Table of Contents

# 2.2 Correlation & Regression with Data Cleaning

## Correlation

Correlation measures the strength and direction of a linear relationship between two variables. The **correlation coefficient** (r) ranges from –1 to +1: +1 indicates a perfect positive linear relationship, –1 a perfect negative linear relationship and 0 no linear relationship. Correlation treats both variables symmetrically and is unaffected by changes in scale【809275587000967†L341-L378】.

## Regression

Regression models how a dependent variable changes as one or more independent variables vary. It is **asymmetric**, designating a response and predictors. Regression coefficients quantify the effect of each predictor; they depend on units and scale【809275587000967†L341-L378】. Regression can be simple (one predictor) or multiple (several predictors).

## Data cleaning

Data cleaning ensures datasets are accurate and consistent before analysis. Key tasks include:

* **Removing duplicates** to avoid over‑representation.
* **Handling missing values** (see Unit 2.3) via deletion or imputation.
* **Correcting errors** such as typos or impossible values.
* **Standardising formats** (e.g., dates, units).
* **Identifying and treating outliers** using boxplots or z‑scores【444968039617338†L448-L474】.

Proper cleaning improves the validity of correlation and regression analyses.

## Example

An analyst studies the relationship between advertising spending and sales. After cleaning the data (removing duplicate records and correcting currency units), she computes the correlation coefficient (for example, r = 0.8) and fits a regression model to estimate sales for a given advertising budget.

## Summary

Correlation quantifies the association between two variables, while regression models how one variable predicts another. Both require clean, high‑quality data for reliable results【809275587000967†L341-L378】【444968039617338†L448-L474】.

## Reflection questions

1. Why is correlation unaffected by changes in units, whereas regression coefficients are not?
2. Describe steps you would take to clean a dataset prior to regression analysis.
3. How can outliers distort correlation and regression results?

## References