Lecture 08

FUNDAMENTALS OF STATISTICS

Probability & Statistics

Francisco Richter, Martina Boschi and Ernst Wit

In previous lectures, we extensively studied random variables, particularly denoted as X. As we transition into statistics, its essential to understand how these random variables relate to the data we analyze.

Definition 1 (Data as Realizations)**.** *Given a random variable X, any observed value of X is called a realization of X, denoted as x. When we have a sequence of independent and identically distributed (i.i.d.) random variables, say $X_1, X_2, \ldots, X_n$, the observed values of these random variables constitute our data $x_1, x_2, \ldots, x_n$.*

Figure 1: Weight stack showing progression in resistancemost workouts center around a comfortable range, with fewer sets at the extremes, resembling a normal distribution of effort.

Definition 2 (Dataset)**.** *A dataset is a collection of realizations from one or more random variables. If these realizations are from i.i.d. random variables, then each realization is an independent observation from the same underlying probability distribution.*

Example (Realizations from Random Variables in a Study of Effort)**.** Consider a study examining the effort exerted by gym users across different weight machines. In this study, each machine in the gym (e.g., chest press, leg press, lat pulldown) represents a different column in our dataset. The variable X represents the amount of wear on the weights, measured in millimeters in the worn zone of each weight plate on a given machine.
For instance, lets focus on one specific machine, represented by the weight stack in the image. Over time, as users select different weights on this machine, certain weights experience more wear than others. This wear, measured in millimeters, reflects the intensity and frequency of use at each weight level.
Suppose we observe the following wear data (in millimeters) for five weight levels on this machine:

$$x_1 = 2.3, \quad x_2 = 3.7, \quad x_3 = 1.8, \quad x_4 = 4.2, \quad x_5 = 3.1$$

These values represent the observations $x_1, x_2, \ldots, x_5$ of the random variables $X_1, X_2, \ldots, X_5$ for this machine, indicating the wear in millimeters on each weight.

Table 1: Wear observed at different weight levels on a gym machine

Weight Level (kg)	Wear (mm)
10	1.5
20	2.4
30	3.1
40	3.8
50	4.5
60	4.0
70	3.6
80	2.9
90	2.2
100	1.8
110	1.2
120	0.9
130	0.5

This table represents a subset of the dataset, capturing the wear levels as realizations of random variables corresponding to the weights on this particular machine. In a broader analysis, data across multiple machines could provide insights into user effort patterns and preferred weight ranges across the gym. □

Descriptive Methods

Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

Definition 3 (Statistic)**.** *A statistic is any function of data that does not depend on any unknown parameters. Formally, given random variables $X_1, X_2, \ldots, X_n$ representing observations drawn from a population with joint probability distribution $f(x_1, x_2, \ldots, x_n; \theta)$, where θ is an unknown parameter (or vector of parameters), a statistic T_n is defined as:*

$$T_n = g(X_1, X_2, \ldots, X_n) \tag{1}$$

where g is a known function that does not involve the parameter θ. The probability distribution of T, induced by the joint distribution of the data, is called the sampling distribution of the statistic.

Measures of central tendency and dispersion are foundational in statistics for summarizing data.

Example (Mean)**.** The mean, denoted as $\bar{x}$, is the sum of all observations divided by the number of observations:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

The mean provides a measure of the central tendency, representing the average value of the data set.
$\bar{x}$ represents the observed value of $T_n = \bar{X} = \sum_{i=1}^{n} X_i$. The mean of random variables is a random variable as well, whose realization corresponds to $\bar{x}$.

□

Example (Median)**.** The median is the value that separates the higher half from the lower half of a data sample. For a dataset with an odd number of observations, it is the middle element, while for an even number of observations, it is the average of the two middle elements:

$$\text{median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{if } n \text{ is even} \end{cases}$$

The median is useful for understanding the central location of the data, especially when the data has outliers or is skewed. As before, the median can be seen as the realization of the related statistics, defined by considering random variables instead of realizations. □

Example (Quantile)**.** A quantile is a cutoff point dividing the range of a probability distribution into continuous intervals with equal probabilities. The q-th quantile, denoted as Q_q, is the value such that a proportion q of the data is less than or equal to Q_q. Quantiles generalize the concept of percentiles:

$$Q_q = P_{q \times 100}$$

Common quantiles include:

- **Quartiles**: Divide data into four equal parts ($q = 0.25, 0.50, 0.75$).
- **Deciles**: Divide data into ten equal parts ($q = 0.1, 0.2, \ldots, 0.9$).
- **Percentiles**: Divide data into one hundred equal parts ($q = 0.01, 0.02, \ldots, 0.99$).

Quantiles are useful for summarizing the distribution of data, especially for understanding its spread and skewness. □

Example (Variance)**.** Variance measures how sampling random variables spread from their expected mean.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Once the variance is evaluated in observed data, it expresses an indication of how spread out the values are around the mean.

□

Example (Standard Deviation)**.** The standard deviation σ is the square root of the variance:

$$\sigma = \sqrt{\sigma^2}$$

This measure provides a sense of the spread or dispersion of the distribution, expressed in the same units as the data itself. □

Example (Range)**.** The range is the difference between the largest and smallest values in a dataset:

$$\text{range} = \max(x_i) - \min(x_i), i = 1, ...n$$

The range gives a quick sense of the spread of the data, showing the extent of variation in the dataset. □

Example (Covariance)**.** Let us consider the situation in which both the random variables X and Y are observed n times. In this case, the covariance between the two can be found as follows:

$$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Covariance indicates whether two variables tend to increase or decrease simultaneously. A positive covariance indicates that the variables increase together, while a negative covariance indicates an inverse relationship. The covariance between their realizations, denoted as σ_{XY}, measures the degree to which the two variables vary together.

□

Example (Histogram)**.** A histogram is a graphical representation of the frequency distribution of numerical data, dividing the data range into intervals, or bins. For continuous data, the histogram $h(x)$ can be mathematically defined as:

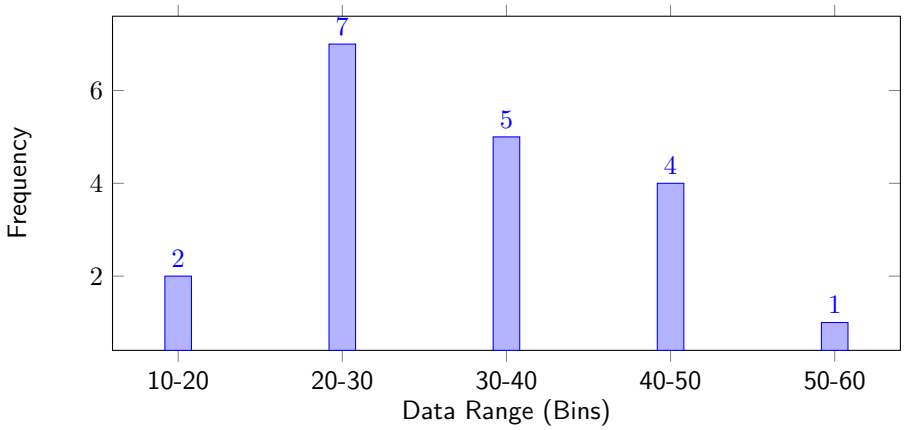
$$h(x) = \sum_{i=1}^{n} I(x \in \text{bin}_k)$$

where $I(x \in \text{bin}_k)$ is an indicator function that takes the value:

$$I(x \in \text{bin}_k) = \begin{cases} 1, & \text{if } x \text{ falls into bin } k \\ 0, & \text{otherwise} \end{cases}$$

Histograms provide a visual summary of the distribution of a dataset, with the height of each bar representing the frequency of values within each bin.

□

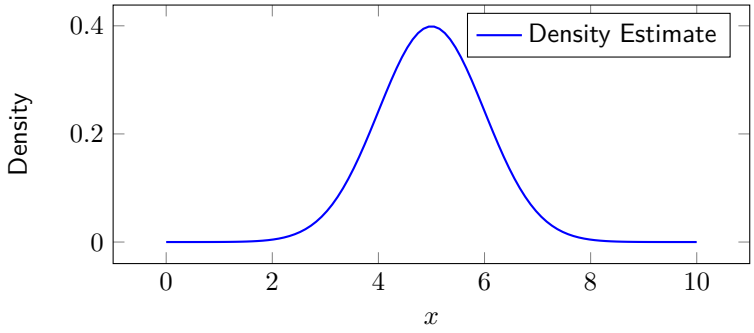


Example (Density Plot)**.** A density plot is a smoothed version of the histogram that provides a continuous estimation of the data distribution, particularly useful for visualizing the shape of distributions without the granularity of discrete bins. The density function $\delta(x)$ is calculated using kernel density estimation, defined as follows:

$$\delta(x) = \frac{1}{nm} \sum_{i=1}^{n} K\left(\frac{x - x_i}{m}\right)$$

where K is a kernel function (e.g., Gaussian) and m is the bandwidth parameter that controls the smoothness of the estimate. Density plots reveal the underlying structure of the data more clearly than histograms. □

Although it may be less obvious, both the histogram and the density plot are still examples of statistics. While we introduced them as functions of the observed data, they are, in principle, functions of the sampling random variabless $X1, ..., X_n$, as defined in 1. This discussion may be performed for box and violin plots.



Example (Box Plot)**.** A box plot summarizes data using five key statistics: the minimum, first quartile (Q_1), median (Q_2), third quartile (Q_3), and maximum. The interquartile range (IQR), defined as $Q_3 - Q_1$, highlights the spread of the central 50% of the data. Box plots are useful for identifying central tendency, dispersion, and potential outliers. □

Example (Violin Plot)**.** A violin plot combines a box plot with a density plot, showing both summary statistics and the distribution shape. Violin plots are particularly valuable when comparing the distribution shapes across different groups, as they provide insights into the spread and concentration of data points in various ranges. □

For a visual comparison, the figure below illustrates data represented in three different ways: as individual data points (dot plot), as a box plot, and as a violin plot. Each representation provides a unique perspective on the data, allowing for analysis of individual values, summary statistics, and overall distribution shape.

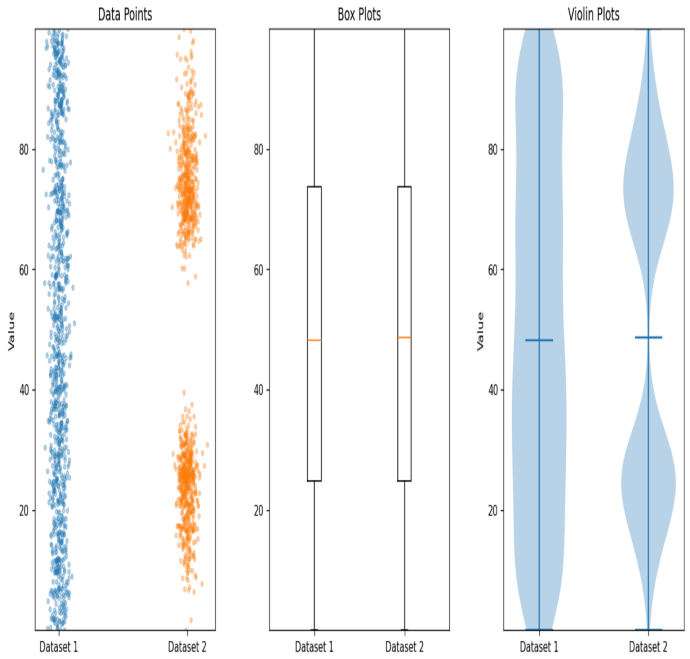


Figure 2: Data represented as data points, box plots, and violin plots for two datasets.

In this figure:

- **Data Points (Dot Plot)**: This representation shows individual observations for each dataset, providing a detailed view of the distribution.
- **Box Plot**: The box plot summarizes central tendency and spread using the median, quartiles, and range, and also highlights potential outliers.
- **Violin Plot**: This plot combines a box plot with a kernel density estimate, showing both summary statistics and the distribution shape for each dataset.

These visualization techniques are useful tools for exploratory data analysis, but each has limitations. Box plots, for example, can obscure underlying differences in data distributions by focusing solely on summary statistics, as seen when two distinct datasets produce identical box plots. Violin plots reveal distribution shapes but may be misleading with small samples, while dot plots show individual values but can become cluttered with large datasets. Therefore, selecting appropriate visualization methods is crucial; combining multiple techniques can provide a more comprehensive view, helping researchers avoid the pitfalls of relying on any single method.

Data Exploratory Analysis

Imagine you want to deepen our understanding of human effort in a gym environment. The gym administrators, recognizing the value of data-driven insights, decide to support this goal by collecting detailed data on equipment usage, user demographics, and session characteristics. By analyzing this data, we aim to uncover patterns in gym usage, examine how different factors contribute to user effort, and explore potential areas for optimizing user experience and equipment maintenance.
Exploratory Data Analysis (EDA) is an approach for summarizing and visualizing key characteristics of data, transforming raw data into meaningful insights. EDA allows us to understand data distributions, relationships among variables, and potential trends or anomalies that might inform further investigation or modeling.
Dataset Overview:
The dataset includes records for different gym sessions across various machines, capturing user demographics, session details, and performance metrics. A sample of the dataset is shown below:

ID	Type	User	Age	Gender	Duration	Weight	Reps	Freq.
1	Chest	101	25	M	20 min	40	12	120
2	Leg	102	30	F	15 min	60	15	90
3	Lat	101	25	M	10 min	50	10	60
4	Shoulder	103	28	F	12 min	30	8	45
5	Row	104	40	M	25 min	50	20	110

Table 2: Sample entries from the gym equipment dataset

This table provides a snapshot of session details, including age, gender, session duration, weight level, repetitions, and frequency of use. This structure enables us to analyze factors influencing user effort.

Descriptive Statistics:
To gain a deeper understanding, we calculate descriptive statistics for the weight levels used on the "Chest Press" machine, broken down by gender:

Gender	Count	Mean	Std Dev	Min	25%	Median	75%	Max
F	11	52.45	14.56	36	41.5	47	61.5	82
M	7	49.86	17.88	28	38	50	59	77

Table 3: Descriptive statistics of weight levels used on the Chest Press machine by gender

The statistics reveal that females and males both tend to use similar average weight levels on the "Chest Press" machine, with means around 50 kg. However, females show a slightly narrower range (36-82 kg), while males display a broader distribution (28-77 kg). This suggests that while weight preferences for this machine are similar across genders, individual variation may be slightly higher among male users.
Correlation Analysis:
The following pairplot and violin plot further illustrate relationships within the data:

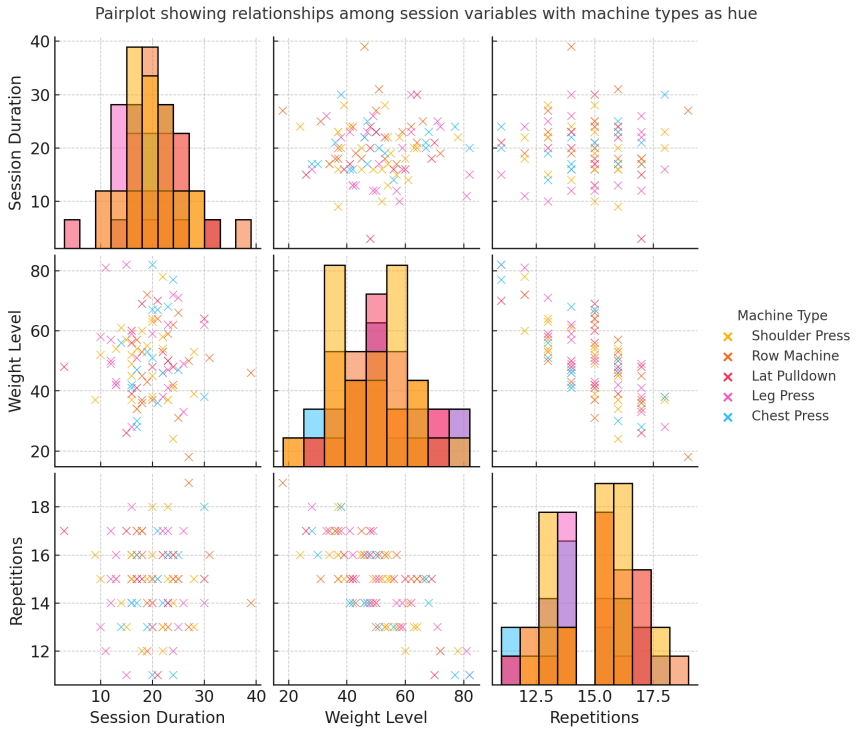


Figure 3: Pairplot showing relationships among session duration, weight level, and repetitions, colored by machine type

The pairplot reveals clustering by machine type and a negative trend between weight level and repetitions, particularly for high-weight sessions on machines like the Chest Press and Leg Press.

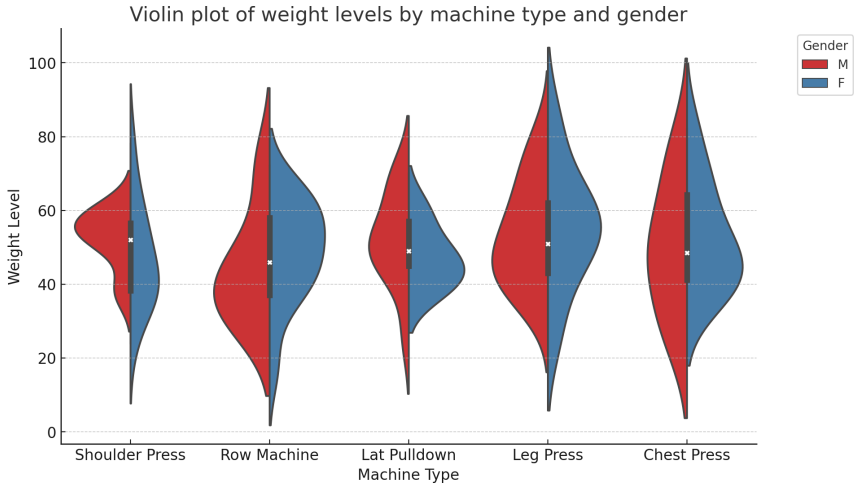


Figure 4: Violin plot of weight levels by machine type, with gender as color coding

The violin plot highlights differences in weight level distributions by machine type, with gender distinctions. Males generally select higher weights across machines, while females show a slightly narrower distribution. The Row Machine and Leg Press show the highest weights overall, indicating their use in more intense sessions.
This exploratory analysis provides insights into gym usage patterns and the relationships among session characteristics. Understanding these patterns may help in guiding equipment maintenance needs, optimizing user experience, and tailoring recommendations based on user demographics.

Statistical Learning

In the theoretical framework of statistical modeling, our objective is to examine the relationship between a dependent variable Y and one (or more) independent variables $\underline{X}$.
Here, Y represents the outcome we seek to predict or explain, while the vector $\underline{X}$ comprises the explanatory or predictor variables that inform our predictions of Y.
This relationship is theoretically expressed as:

$$Y = f(\underline{X}) + \epsilon$$

where:

- $f(\underline{X})$ embodies the systematic information conveyed by $\underline{X}$ regarding Y. This function f is unknown and reflects the true underlying process that we aim to understand.
- ϵ captures the random error, accounting for variations in Y not explained by $\underline{X}$, with an expected value of zero.

Definition 4 (Statistical Model)**.** *A statistical model is a mathematical representation of observed data. In the context of regression, we often describe the model as $Y = f(\underline{X}) + \epsilon$, where Y is the dependent variable, $\underline{X}$ is the independent variable, f is the function that represents the systematic relationship between $\underline{X}$ and Y, and ϵ represents the error term, capturing all other factors affecting Y that are not included in $\underline{X}$.*

Transitioning from this theoretical construct to empirical application, we collect a dataset with n observations. Each observation i in the dataset comprises an actual outcome y_i and the corresponding values of independent variables $\mathbf{x}_i$. The practical challenge in regression analysis lies in estimating a function $\hat{f}$ from the observed data that serves as a surrogate for the true function f. This estimated function $\hat{f}$ is what we use to predict new values of y based on observed $\mathbf{x}$.
Accordingly, the estimated relationship is given by:

$$y_i = \hat{f}(\mathbf{x}_i) + e_i$$

where $\hat{f}(\mathbf{x}_i)$ is the predicted value of y based on the i-th observation's values of $\mathbf{x}$, and e_i is the estimation error for that observation.
Naturally, to evaluate how well the model describes the data, we can consider metrics like the Mean Squared Error (MSE), which is computed as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where:

- $\hat{y}_i$ represents the predicted value of the dependent variable for the i-th observation, given by the model as $\hat{y}_i = \hat{f}(\mathbf{x}_i)$.

The goal of the regression analysis is, therefore, to find the estimated function $\hat{f}$ that minimizes the MSE, reflecting the closest approximation to the true function f that generated the observed data.

Estimation

In the case of linear regression, $f(\mathbf{X})$ is a linear function of the independent variables $\mathbf{X}$, as follows:

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$$

Once we observe the realizations of the random variables, we obtain:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + e_i, i = 1, ..., n$$

We can express this relationship, for all the observations, by expressing the matrix form of the residuals, as:

$$\mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{y} - \hat{\mathbf{y}}$$

where:

- $\mathbf{y}$ be the $n \times 1$ vector of observed dependent variable values $[y_1, y_2, \ldots, y_n]^\top$,
- $\mathbf{X}$ be the $n \times k$ matrix of predictor variables, where each row $\mathbf{x}_i^\top$ corresponds to the i-th observation.

  REMARK: If our statistical model includes an intercept term (i.e., β_0 is the first element of $\boldsymbol{\beta}$), then the first element of each vector $\mathbf{x}_i^\top$ is 1. Consequently, the first column of $\mathbf{X}$ consists entirely of 1s.
- $\boldsymbol{\beta}$ be the $k \times 1$ vector of coefficients to be estimated.

Consider the problem of estimating the parameters $\boldsymbol{\beta}$ within a linear regression framework, where the goal is to minimize the Mean Squared Error, that, in matrix form, is given by:

$$\text{MSE}(\boldsymbol{\beta}) = \frac{1}{n}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

To derive the Ordinary Least Squares (OLS) estimator, we look for the value of $\boldsymbol{\beta}$ that minimizes the MSE.
We start by setting the gradient of the MSE with respect to $\boldsymbol{\beta}$ equal to zero:

$$S(\beta) = \frac{\partial}{\partial \boldsymbol{\beta}}\text{MSE}(\boldsymbol{\beta}) = -\frac{2}{n}\mathbf{X}^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = 0$$

Solving for $\boldsymbol{\beta}$ yields the normal equations:

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

If $\mathbf{X}^\top\mathbf{X}$ is invertible, we can isolate $\boldsymbol{\beta}$, allowing us to express the OLS estimator in closed form:

$$\hat{\boldsymbol{\beta}} = \max_{\boldsymbol{\beta}} \text{MSE}(\boldsymbol{\beta}) = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$$

This expression for $\boldsymbol{\beta}$ minimizes the MSE and is known as the OLS estimator.
Residuals of the model $\mathbf{e}$ are defined as portion of the dependent variable that is not explained by the model. They are a crucial diagnostic tool in regression analysis, as they allow us to assess the validity of the model assumptions such as homoschedasticity (constant variance of the error terms) and the absence of autocorrelation (the error terms are not correlated with each other).
It is also important to check the normality of the residuals, as the OLS method relies on the assumption that the error terms are normally distributed. This can be done using various statistical tests and graphical methods such as a Q-Q plot.

Theorem 5. *The sum of the OLS residuals is zero, i.e.,* $\mathbf{1}^\top \boldsymbol{e} = 0$ *where* $\mathbf{1}$ *is a vector of ones.*

Proof. Given the OLS estimator $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$, the predicted values can be written as $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Hence, the residuals are:

$$\boldsymbol{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

By definition of the OLS estimator, $\mathbf{X}^\top (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$. This implies that:

$$\mathbf{X}^\top \boldsymbol{e} = \mathbf{0}$$

Considering that $\mathbf{X}$ includes a column of ones (the intercept term), this leads to:

$$\mathbf{1}^\top \boldsymbol{e} = 0$$

which proves that the sum of the OLS residuals is zero. □

Example (Repetitions and Weight Level Regression Analysis)**.** Consider a gym dataset focused on the "Chest Press" machine, which includes attributes like weight level, repetitions, age, and gender. To understand the relationship between weight level and repetitions, we start with a simple linear regression and then introduce more complexity to improve the model.

Simple Linear Regression

The initial model examines the influence of weight level on the number of repetitions performed:

$$\text{Repetitions} = \beta_0 + \beta_1 \times \text{Weight Level} + \varepsilon \tag{2}$$

where β_0 is the y-intercept, β_1 is the slope (indicating the effect of weight on repetitions), and ε represents the error term. β_0 and β_1 compose vector β.

The fitted model parameters are:

$$\hat{\beta}_0 = 18.36 \quad \text{(Intercept)}$$
$$\hat{\beta}_1 = -0.077 \quad \text{(Slope)}$$

The negative slope indicates that higher weight levels are associated with fewer repetitions. The goodness-of-fit is assessed with $R^2 \approx 0.27$, suggesting a moderate relationship. The MSE for this model is 1.91.

Figure 5: Simple Linear Regression of Repetitions on Weight Level for the Chest Press machine.

Gender-specific Regression

To account for potential differences in repetitions between genders, we fit separate linear regression models for male and female users:

$$\text{Repetitions} = \beta_0 + \beta_1 \times \text{Weight Level} + \varepsilon \tag{3}$$

Separate models for each gender provide the following parameters:

$$\text{Male Model:} \quad \hat{\beta}_0 = 17.92, \quad \hat{\beta}_1 = -0.065$$
$$\text{Female Model:} \quad \hat{\beta}_0 = 18.74, \quad \hat{\beta}_1 = -0.083$$

The slopes indicate that weight affects repetitions similarly for both genders, with a slightly stronger negative effect for females. The average MSE for the gender-specific models is 1.74.

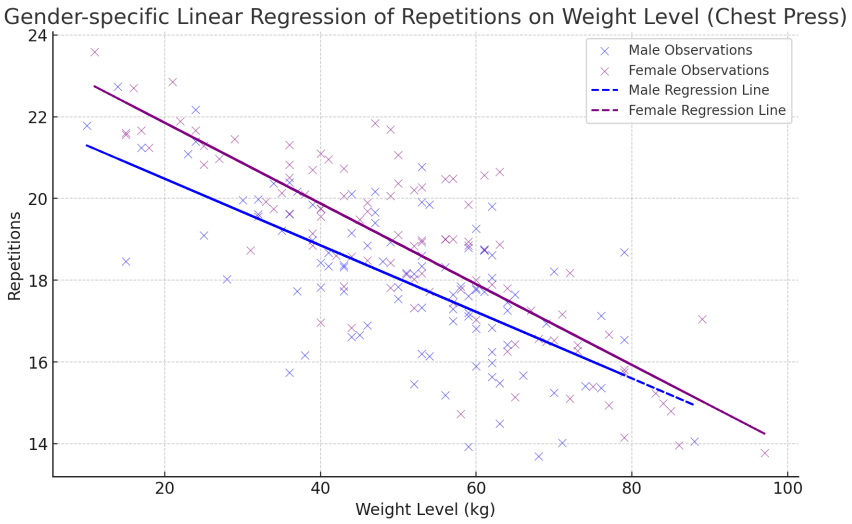


Figure 6: Gender-specific Linear Regression of Repetitions on Weight Level for the Chest Press machine. Red line represents male model, green line represents female model.

Multivariate Regression

Finally, we incorporate additional variables, such as age and gender, into a multivariate regression model:

$$\text{Repetitions} = \beta_0 + \beta_1 \times \text{Weight Level} + \beta_2 \times \text{Age} + \beta_3 \times \text{Gender (M)} + \varepsilon \tag{4}$$

where β_3 is an indicator variable for gender (1 for male, 0 for female). The MSE for this multivariate model is 1.71. In this case the true vector of parameters β is composed by $\beta_k, k = 0, 1, 2, 3$.

Comparison of Model Performance

The MSE values for each model are summarized below:

Model	MSE
Simple Linear Regression	1.91
Gender-specific Regression	1.74
Multivariate Regression	1.71

Table 4: MSE for Repetitions Prediction on Chest Press by Model Type.

This analysis shows that while the simple model provides a baseline understanding, incorporating gender and additional variables in the multivariate model slightly improves fit. Each approach provides insights into gym usage patterns, helping gym managers tailor equipment recommendations based on user attributes. □

Error Measurement

The MSE is a key metric for evaluating an estimator $\hat{f}$ when predicting outcome Y for new inputs $\underline{X}$. For simplicity of notation, let us consider a single random input X, though the discussion extends naturally to the multivariate case $\underline{X}$.

The MSE of an estimator $\hat{f}(\cdot)$ is defined as the expected value of the squared difference between the true output Y and the predicted value $\hat{f}(X)$:

$$\text{MSE}(\hat{f}) = E\left[(Y - \hat{f}(X))^2\right] \tag{5}$$

To analyze the sources of error, we expand Y as $f(X) + \epsilon$, where $f(X)$ is the true function and ϵ is the noise term. The squared error term then becomes:

$$\begin{aligned}(Y - \hat{f}(X))^2 &= (f(X) + \epsilon - \hat{f}(X))^2 \\ &= (f(X) - \hat{f}(X))^2 + 2\epsilon(f(X) - \hat{f}(X)) + \epsilon^2\end{aligned}$$

Since we work under the following assumptions:

- The error term is unbiased, $E[\epsilon] = 0$.
- ϵ is independent of $\hat{f}(X)$.

The MSE simplifies:

$$E\left[(Y - \hat{f}(X))^2\right] = E\left[(f(X) - \hat{f}(X))^2\right] + E[\epsilon^2] \tag{6}$$

The term $E\left[(f(X) - \hat{f}(X))^2\right]$ decomposes further into the bias and variance of the estimator:

$$E\left[(f(X) - \hat{f}(X))^2\right] = \text{Bias}(\hat{f}(X))^2 + \text{Variance}(\hat{f}(X)) \tag{7}$$

where:

$$\text{Bias}(\hat{f}(X))^2 = (E[\hat{f}(X)] - f(X))^2$$
$$\text{Variance}(\hat{f}(X)) = E[\hat{f}(X)^2] - (E[\hat{f}(X)])^2$$

The irreducible error, given by the variance of the noise ϵ, represents the portion of error that no model can reduce:

$$\text{Irreducible Error} = E[\epsilon^2] = E[\epsilon^2] - E[\epsilon]^2 = V[\epsilon] = \sigma_\epsilon^2 \tag{8}$$

Thus, the expected MSE decomposes into:

$$E\left[(Y - \hat{f}(X))^2\right] = \text{Bias}(\hat{f}(X))^2 + \text{Variance}(\hat{f}(X)) + \text{Irreducible Error} \tag{9}$$

This decomposition illustrates a fundamental tradeoff between bias and variance. Simplifying a model may introduce bias, while overly complex models may lead to high variance, especially when fitting data too closely. Minimizing MSE involves balancing bias and variance while acknowledging irreducible error.

Bootstrap Techniques

Bootstrap resampling is a powerful method for estimating the sampling distribution of a statistic. By repeatedly drawing samples with replacement from the original dataset, bootstrap resampling allows us to approximate the distribution of an estimator and compute metrics like confidence intervals and standard errors.
Given a dataset of size n, the bootstrap method involves drawing n observations with replacement to form a new sample. For each bootstrap sample $\mathbf{X}_b^*$, we calculate the statistic of interest θ_b^*:

$$\theta_b^* = g(X_b^*), \quad b = 1, 2, \ldots, B \tag{10}$$

where $g(\cdot)$ represents the statistic, and B is the number of bootstrap samples (often 1000 or more).
For example, to estimate the MSE of our multivariate model, we resample the data and refit the model for each bootstrap sample. This provides a distribution of MSE values from which we can compute the standard error:

$$SE(\text{MSE}) \approx \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\text{MSE}_b - \overline{\text{MSE}})^2} \tag{11}$$

To construct a $(1-\alpha) \times 100\%$ confidence interval for the MSE, we use the percentiles of the bootstrap replicates:

$$CI_{(1-\alpha)\times 100\%} = \left(\text{MSE}_{(\alpha/2)}, \text{MSE}_{(1-\alpha/2)}\right) \tag{12}$$

In our analysis, we computed the following values for the MSE from 1000 bootstrap samples:
These results highlight the stability and reliability of the multivariate model's performance across different data samples, as seen from the bootstrap estimates. The multivariate model effectively balances bias and variance, minimizing MSE and providing robust predictions for repetitions based on weight, age, and gender.

Metric	Value
Multivariate Model MSE	1.71
Bootstrap MSE SE	0.17
Bootstrap 95% CI Lower	1.36
Bootstrap 95% CI Upper	2.02

Table 5: Summary of MSE and Bootstrap Estimates for Multivariate Model.

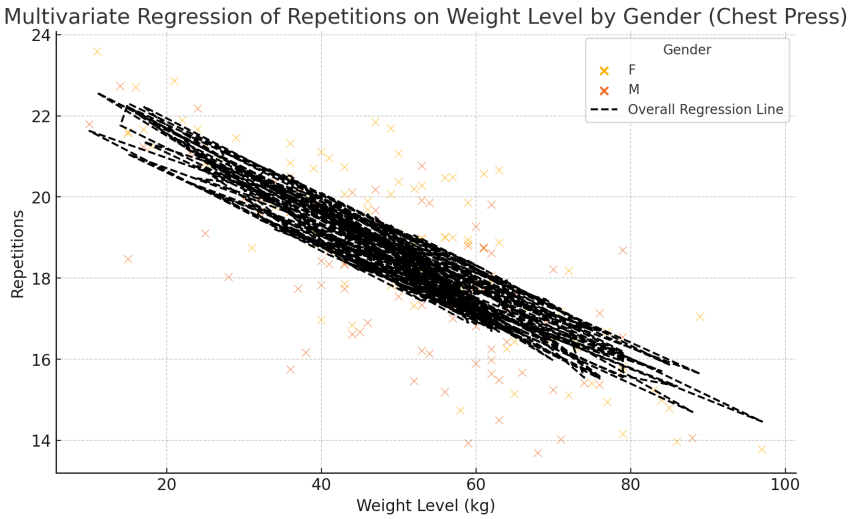


Figure 7: Multivariate Regression of Repetitions on Weight Level, Age, and Gender with Observations and Gender-Specific Trends.

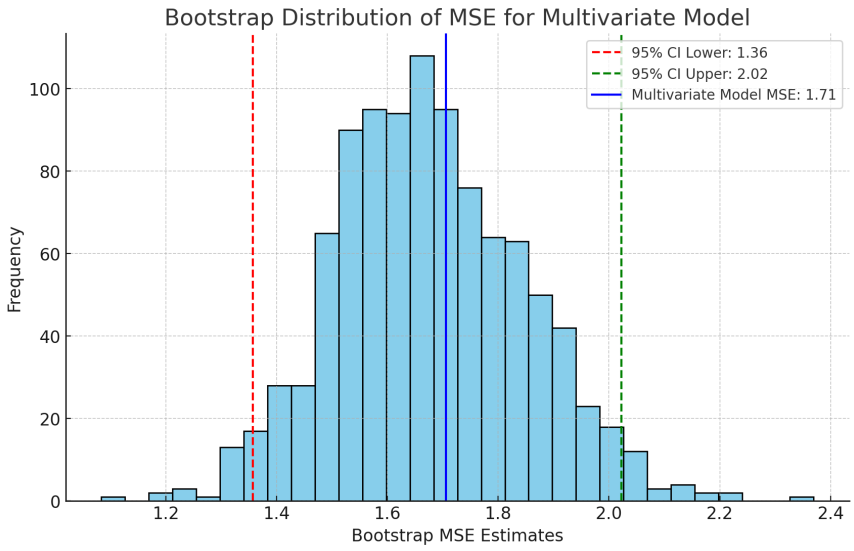


Figure 8: Bootstrap Distribution of MSE for Multivariate Model. Red and green lines represent the 95% confidence interval bounds, while the blue line indicates the observed model MSE.