Lecture Notes: From Random Variables to Data

In our previous lectures, we have extensively studied random variables, particularly denoted as X. As we transition into statistics, it's crucial 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 (resulting from an actual experiment or observation) is called a realization of 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.*

Mathematically, if $x_1, x_2, \ldots, x_n$ are the observed values from the random variables $X_1, X_2, \ldots, X_n$ respectively, then $x_1, x_2, \ldots, x_n$ are the realizations of these random variables and form our dataset.

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 1 (Realizations from Random Variables)**.** *Imagine a scenario where an ecologist is studying the heights (in meters) of a particular species of tree in a forest. She randomly selects five trees and measures their heights. Let the random variable X denote the height of a randomly chosen tree from this forest. If the five trees have heights 3.2m, 4.1m, 3.9m, 4.3m, and 3.8m, these heights are the realizations $x_1 = 3.2, x_2 = 4.1, \ldots, x_5 = 3.8$ of the random variables $X_1, X_2, \ldots, X_5$ respectively.*

Tree	*Height*
1	*3.2m*
2	*4.1m*
3	*3.9m*
4	*4.3m*
5	*3.8m*

This table, thus, represents a dataset of the realizations of 5 i.i.d. random variables, all showcasing the distribution of X.

This table represents a dataset of the realizations of 5 i.i.d. random variables, all following the distribution of X.

Statistics

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

In the realm of data analysis, *descriptive statistics* refers to the arsenal of tools and techniques used to summarize the main characteristics of a dataset. This includes graphical representations, measures of central tendency like the mean, median, and mode, and measures of dispersion such as variance and standard deviation. Within this domain, *summary statistics* provide a high-level overview, offering a quick snapshot of the data by reporting key numerical measures that capture the essence of the dataset's distribution, central tendency, and variability. While all summary statistics are descriptive, not all descriptive statistics can be considered as summary statistics, as the former may also encompass detailed graphical analyses.

Definition 3 (Statistic/Statistical Function)**.** *A statistic (also known as a statistical function) is any function of the data from a sample. More formally, given a sample $X_1, X_2, \ldots, X_n$ drawn from a population with a joint distribution $f(x_1, x_2, \ldots, x_n; \theta)$, where θ is a parameter, a statistic T is a function:*

$$T = g(X_1, X_2, \ldots, X_n)$$

where g is a function that does not depend on the unknown parameter θ. The distribution of T is called the sampling distribution of the statistic.

Measures of central tendency and dispersion are foundational in statistics for summarizing data. 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 median is the value separating 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 mode is the most frequently occurring value in a dataset. While it is straightforward to identify in a discrete data set, a mode is not always definable in continuous data distributions.

Variance σ^2 measures how far a set of numbers are spread out from their average value:

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

The standard deviation σ is the square root of the variance, providing a measure of the spread of the distribution of numbers:

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

The range is the difference between the largest and smallest values in a dataset:

$$\text{range} = \max(X_i) - \min(X_i)$$

The covariance between two random variables X and Y, denoted as σ_{XY}, measures the degree to which the two variables vary together:

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

The histogram is a statistical function that represents the frequency distribution of numerical data. It is mathematically described as:

$$h(x) = \begin{cases} f_k, & x \text{ in the k-th bin} \\ 0, & \text{otherwise} \end{cases}$$

where f_k is the frequency of data points in the k-th bin, and x is a value within the range of the data.

For visual representations, data visualization techniques like distribution plots (including histograms and density plots), box plots, scatter plots, pair plots, correlation matrices for relationships, and bar plots and count plots for categorical data are used to understand complex data structures more effectively.

Descriptive statistics and data visualization together provide a comprehensive understanding of data by revealing patterns, detecting outliers, and identifying relationships between variables.

Example 2. ***Exploratory Data Analysis of Tree Heights and Diameters***

The dataset consists of 100 entries, each representing a tree characterized by its ID, type (Oak, Pine, Maple), height in meters, age in years, and diameter at breast height (DBH) in centimeters. The purpose of this exploratory analysis is to understand the underlying distributions and relationships between the various attributes of these three tree species.

Summary of the Dataset:

The dataset begins with the following entries:

Tree ID	Type	Height (m)	Age (years)	DBH (cm)
1	Maple	13.73	12	26.99
2	Oak	24.04	92	34.52
3	Maple	16.31	67	28.36
4	Maple	14.06	15	27.17
5	Oak	23.89	85	34.44

Table 1: Sample entries from the tree dataset

Exploratory Analysis:

Exploratory Data Analysis (EDA) is conducted to summarize the main characteristics of the data, often using visual methods. In our case, we perform EDA to understand the distribution and relationship of tree heights and DBH across different ages and species.

Statistic	Height (m)	DBH (cm)
Mean	23.65	27.75
Standard Deviation	7.06	5.03
25% Quantile	17.77	22.39
50% Quantile (Median)	23.08	27.91
75% Quantile	30.09	32.65
Max	39.07	36.06
Min	10.58	20.15

Table 2: Descriptive statistics of the tree dataset

The correlation matrix for our dataset, showcasing the relationships between tree height, age, and DBH, is as follows:

$$Corr(\mathbf{X}) = \begin{bmatrix} 1.00 & 0.23 & -0.51 \\ 0.23 & 1.00 & 0.10 \\ -0.51 & 0.10 & 1.00 \end{bmatrix}$$

This matrix provides insight into how the variables correlate with each other. A value of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other also increases. A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other decreases. The values in our matrix suggest moderate correlations between

the variables, with tree height and DBH showing a moderate negative correlation, indicating that in this dataset, taller trees do not always have larger diameters.

Visualizations:

The pairplot provides a pairwise relationships visualization in the dataset, offering immediate insights into correlations between variables and distributions of single variables. The histograms along the diagonal show the distribution of each variable, and the scatter plots show the relationships between them.

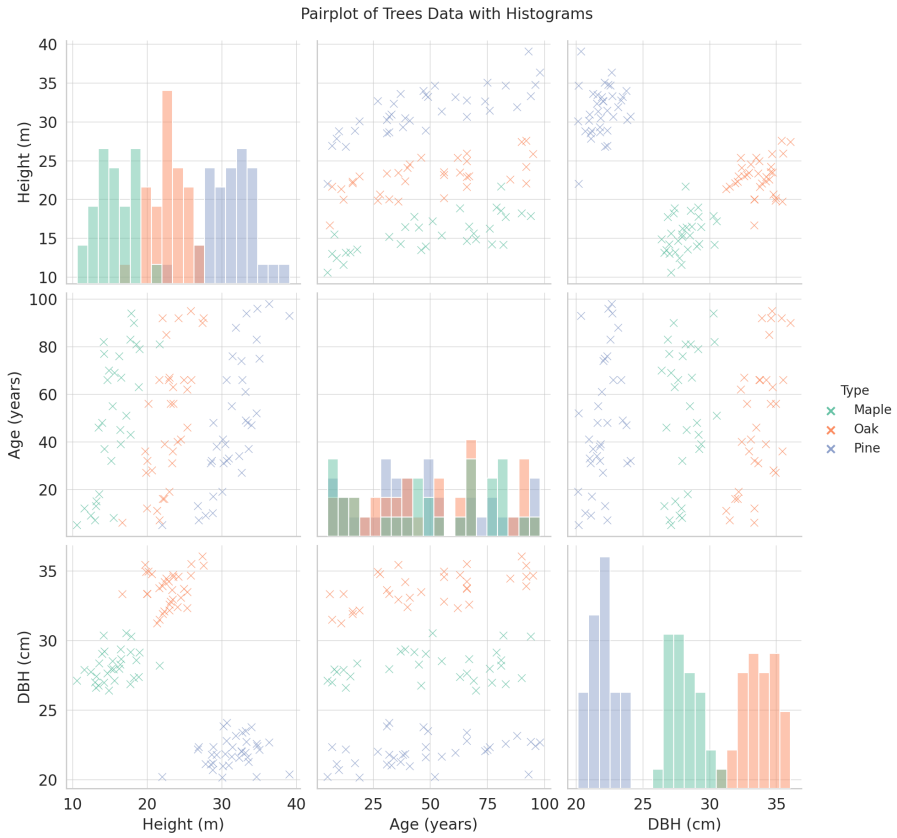

Figure 1: Pairplot showing the relationships between tree height, age, and DBH, colored by species type

Violin plots offer a more nuanced view of the data distribution than box plots. They display the probability density of the data at different values, with similar widths indicating similar frequencies. The inner markings show the median and quartiles.

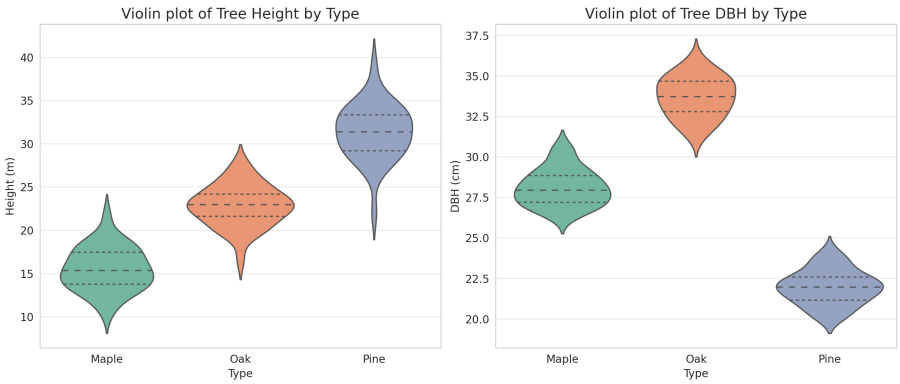

Figure 2: Violin plots of tree height and DBH by species type

Through these visualizations, it is evident that species type has a significant impact on both the height and DBH of the trees. Moreover, the variability within each type can also be assessed, which is crucial for further statistical analysis or predictive modeling.

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 $\mathbf{x}$. Here, y represents the outcome we seek to predict or explain, while the vector $\mathbf{x}$ comprises the explanatory or predictor variables that inform our predictions of y.

This relationship is theoretically expressed as:

$$y = f(\mathbf{x}) + \epsilon$$

where:

- $f(\mathbf{x})$ embodies the systematic information conveyed by $\mathbf{x}$ 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 $\mathbf{x}$, 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 $Y = f(X) + \epsilon$, where Y is the dependent variable, X is the independent variable, f is the function that represents the systematic relationship between X and Y, and ϵ represents the error term, capturing all other factors affecting Y that are not included in 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:

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

where $\hat{f}(\mathbf{x}_i)$ is the predicted value of y based on the i-th observation's values of $\mathbf{x}$, and ϵ_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:

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

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.

Linear regression

In the case of linear regression, $f(\mathbf{x})$ is a linear function of the independent variables $\mathbf{x}$. When we have multiple observations, we can express the linear regression model in matrix form as:

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

where:

- $\mathbf{y}$ is a $n \times 1$ vector of the observed dependent variable.
- $\mathbf{X}$ is a $n \times k$ matrix (with n observations and k predictors) of the independent variables.
- $\boldsymbol{\beta}$ is a $k \times 1$ vector of unknown parameters to be estimated.
- $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of the error terms.

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

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

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:

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

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

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

If $\mathbf{X}'\mathbf{X}$ is invertible, the OLS estimator $\boldsymbol{\beta}$ is given by:

$$\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

This expression for $\boldsymbol{\beta}$ minimizes the MSE and is known as the OLS estimator.

The residuals of the model are defined as:

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

where:

- $\mathbf{y}$ is the vector of observed values of the dependent variable.
- $\hat{\mathbf{y}}$ is the vector of predicted values obtained from the model, given by $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the vector of OLS estimates.

The residuals represent the 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 homoscedasticity (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 1. *The sum of the OLS residuals is zero, i.e.,* $\mathbf{1}'\boldsymbol{e} = 0$ *where* $\mathbf{1}$ *is a vector of ones.*

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

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

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

$$\mathbf{X}'\boldsymbol{e} = \mathbf{0}$$

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

$$\mathbf{1}'\boldsymbol{e} = 0$$

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

Example 3 (Tree Height and Age Regression Analysis)**.** *Consider the trees dataset comprising various species of trees and their corresponding attributes such as age, height, and diameter at breast height (DBH). To understand the influence of age on the height of a tree, a linear regression analysis is performed.*

The linear regression model is represented as:

$$\textit{Height} = \beta_0 + \beta_1 \times \textit{Age} + \varepsilon \tag{1}$$

where β_0 is the y-intercept, β_1 is the slope of the line (representing the effect of age on height), and ε represents the error term of the model.

Upon fitting the model to our data, the estimated parameters are:

$$\hat{\beta}_0 = 7.114 \quad (\textit{Intercept})$$
$$\hat{\beta}_1 = 0.076 \quad (\textit{Slope})$$

The coefficient $\hat{\beta}_1$ indicates that for each additional year of age, the tree's height is expected to increase by an average of 0.076 meters.

The goodness-of-fit for the model is assessed by the R^2 statistic, which is 0.78. This suggests that approximately 78% of the variability in tree height is explained by the age of the tree.

Residuals, which are the differences between the observed values and the values predicted by the model, are calculated as follows:

$$e_i = y_i - \hat{y}_i \tag{2}$$

where y_i are the observed heights and $\hat{y}_i$ are the heights predicted by the model.

A plot of the observed data points and the regression line is shown below:

The linear regression model's residuals for the first five trees are presented below:

Actual Height (m)	Predicted Height (m)	Residuals
11.58	21.44	-9.86
22.06	26.23	-4.18
16.51	24.73	-8.22
13.20	21.62	-8.42
22.57	25.81	-3.25

Table 3: Residuals of the linear regression model for tree height prediction.

The Mean Squared Error (MSE) for the model is calculated as follows:

$$\textit{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

Figure 3: Linear regression of tree height on age. The red line represents the best fit line obtained from the regression analysis.

For our model, the MSE is approximately 46.6671, which provides an estimate of the average squared difference between the observed actual outcomes and the outcomes predicted by the model.

we consider the different types of trees separately to determine if there are distinct growth patterns associated with each type. This approach allows us to tailor the regression model for each tree type, potentially improving the fit of the model and gaining better insights into species-specific growth characteristics.

We fit separate linear regression models to predict the height of Oak, Pine, and Maple trees based on their age. Each model's performance is assessed by calculating the Mean Squared Error (MSE), which indicates the average of the squares of the errors, i.e., the average squared difference between the observed actual outcomes and the predicted values by the model.

Figure 1 presents the scatter plot of tree heights by age, along with separate regression lines for each tree type, with the lines color-coded according to the "Set2" palette.

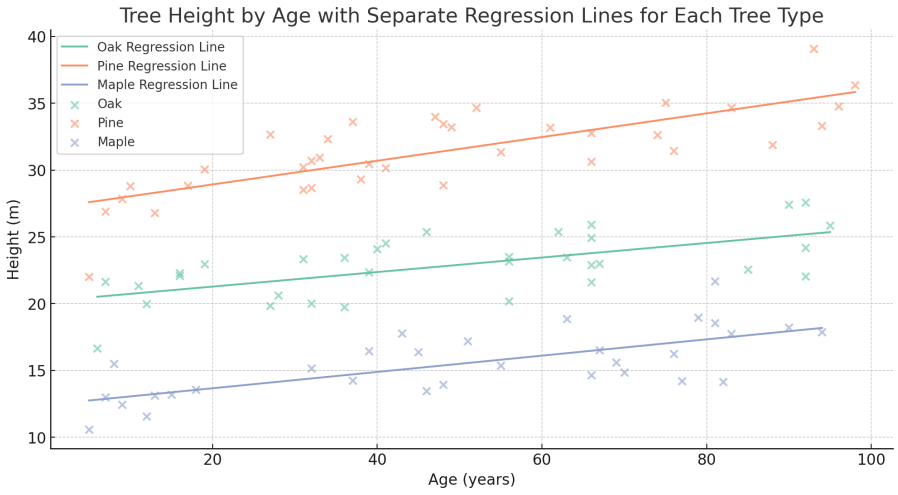

Figure 4: Tree Height by Age with Separate Regression Lines for Each Tree Type (Oak, Pine, Maple).

The MSE values for each tree type and the overall model are summarized in Table 1. These values demonstrate the effectiveness of the regression models for each species, with lower MSE values indicating a better fit.

Tree Type	MSE
Oak	3.2403
Pine	3.9088
Maple	2.9567
Overall	46.6671

Table 4: MSE for tree height prediction by tree type and overall.

The mean squared error (MSE) is a common measure used to evaluate the performance of an estimator. It captures the average of the squares of the errors, which are the differences between the predicted values by the model and the actual values. The expected MSE can be decomposed into three components: bias squared, variance, and irreducible error, corresponding to different sources of error in the prediction.

The MSE of an estimator $\hat{f}$ when predicting outcomes Y for new inputs x can be defined and decomposed as follows:

The MSE at a point x is the expected value of the squared difference between the true output Y and the predicted value $\hat{f}(x)$:

$$\text{MSE}(x) = \mathbb{E}\left[(Y - \hat{f}(x))^2\right] \tag{4}$$

The bias of an estimator is the difference between the expected prediction of our model and the true value. Variance measures the variability of the model prediction across different data sets. Assuming $f(x)$ is the true function, and ϵ is the error term (noise), we can expand Y as $f(x) + \epsilon$.

Expanding the squared term in the MSE equation yields:

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

Given that $\mathbb{E}[\epsilon] = 0$ and ϵ is independent of $\hat{f}(x)$, the cross-product term has an expected value of zero, simplifying our expression.

Taking the expected value over the squared difference, we get:

$$\mathbb{E}\left[(Y - \hat{f}(x))^2\right] = \mathbb{E}\left[(f(x) - \hat{f}(x))^2\right] + \mathbb{E}[\epsilon^2] \tag{5}$$

The term $\mathbb{E}\left[(f(x) - \hat{f}(x))^2\right]$ is decomposed into the bias and variance of the estimator:

$$\mathbb{E}\left[(f(x) - \hat{f}(x))^2\right] = \text{Bias}(\hat{f}(x))^2 + \text{Variance}(\hat{f}(x)) \tag{6}$$

where:

$$\begin{aligned}\text{Bias}(\hat{f}(x))^2 &= (\mathbb{E}[\hat{f}(x)] - f(x))^2 \\ \text{Variance}(\hat{f}(x)) &= \mathbb{E}[\hat{f}(x)^2] - (\mathbb{E}[\hat{f}(x)])^2\end{aligned}$$

The irreducible error is the variance of the error term ϵ, which cannot be reduced by any estimator:

$$\text{Irreducible Error} = \mathbb{E}[\epsilon^2] = \sigma_\epsilon^2 \tag{7}$$

Putting it all together, the expected MSE can be decomposed as:

$$\mathbb{E}\left[(Y - \hat{f}(x))^2\right] = \text{Bias}(\hat{f}(x))^2 + \text{Variance}(\hat{f}(x)) + \text{Irreducible Error} \tag{8}$$

The mean squared error (MSE) is a pivotal metric in statistical estimation, quantifying the average of the squares of errors between an estimator's predictions and the actual values. It decomposes into three components: bias, variance, and irreducible error. Bias represents systematic errors from incorrect model assumptions, variance measures prediction sensitivity to different data sets, and irreducible error is the noise inherently present in data that cannot be reduced by the model. A fundamental tradeoff exists between bias and variance; simplifying a model may increase bias due to oversimplification, while complex models may fit data too closely, leading to high variance. Minimizing MSE involves balancing bias and variance to acknowledge the presence of irreducible error and achieve reliable predictions within the given context.

Bootstrap Resampling

Bootstrap resampling is a statistical method for estimating the sampling distribution of a statistic by generating multiple samples from the original dataset. It involves drawing samples with replacement, which means that each observation can be selected more than once. The principle of bootstrap is to approximate the population distribution by the empirical distribution of the observed data.

Given an original dataset of size n, the bootstrap method consists of repeatedly sampling n observations with replacement to form a bootstrap sample. The statistic of interest, let's denote it as θ, is calculated for each bootstrap sample. This process is mathematically represented as:

$$\theta_b^* = s(X_b^*), \quad b = 1, 2, \ldots, B \tag{9}$$

where X_b^* is the b-th bootstrap sample and $s(\cdot)$ is the statistic of interest calculated on that sample. The index b runs over the number of bootstrap samples B, which is often a large number, such as 1000 or 10,000, to ensure the stability of the resulting bootstrap distribution.

The bootstrap distribution of θ is then used to estimate various properties such as its mean, standard error, and bias. The standard error of θ, for instance, can be estimated using the standard deviation of the bootstrap replicates:

$$SE(\theta) \approx \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\theta_b^* - \bar{\theta}^*)^2} \tag{10}$$

where $\bar{\theta}^*$ is the mean of the bootstrap replicates θ_b^*.

For constructing a $(1-\alpha) \times 100\%$ confidence interval for θ, one can use the percentiles of the bootstrap replicates. For example, the lower and upper bounds of the confidence interval can be determined by the $\alpha/2$ and $1-\alpha/2$ quantiles of the bootstrap distribution, respectively:

$$CI_{(1-\alpha)\times 100\%} = \left(\theta_{(\alpha/2)}^*, \theta_{(1-\alpha/2)}^*\right) \tag{11}$$

It's important to note that bootstrap methods assume that the sample is representative of the population. If the original sample is biased, the bootstrap estimates will reflect that bias. Also, while bootstrap is versatile and applicable to a wide range of problems, it can be computationally intensive.

Regression and Bootstrap Analysis for House Price Prediction

In this analysis, we aim to predict house prices based on several features such as the size of the house in square meters, the number of bedrooms and bathrooms, and the year the house was built. We perform a regression analysis to estimate the relationship between these features and the house prices. Furthermore, we employ a bootstrap method to estimate the distribution of the regression coefficients and assess their variability.

The dataset for this analysis consists of simulated house prices with the following features: Size in square meters, number of bedrooms, number of bathrooms, and year built. The prices are simulated with added noise to reflect a more realistic scenario.

Table 5: Head of the simulated house data

Size (m^2)	Bedrooms	Bathrooms	Year Built	Price (USD)
161.70	2	4	1999	\$432,364.87
204.00	5	2	2009	\$473,637.06
168.98	3	1	1987	\$411,023.93
185.15	2	1	2020	\$425,185.60
131.01	1	2	2017	\$376,976.01

A pairplot provides a pairwise visualization of the relationships between the features and the price. It is useful for identifying patterns and potential correlations within the data.

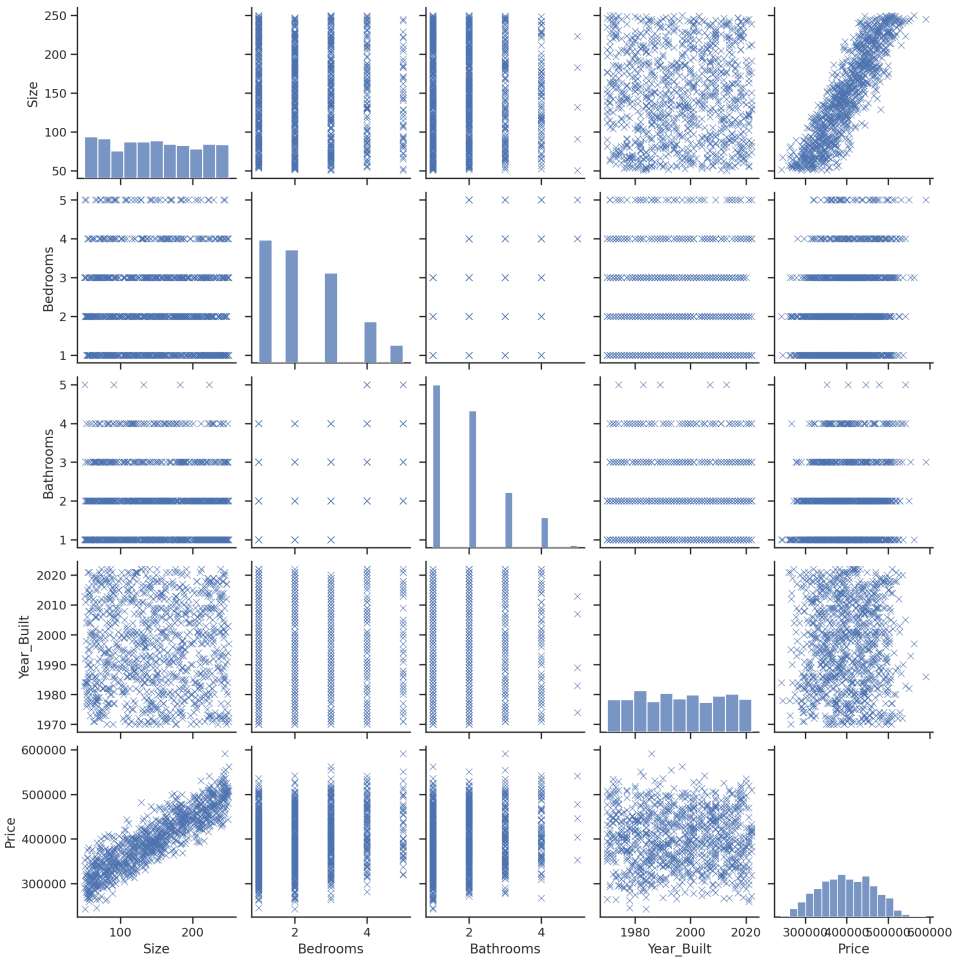

Figure 5: Pairwise relationships between house features and price

We performed a multivariate linear regression to predict house prices based on the features. The following coefficients were estimated:

- Size: 980.81 per square meter
- Bedrooms: 9970.05
- Bathrooms: 5821.01
- Year Built: 9.34 per year

The Mean Squared Error (MSE) of the regression model was found to be $637,014,199.88, indicating the average squared difference between the observed actual outcomes and the outcomes predicted by the model.

To assess the stability of our regression coefficients, we applied a bootstrap approach with 1000 resamples. This method provides a non-parametric estimation of the sampling distribution of the estimator by resampling with replacement from the original data and recalculating the estimator for each resample.

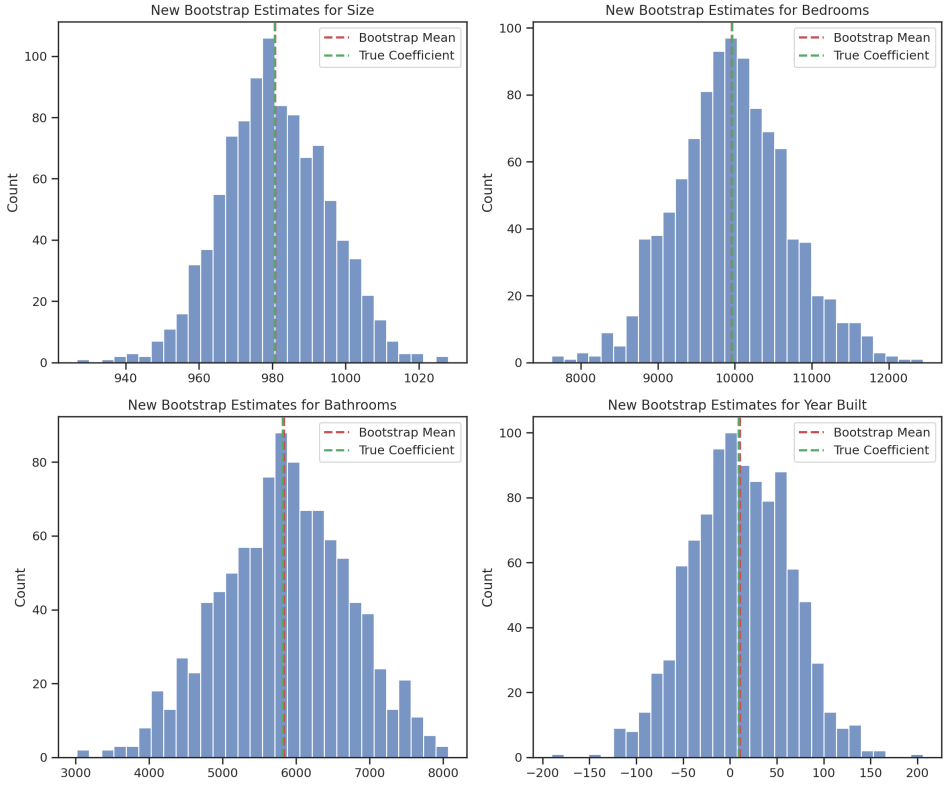

Figure 6: Bootstrap distributions of regression coefficients with mean estimates

The mean bootstrap estimates closely matched the original regression coefficients, which suggests that our model is relatively stable. The red dashed lines in Figure **??** represent the mean of the bootstrap estimates, while the green dashed lines indicate the estimated coefficients from the original regression.

The regression analysis suggests that size, number of bedrooms, and number of bathrooms are significant predictors of house prices, while the year built has a relatively small effect. The bootstrap analysis confirms the stability of these coefficients, providing confidence in the robustness of the model.