Introduction to Data Science

- Explanatory Data Analysis: Correlation, Covariance & Simple Linear Regression -

Mauricio Molina

Keio University, Faculty of Economics

May 14, 2025

慶應義塾
Keio University

Contents

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the process of summarizing and visualizing a dataset to uncover its main characteristics and inform subsequent modeling steps.

- **Data Cleaning & Preprocessing:** Handle missing values, correct data types, and remove or impute erroneous entries.
- **Descriptive Statistics:** Compute measures of center (mean, median), dispersion (variance, IQR), and shape (skewness, kurtosis).
- **Visual Exploration:** Use histograms, box plots, scatter plots, and bar charts to reveal distributions, patterns, and relationships.
- **Correlation & Dependency Analysis:** Examine pairwise relationships with correlation matrices and scatter-plot matrices.
- **Outlier & Anomaly Detection:** Identify unusual observations that may distort analyses or signal data issues.
- **Feature Engineering & Selection:** Create new variables, transform features, and select the most informative ones.
- **Iterative Process:** Refine each step based on insights until the data is ready for modeling.

What is Covariance?

- Covariance measures how two variables change together.
- A positive covariance indicates that the variables tend to increase or decrease simultaneously.
- A negative covariance indicates that one variable tends to increase while the other decreases.
- A covariance near zero suggests no predictable pattern of change.
- Essential for understanding the relationship between variables in EDA.

Formal Definition of Covariance

Definition (Covariance)

The covariance between two random variables X and Y is given by:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \tag{1}$$

where:

- $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$ are the means of X and Y, respectively.

- Covariance units depend on the variables and can vary widely, making it difficult to interpret its magnitude directly.
- This leads to the use of correlation for standardized measurement.

Sample Covariance

The sample covariance for paired observations (x_i, y_i), $i = 1, 2, \dots, n$, is given by:

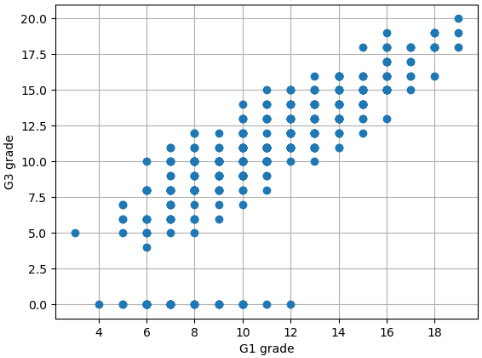
$$S_{x,y} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{2}$$

where:

- $\bar{x}$ and $\bar{y}$ are the sample means of X and Y, respectively.
- n is the number of paired observations.

Example on the Student's dataset

Load the student's dataset of the previous lesson. Consider the variables G1: First period grade, and G3: Third period grade. Let's plot the scatterplot for these variables.



Example Code

```
# Scatter plot
plt.plot(student_data['G1'],
 student_data['G3'], 'o')

# Add labels
plt.xlabel('G1 grade')
plt.ylabel('G3 grade')
plt.grid(True)
```

Sample Covariance Matrix

The sample covariance matrix **S** for a set of variables $X_1, X_2, \ldots, X_p$ is defined as:

$$\mathbf{S} = \begin{pmatrix} S_{X_1,X_1} & S_{X_1,X_2} & \cdots & S_{X_1,X_p} \\ S_{X_2,X_1} & S_{X_2,X_2} & \cdots & S_{X_2,X_p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{X_p,X_1} & S_{X_p,X_2} & \cdots & S_{X_p,X_p} \end{pmatrix} \tag{3}$$

where each element S_{X_i,X_j} is the sample covariance between variables X_i and X_j.

np.cov

Numpy offers a method to compute the covariance matrix

Example Code

```
# Covariance matrix
np.cov(student_data['G1'], student_data['G3'])
```

By definition, the covariance matrix is symmetric (i.e., $S = S^{'}$), and the diagonal elements are the variances.

What is Correlation?

- Correlation measures the strength and direction of the relationship between two variables.
- Helps in understanding if and how strongly pairs of variables are related.
- Commonly visualized using scatter plots.
- Examples:
  - Income and consumption
  - Height and weight
  - Stock prices and economic indicators
- Important in exploratory data analysis (EDA) to inform modeling decisions.

Correlation Coefficient

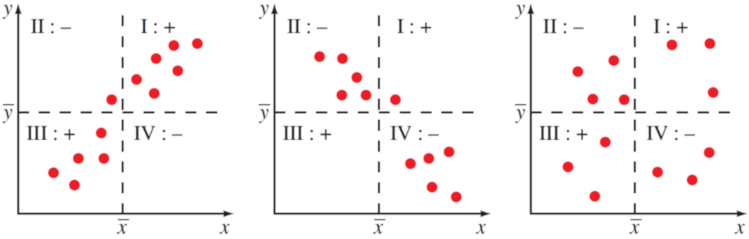
For the data in the previous example, you could describe each variable, x and y, individually using the means ($\bar{x}$ and $\bar{y}$) or the standard deviations (s_x and s_y).

A simple measure that helps to describe this relationship is called the **correlation coefficient**, denoted by r, and defined as:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Patterns and Interpretation of s_{xy} and r

- If most points are in areas I and III (positive pattern), s_{xy} and r will be positive.
- If most points are in areas II and IV (negative pattern), s_{xy} and r will be negative.
- If points are scattered across all four areas (no clear pattern), s_{xy} and r will be close to 0.



(a) Positive pattern (b) Negative pattern (c) No pattern

Interpretation of the Correlation Coefficient

- Correlation coefficient (r) always lies between -1 and 1.
- When $r > 0$, x increases as y increases.
- When $r < 0$, x decreases as y increases (or vice versa).
- If $r = 1$ or $r = -1$, points lie exactly on a straight line.
- If $r = 0$, there is no apparent linear relationship between the two variables.
- The closer r is to 1 or -1, the stronger the linear relationship.

np.corrcoef

Numpy offers a method to compute the correlation coefficient matrix

Example Code

```
# Correlation coeficient
np.corrcoef(student_data['G1'], student_data['G3'])
```

By definition, the covariance matrix is symmetric (i.e., $S = S^{'}$), and the diagonal elements are the variances.

Coefficient of Variation (CV)

Definition (Coefficient of Variation)

The **Coefficient of Variation**, denoted by **CV**, measures the relative variability of data and is defined as the ratio of the standard deviation to the mean:

$$CV = \frac{s}{\bar{x}}$$

It is often expressed as a percentage:

$$CV(\%) = \frac{s}{\bar{x}} \times 100$$

where:

- s is the sample standard deviation.
- $\bar{x}$ is the sample mean.

CV is useful for comparing the relative variability between datasets with different units or scales.

CV

Example Code

```
# Coefficient of Variation
cv1 = student_data['G1'].std()/student_data['G1'].mean()
cv3 = student_data['G3'].std()/student_data['G3'].mean()
print(f"The CV for G1: {cv1*100 : .2f}%")
print(f"The CV for G3: {cv3*100 : .2f}%")
#Output
The CV for G1:  30.43%
The CV for G3:  43.99%
```

The CV of `G1` is 30.43%, indicating moderate variability relative to its mean. The CV of `G3` is higher (43.99%), suggesting greater relative variability and dispersion in the final grades compared to the first evaluation. A higher CV in `G3` could imply that performance became more heterogeneous by the end of the period.

Exploring Relationships with Pair Plots

Variables of Interest:

- `Dalc` – Workday alcohol consumption (1 = very low, 5 = very high)
- `Walc` – Weekend alcohol consumption (1 = very low, 5 = very high)
- `G1` – First period grade (0 to 20)
- `G3` – Final period grade (0 to 20)

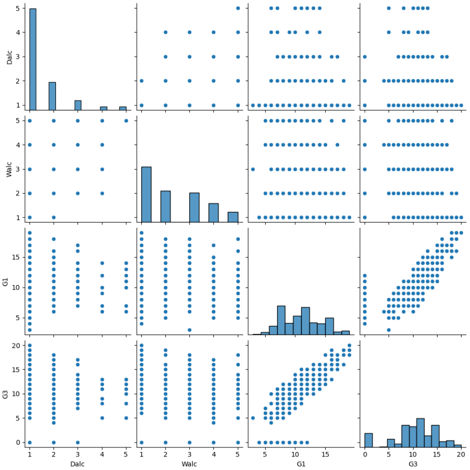
Example Code

```
sns.pairplot(student_data[['Dalc', 'Walc', 'G1', 'G3']])
plt.grid(True)
```

Interpretation:

- The pairplot provides scatter plots for each pair of variables and histograms on the diagonals.
- It helps visualize potential relationships (e.g., negative correlation between alcohol consumption and grades).
- Patterns suggest whether variables are linearly related, clustered, or independent.
- Look for clear downward trends between `Dalc`/`Walc` and `G1`/`G3` to evaluate the impact of alcohol on academic performance.

Pairplot: Alcohol Consumption and Grades



Key Observations:

- Strong positive linear relationship between `G1` and `G3`.
- No strong visual correlation between `Dalc`/`Walc` and grades.
- Most students report low alcohol consumption on both weekdays and weekends.
- Some clusters suggest lower grades may coincide with higher alcohol consumption, but pattern is not strong.
- Pairplot is useful for quickly spotting relationships and data concentration.

Simple Linear Regression

Goal: Model the relationship between two numerical variables.

- Predict a response variable y using an explanatory variable x.
- The model assumes a linear relationship:

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

- β_0: intercept (value of y when $x = 0$)
  β_1: slope (change in y for a one-unit increase in x)
  ε: error term (difference between observed and predicted values)

The Least Squares Line

How is the best line chosen?

- The **least squares line** minimizes the sum of squared residuals:

$$\text{Residual} = y_i - \hat{y}_i$$

$$\text{Objective:} \quad \min \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$

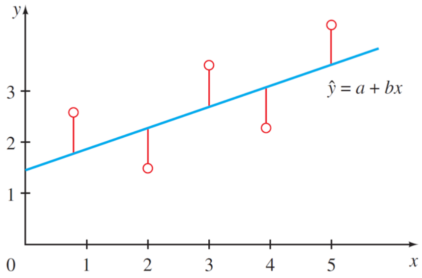
- This gives the line that best fits the data in terms of minimizing prediction errors.

The Best Fitting Line

- The **best fitting line** minimizes the overall prediction error.
- It tries to reduce:

$$\sum(\text{actual } y_i - \text{predicted } \hat{y}_i)^2$$

- These differences are called **residuals**.
- A good fit shows residuals that are randomly scattered (no pattern).



Formulas for Slope and Intercept

Given a set of paired data $(x_1, y_1), \ldots, (x_n, y_n)$, the least squares regression line:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

has coefficients:

Slope:

$$\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}$$

where s_{xy} is the sample covariance between x and y, and s_x^2 is the sample variance of x.

Intercept:

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

where $\bar{x}$ and $\bar{y}$ are the sample means of x and y, respectively.

These formulas define the **best-fitting line** that minimizes squared prediction errors.

Formula for R^2 (Coefficient of Determination)

Definition: The R^2 value measures the proportion of variance in the response variable y that is explained by the explanatory variable x in the regression model.

Formula:

$$R^2 = \frac{\text{Explained Sum of Squares (SSR)}}{\text{Total Sum of Squares (SST)}} = 1 - \frac{\text{Residual Sum of Squares (SSE)}}{\text{Total Sum of Squares (SST)}}$$

Equivalently, using notation:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

Interpretation:

- $R^2 = 0$: The model explains none of the variability in y.
- $R^2 = 1$: The model explains all the variability in y.
- Higher R^2 means better model fit (in terms of variance explained).

Creating a Linear Regression Model

Code Example

```
from sklearn import linear_model

# Define instance of linear regression
reg = linear_model.LinearRegression()
```

Explanation:

- We import the `linear_model` module from `scikit-learn`.
- `LinearRegression()` creates an instance of a linear regression model.
- This object `reg` will be used to fit the model to data and make predictions.

Note: No model is trained yet — this just creates the object.

Fitting a Simple Linear Regression Model

Code Walkthrough

```
# Extract explanatory variable (as 2D array)
X = student_data.loc[:, ['G1']].values

# Extract response variable (as 1D array)
Y = student_data['G3'].values

# Fit the model
reg.fit(X, Y)

# Print model parameters
print(f"Slope: {reg.coef_}")
print(f"y-Intercept: {reg.intercept_}")
```

- `X` must be a 2D array — that's why we use double brackets.
- `Y` is the target (1D).
- `reg.fit(X, Y)` trains the model to find the best-fitting line.
- `reg.coef_` gives the slope, and `reg.intercept_` gives the y-intercept of the line.

Evaluating the Model: R^2 Score

Code Example

```
# Print the R^2
print(f"R-squared: {reg.score(X, Y)}")
# Output: R-squared: 0.64235084605227
```

Interpretation:

- The R^2 (coefficient of determination) measures how well the model explains the variation in the response variable.
- Value ranges from 0 to 1:
  - $R^2 = 1$: perfect fit
  - $R^2 = 0$: model explains none of the variability
- In this case, $R^2 \approx 0.64$ means that:

  Around 64% of the variability in $G3$ is explained by $G1$.

- This indicates a moderate-to-strong linear relationship.

Assignment Questions

Solve the following problems in your Jupyter Notebook. Ensure that all code runs without errors. Upload the completed `.ipynb` file to K-LMS by next Tuesday at midnight.

Q1: Load the Kaggle dataset on Car Price Prediction (Multiple Linear Regression) from the previous assignment.

- Print the covariance matrix for the variables `price` and `enginesize`.
- Print the correlation coefficient between `price` and `enginesize`.
- Based on these calculations, what can you conclude about the relationship between these variables?
- Create a pair plot of the two variables.

Q2: Using the same dataset:

- Consider the variables: `price`, `enginesize`, `wheelbase`, `horsepower`, and `carheight`. Create a pair plot for all of them.
- Read the documentation for the `seaborn.heatmap` function. Use it to generate a heatmap for the variables above.
- What can you conclude about the relationships among these variables based on the visualizations?

Q3: Using the same dataset:

- Use `price` as the response variable. Perform simple linear regression using each of the variables listed in Q2 as explanatory variables.
- Print the slope, intercept, and R^2 value for each regression.
- Create scatter plots and plot the corresponding least squares regression line for each case.
- Which regression provides the best fit? Explain why.