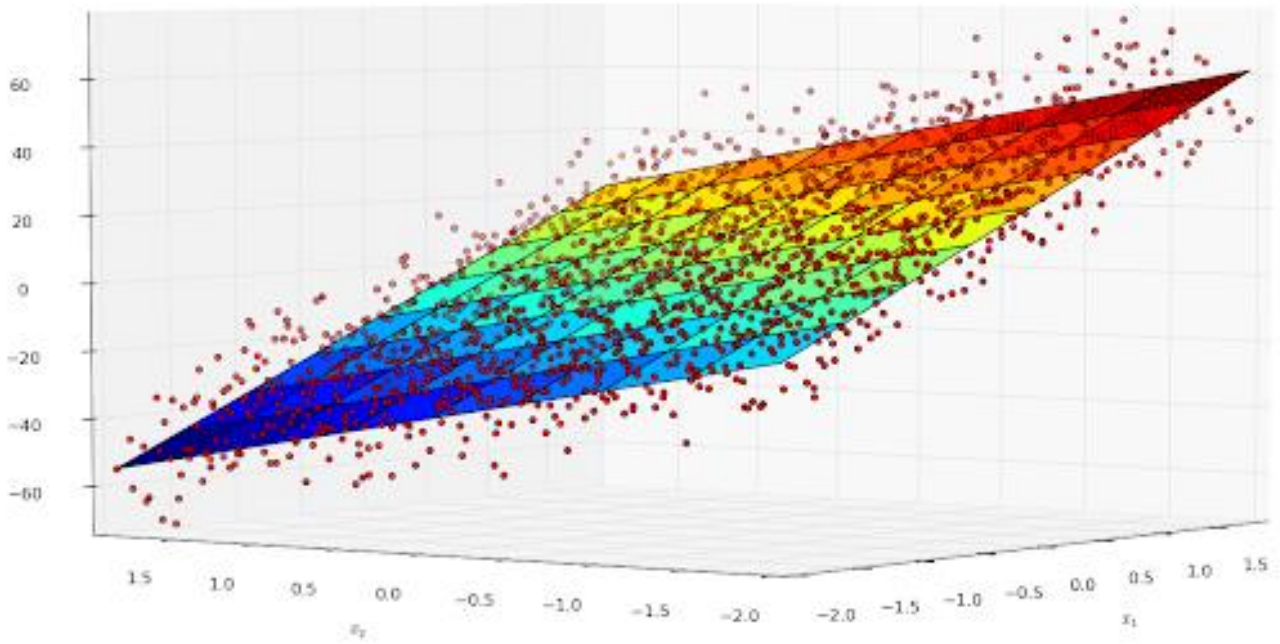
- Independence: The observations must be independent of each other.

Linear regression



$$y = \beta_0 + \beta_1 x + \varepsilon$$

Multiple linear regression



$$y = \beta_0 + \beta_1 x + \beta_2 x + \beta_3 x + \varepsilon$$