# CS573 - Optimization Methods

Fall Semester 2023
Assignment 2

*Papageridis Vasileios - 4710*
*csd4710@csd.uoc.gr*

January 31, 2024

## Introduction

This assignment presents an in-depth exploration of regression analysis, starting with the theoretical constructs that support the Unconstrained Least Squares Regression method. We will use the *Boston Housing* dataset to see how different factors, like crime rates or taxes, can affect house prices. The first part of our work is all about making a regression model from scratch, where we pay close attention to how each piece of data can change the predictions we make about house prices. The first part of our work is all about making a regression model from scratch, where we pay close attention to how each piece of data can change the predictions we make about house prices.

Next, we move on to Lasso Regression. This method is a step up from the basic linear regression as it incorporates a technique known as L1 regularization. The key advantage of Lasso Regression lies in its ability to simplify our model by prioritizing the most impactful variables from our dataset. This process of feature selection is critical, especially when dealing with datasets that have many variables, as it helps in creating a more efficient model. We'll be manually implementing Lasso Regression to observe its approach to handling data and to understand how it compares to the basic least squares model of exercise 1.

Finally, we tackle a common problem in data analysis: missing data. We intentionally leave out some data and watch how it messes with our models. This helps us learn how strong our models are and what happens when they don't have all the information they need. This exercise not only tests the robustness of our regression techniques but also sets the stage for the deployment of matrix completion strategies. The attempt of data imputation through matrix completion underpins the final analysis, showcasing the method's efficiency in restoring integrity to the dataset and re-calibrating our regression models for optimum performance.

Throughout this assignment, we aim to draw meaningful conclusions about the behavior of these regression techniques under various conditions.

Following this introduction, we will detail the methodology adopted in each exercise, present our findings, and discuss the implications of our results in the broader context of data-driven modeling and decision-making.

## Exercise 1: Unconstrained Least Squares Regression

This exercise involves the application of the Unconstrained Least Squares Regression method to the Boston Housing Dataset. The dataset contains various features that are used to predict the median value of houses within the Boston area.

### Dataset Description

The Boston Housing Dataset comprises information on 506 housing entities, with the following 14 features:

1. **CRIM** - Per capita crime rate by town.
2. **ZN** - Proportion of residential land zoned for lots over 25,000 sq.ft.
3. **INDUS** - Proportion of non-retail business acres per town.
4. **CHAS** - Charles River dummy variable (1 if tract bounds river; 0 otherwise).
5. **NOX** - Nitric oxides concentration (parts per 10 million).
6. **RM** - Average number of rooms per dwelling.
7. **AGE** - Proportion of owner-occupied units built prior to 1940.
8. **DIS** - Weighted distances to five Boston employment centres.
9. **RAD** - Index of accessibility to radial highways.
10. **TAX** - Full-value property-tax rate per $10,000.
11. **PTRATIO** - Pupil-teacher ratio by town.
12. **B** - $1000(Bk - 0.63)^2$ where $Bk$ is the proportion of black population by town.
13. **LSTAT** - $\%$ lower status of the population.
14. **MEDV** - Median value of owner-occupied homes in $1000's (the target variable).

An important aspect to note is the dataset's completeness, as it contains no null values. This integrity is crucial for ensuring accurate and reliable predictive modeling. Without missing data, we can proceed with our analysis with confidence in the dataset's quality, allowing us to focus purely on the regression modeling.

## Exploratory Data Analysis

In this exercise, we explore the distribution of the features of the Boston Housing Dataset through histograms, which provide a visual representation of the data. The following insights have been gathered:

- **CRIM:** The per capita crime rate by town is predominantly low, but with some towns experiencing higher crime rates, indicated by a long tail to the right.
- **ZN:** Residential land zoned for large lots over 25,000 sq.ft. is limited, suggesting that such zoning is not common.
- **INDUS:** Proportion of non-retail business acres per town is also right-skewed, with most towns having a lower industrial presence.
- **CHAS:** The Charles River dummy variable indicates that few tracts border the river.
- **NOX:** Nitric oxides concentration is somewhat normally distributed, with one particular concentration being most common.
- **RM:** Average number of rooms per dwelling appears to be normally distributed, indicating a variety of dwelling sizes.
- **AGE:** There is a high proportion of owner-occupied units built prior to 1940, with a decreasing number of newer units.
- **DIS:** Weighted distances to employment centers decay, implying closer proximity for most houses.
- **RAD:** Accessibility to radial highways is bimodal, reflecting differences in town accessibility.
- **TAX:** Full-value property-tax rate distribution is bimodal, which may indicate different tax policies.
- **PTRATIO:** The pupil-teacher ratio by town is multimodal, showing diversity in educational resources.
- **B:** The proportion of black population indicates potential segregation in certain areas.
- **LSTAT:** The percentage of lower-status population is right-skewed, meaning higher-status residents predominate.
- **MEDV:** Median value of owner-occupied homes is normally distributed, skewed to the right, suggesting that more towns have lower median home values.

The histogram plots for each feature are shown below, providing a comprehensive view of the data distribution. These visualizations are instrumental in understanding the dataset's characteristics and guiding the subsequent modeling efforts.

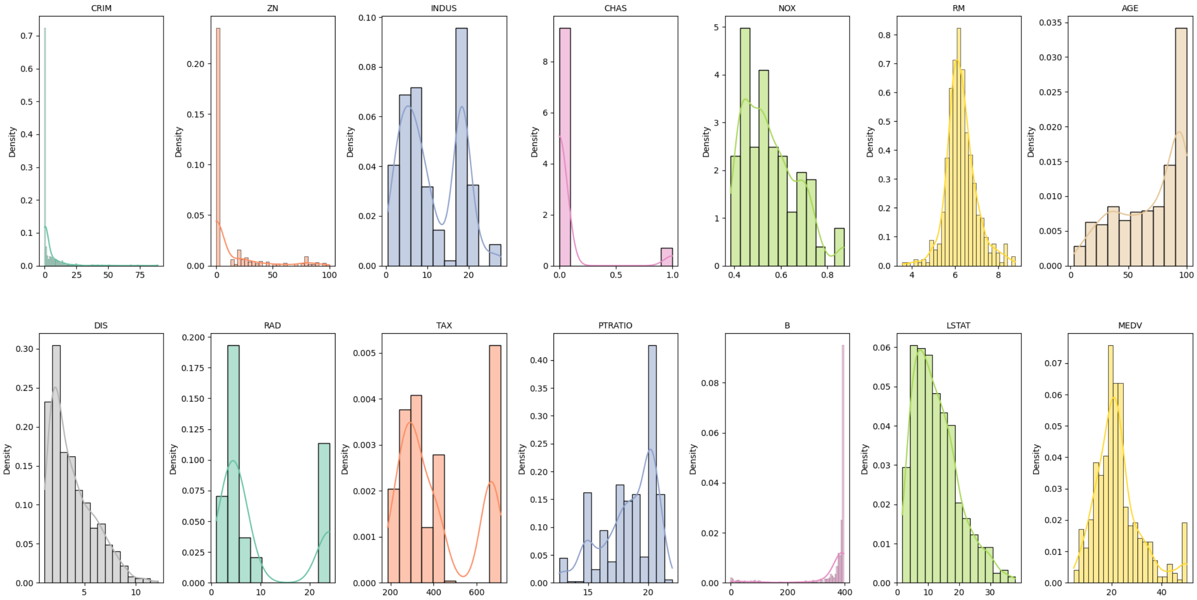

**Figure 1:** Histograms showing the distribution of each feature in the Boston Housing Dataset.

In the initial phase of our analysis, we visualized the distribution of features within the Boston Housing Dataset. Our objective was to comprehend the average scale of each feature and to observe any notable patterns that could influence the regression model.

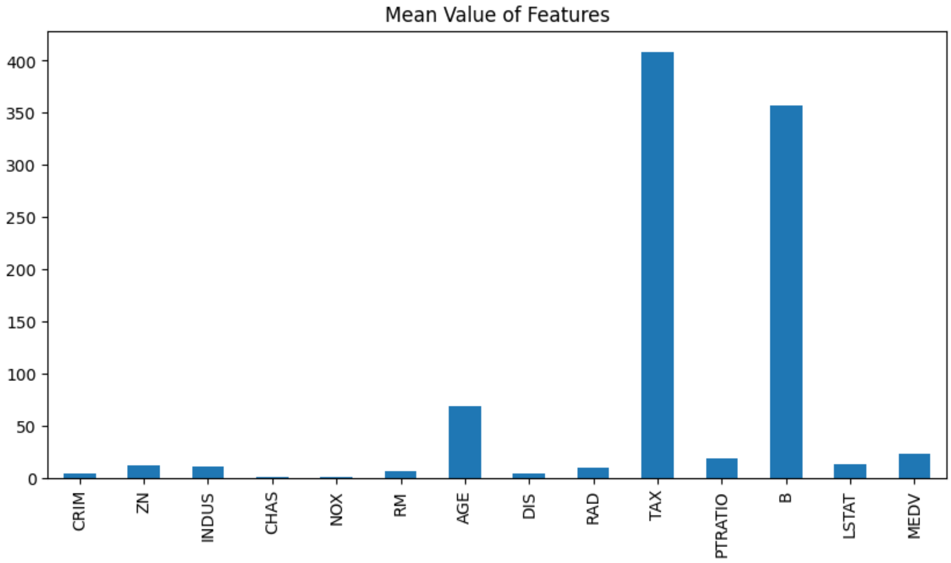

**Figure 2:** Bar plot showing the mean values of features in the Boston Housing Dataset.

**Insights from Mean Value Bar Plot:** The bar plot depicting the mean values of the features reveals several interesting aspects:

- The property-tax rate (**TAX**) and the proportion of black residents (**B**) exhibit the highest mean values, suggesting significant property taxation and a high representation of black residents in the towns included in the dataset.

- Conversely, the per capita crime rate (**CRIM**) and the proportion of residential land zoned for large lots (**ZN**) have low mean values, indicating a generally low crime rate and a lesser prevalence of large residential lots.

- The average number of rooms per dwelling (**RM**) portrays moderately-sized housing, which could be indicative of middle-sized family units being prevalent.

- Mean values for weighted distances to employment centers (**DIS**) and the proportion of pre-1940s owner-occupied units (**AGE**) are on the lower side, hinting at closer proximities to workplaces and a smaller share of older housing stock.

- The pupil-teacher ratio (**PTRATIO**) stands out with one of the higher mean values, pointing to potentially larger class sizes in the local schools.

- The percentage of lower status population (**LSTAT**) and median home values (**MEDV**) show higher mean values but not as pronounced as **TAX** and **B**, possibly reflecting a moderate socioeconomic status and housing value within the dataset.

Following the initial evaluation of mean values, we performed a boxplot analysis for a more detailed examination of the variables **ZN** and **INDUS**. The boxplots provide a visual summary of the distributions, highlighting the median, quartiles, and potential outliers.

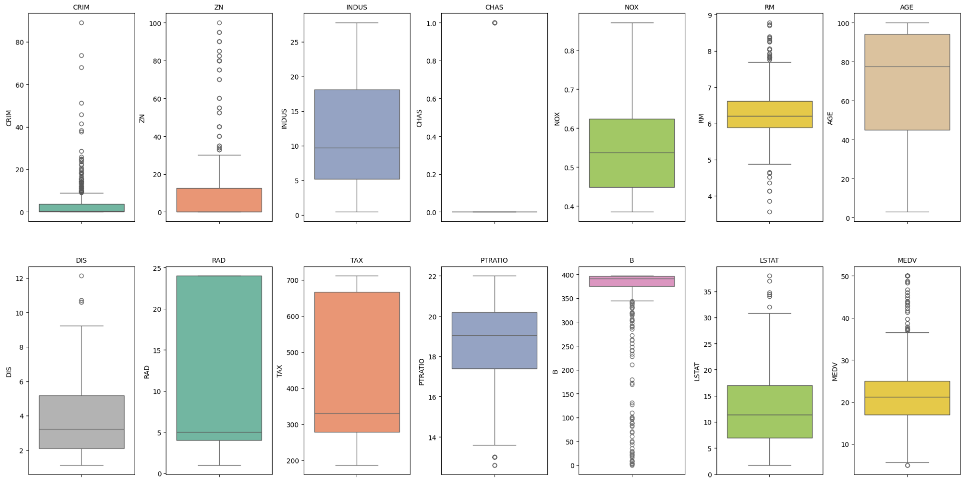

**Figure 3:** Boxplots illustrating the distributions of the **ZN** and **INDUS** features.

**Insights from Boxplot Analysis:**

- The boxplot for **ZN** indicates that a substantial number of observations fall into the lower quartile, with the median close to the bottom of the data range. This reflects a dominance of zones with smaller residential land areas over 25,000 sq.ft.

- Outliers are present in the **ZN** variable, showcasing the existence of areas with significantly larger residential land zoning, although they are relatively few.

- The **INDUS** feature's boxplot displays a similar concentration of values in the lower half of the range, implying that industrial business acres are less prevalent in the towns. The upper quartile extends further, showing a more uniform spread and indicating that some towns have a higher proportion of non-retail business acres.

These observations are critical as they provide a nuanced understanding of the residential and industrial composition of the areas within our dataset. The skewness and spread of values could significantly influence the pricing model of houses, which we shall explore in our regression analysis.

**Correlation Analysis of Features** The correlation heatmap is an essential tool for understanding the interdependencies among the various features in the Boston Housing Dataset. By examining the correlation coefficients, we can infer the strength and direction of the relationship between the features and the target variable, MEDV (Median value of owner-occupied homes in $1000's).

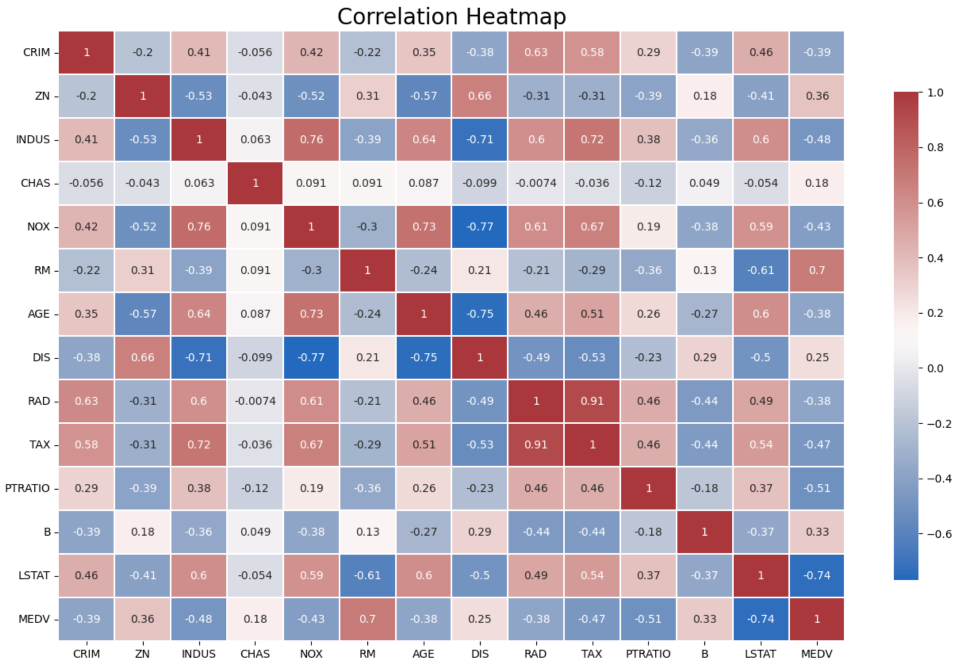

**Figure 4:** Correlation heatmap showing the relationship between features and the target variable in the Boston Housing Dataset.

**Significant Positive Correlation:**

- The number of **rooms per dwelling (RM)** has a strong positive correlation with the median value of homes ($\rho = 0.695$). This suggests that houses with more rooms are typically more valuable, which is consistent with the general understanding that larger homes command higher prices.

**Significant Negative Correlations:**

- The **pupil-teacher ratio by town (PTRATIO)** shows a significant negative correlation with MEDV ($\rho = -0.508$), indicating that higher values of PTRATIO (which can be a proxy for school overcrowding or lower educational resources) are associated with lower house values.

- The percentage of lower status population (**LSTAT**) is also strongly inversely related to MEDV ($\rho = -0.738$). A higher percentage of lower status population within the town is likely to decrease the median value of the homes, possibly due to socio-economic factors.

These insights reveal that both the size and quality of homes (represented by RM), as well as the socio-economic conditions of the neighborhood (captured by LSTAT and PTRATIO), significantly impact house prices. It is evident from the heatmap that as the quality of education and socio-economic status decline, the value of houses in the area tends to decrease as well. This correlation analysis is instrumental in guiding feature selection for our regression models, where we aim to predict house prices with higher accuracy.

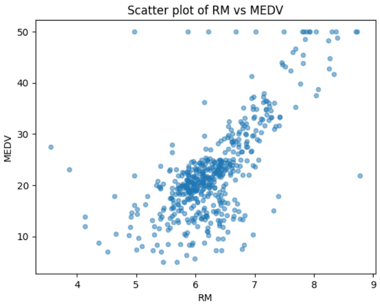

**Figure 5:** RM vs MEDV

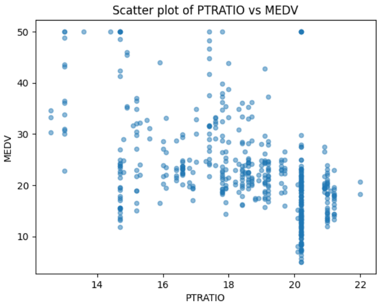

**Figure 6:** PTRATIO vs MEDV

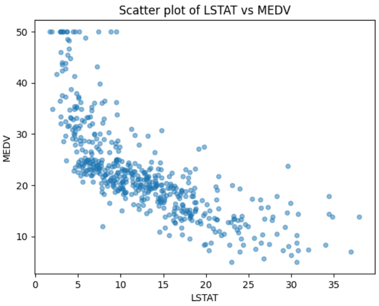

**Figure 7:** LSTAT vs MEDV

**Scatter Plots Analysis** The scatter plots visualize the relationships between the most positively and negatively correlated features with the target variable *MEDV*. These relationships are pivotal in validating the linearity assumption essential for linear regression.

- **RM vs MEDV:** The plot indicates a positive correlation, suggesting that homes with more rooms generally command higher median values. The data points exhibit a clear upward trend, reinforcing the strength of the linear relationship between room number and house value.

- **PTRATIO vs MEDV:** The scatter plot for pupil-teacher ratio does not display a distinct linear pattern, suggesting that *PTRATIO* may not be a strong standalone predictor for house value. The dispersion of data points indicates a need for additional contextual features for better prediction.

- **LSTAT vs MEDV:** There is a strong negative correlation observed, where areas with a higher percentage of lower-status population tend to have lower median house values. The downward trend in the plot is indicative of the inverse relationship between socioeconomic status and housing price.

These plots underscore the importance of feature selection in predictive modeling and highlight the necessity of considering linear relationships when building regression models.

## Developing a Least Squares Regression Model without Regularization

In this section, we focus on the development of a Least Squares Regression Model without any regularization. The mathematical formulation of the model is based on the Normal Equation, which is derived from the minimization of the sum of squared residuals. The regression coefficients are calculated as follows:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Where:

- $X$ is the feature matrix with an added intercept term (a column of ones).
- $X^T$ represents the transpose of the feature matrix.
- $(X^T X)^{-1}$ is the inverse of the matrix product of $X^T$ and $X$.
- $y$ is the vector of the target variable.

Predictions are made by the dot product of the feature matrix and the coefficient vector, $\hat{y} = X\hat{\beta}$. The dataset is randomly split into training and testing sets, with 80% of the data used for training the model.

## Observations from the Least Squares Regression Model

The implementation of the Least Squares Regression Model yields several insights:

- The Root Mean Square Error (RMSE) for our manually developed model is approximately $5.27$. This RMSE is consistent with the RMSE obtained from the ‘sklearn‘ library’s implementation of the Least Squares model, which suggests that our manual model performs on par with standard library implementations.
- Given that the RMSE reflects the median value of owner-occupied homes in thousands of dollars (MEDV), an RMSE of $5.27$ indicates that the average error from our model is approximately $5,270$, considering the scale of the target variable.

These observations indicate that our model is both accurate and robust, providing reliable predictions for the median value of homes based on the given features.

## Exercise 2: Lasso Regression

### Mathematical Foundation of the Custom Lasso Regression Model

The Lasso Regression model integrates L1 regularization into the linear regression framework. The objective function, which the model aims to minimize, is given by:

$$L(w, b) = \left( \sum_{i=1}^{n} \left( y^{(i)} - \left( w \cdot x^{(i)} + b \right) \right)^2 \right) + \lambda \sum_{j=1}^{d} |w_j| \tag{1}$$

where:

- $w$ denotes the weight vector.
- $b$ represents the bias term.
- $x^{(i)}$ and $y^{(i)}$ are the feature vector and the target value for the $i$-th example, respectively.
- $n$ is the total number of training examples.
- $d$ is the dimensionality of the feature space.
- $\lambda$ is the regularization parameter controlling the magnitude of the L1 penalty.

### Coordinate Descent Algorithm for Lasso Regression

The coordinate descent algorithm offers a strategic approach to optimizing the Lasso Regression model by iteratively updating each weight. The steps are as follows:

1. Calculate the partial residual for each feature $j$, excluding its current contribution from the overall model prediction.
2. Determine the correlation $\rho_j$ between the calculated partial residual and the feature $j$ itself.
3. Employ the soft-thresholding operator to $\rho_j$ with $\lambda$ as the threshold parameter to obtain the updated weight $w_j$.

The soft-thresholding operator is mathematically expressed as:

$$S(\rho, \lambda) = \text{sign}(\rho) \cdot \max(|\rho| - \lambda, 0) \tag{2}$$

The update rule for $w_j$ is formalized by:

$$w_j = \frac{S(\rho_j, \lambda)}{z_j} \tag{3}$$

where $z_j$ is the sum of the squares of all elements in feature column $j$. This iterative process continues for each weight until convergence is achieved, effectively minimizing the loss function while applying the L1 regularization to enhance model sparsity and feature selection. The implementation of the model can be found in the **Jupyter Notebook** at the second section. There is a comprehensive explanation about how the code is implemented based on the method above.

## Visualization of Lasso Model

The bar graph demonstrates the influence of each feature on the model's predictions. Features such as RM (average number of rooms per dwelling) and CHAS (Charles River dummy variable) exhibit the largest coefficients, indicating they are significant predictors in the dataset. The lasso regression inherently performs feature selection, as indicated by features with zero or near-zero coefficients being ”selected out”. This property enhances the interpretability of the model, enabling a straightforward understanding of feature importance.

Positive coefficient values suggest a direct relationship with the target variable, while negative values suggest an inverse relationship. The bar graph thus serves as a visual tool to understand feature relevance and the lasso model's decision-making process.

## Residuals Plot Analysis

The residuals plot is an essential diagnostic tool. It illustrates the discrepancies between the observed values and those predicted by the lasso regression model. The dashed line at zero signifies the ideal match between predicted and actual values. The scattered distribution of residuals around this line indicates the model's errors lack systematic bias.

The random spread of residuals suggests that the model assumptions hold true. However, the presence of outliers indicates potential model limitations. Homoscedasticity is a desired characteristic in such plots; any apparent patterns may suggest the need for model refinement. The absence of distinct patterns in the residuals plot indicates the robustness of the model, yet the outliers hint at possibilities for further improvement.

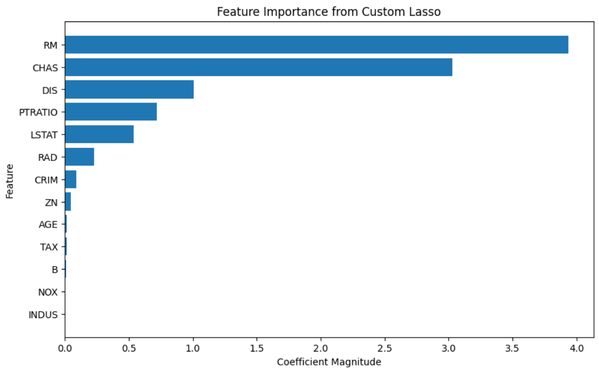

**Figure 8:** Feature Importance from Lasso

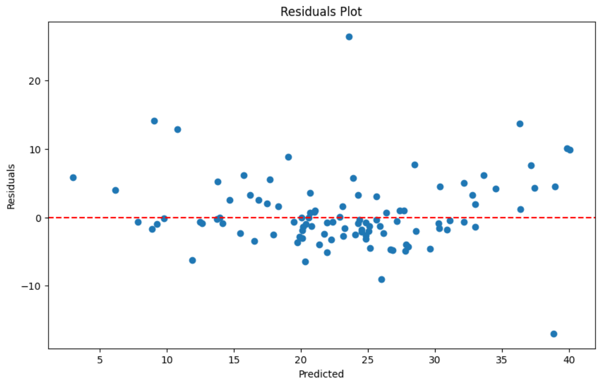

**Figure 9:** Residuals plot from Lasso

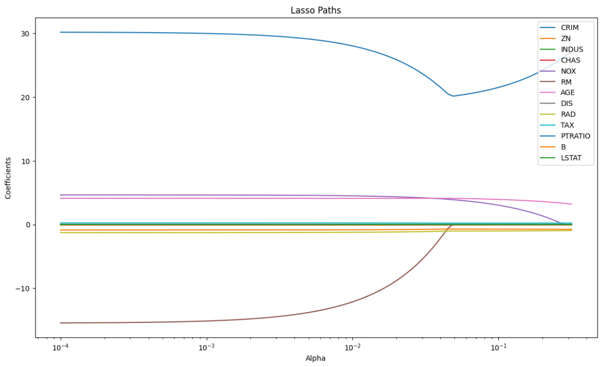

**Figure 10:** Lasso Path 1

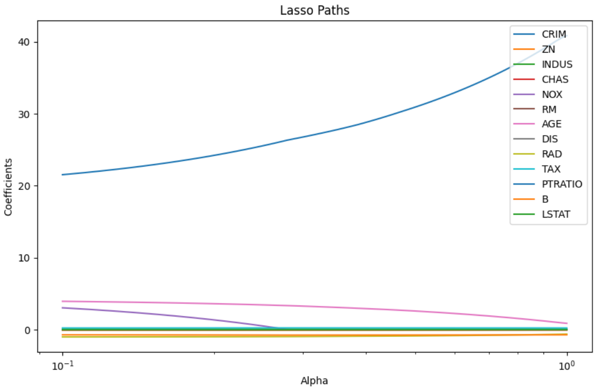

**Figure 11:** Lasso Path 2

## Analysis of Lasso Paths

The Lasso Paths graphically represent the evolution of the Lasso regression model's coefficients across various levels of the regularization parameter, denoted as $\alpha$. Each trajectory in the plot is associated with a specific attribute from the dataset, illustrating the modification of its coefficient in response to the incremental changes in $\alpha$.

### Coefficient Contraction

With an increase in $\alpha$ (leftward movement along the x-axis), we observe a diminution in the absolute value of the coefficients. This phenomenon is indicative of Lasso regression's inherent property where a surge in regularization magnitude precipitates a pronounced contraction of the coefficients.

### Attribution Selection

Intersecting the horizontal axis at the origin, the trajectories signify the attributes whose influence has been nullified by the model. The intersection point corresponds to the threshold $\alpha$ beyond which the attribute is deemed inconsequential by the model.

### Predominant Attributes

Attributes retaining non-zero coefficients over an extensive array of $\alpha$ values are deemed pivotal within the model framework. Notably, the attribute represented by the prominent blue trajectory (presumably 'RM', denoting the average number of rooms) remains influential throughout, maintaining a substantial and non-zero coefficient even at elevated $\alpha$ levels.

### Attributes of Lesser Significance

In contrast, attributes whose coefficients swiftly approach zero as $\alpha$ escalates are considered of minor significance. For example, the trajectory in brown (potentially signifying 'LSTAT', the proportion of lower socioeconomic status) diminishes markedly with increasing $\alpha$, highlighting its reduced stability as a predictor under regularization.

### Interpretative Nuance of the Alpha Axis

The logarithmic scale of the x-axis facilitates an examination of coefficient trends over a broad spectrum of $\alpha$ values. Diminutive values of $\alpha$‘ reflect minimal regularization, aligning the Lasso model closely with traditional linear regression. Conversely, substantial values signify intensified regularization, cultivating model sparsity by relegating more coefficients to insignificance.

### Practical Considerations

The visualization serves as a strategic tool in selecting an optimal $\alpha$ for the Lasso model. Ideally, one might opt for an $\alpha$ that strikes a balance between model parsimony and the retention of significant predictors, ideally situated just prior to the convergence of critical feature coefficients towards zero.
Cross-validation of $\alpha$ selection is paramount to avert overfitting, with a preference for the $\alpha$ yielding the minimal cross-validated RMSE or a comparable metric of error.
Moreover, the coefficient trajectories relative to differing $\alpha$ values yield insights into the dataset's structure, suggesting that attributes resistant to regularization may possess a more pronounced or direct correlation with the dependent variable.

## Comparison of the Lasso Regression Model with the Least Squares Regression Model

Within the framework of the Boston Housing dataset analysis, we assessed the performance of both the Least Squares and the Lasso regression models. The principal metric for comparison is the Root Mean Squared Error (RMSE) on the test data, providing a measure of the predictive accuracy for each model.

### Model Performance

- **Sklearn Least Squares RMSE:** 5.2753
- **Sklearn Lasso RMSE:** 5.2808

The proximate RMSE values suggest that both models deliver comparable predictive precision on this dataset.

### Feature Selection

- **Least Squares:** This model incorporates all available features, assuming each contributes value. While this can be effective when features are relevant and collinearity is minimal, it may complicate the interpretability when dealing with extensive feature sets.
- **Lasso:** By integrating an L1 penalty, the Lasso model inherently performs feature selection, which can simplify the model by nullifying coefficients of less contributory features.

### Interpretability

- **Least Squares:** Utilizing all features without selection can lead to a complex model, particularly when the dataset includes a substantial number of predictors.
- **Lasso:** The sparsity induced by the Lasso model enhances interpretability, elucidating significant predictive factors by excluding non-contributory variables.

### Conclusion

In this analysis, the Least Squares and Lasso models have demonstrated very similar performance in terms of RMSE. The choice between the two would then be influenced by the preference for model simplicity or the need for feature selection. If interpretability and understanding which features are most influential is important, Lasso may be the preferred model despite the negligible increase in RMSE. However, if the primary goal is prediction with all available information, and all features are assumed relevant, the Least Squares model is just as suitable.

## Exercise 3: Handling Missing Data

In this exercise the goal is to handle missing data. In order to achieve the goal of the exercise we will have first to introduce missing values in the dataset. We implement a function that will add a specific percentage of missing values in our dataset, so we can observe how different percentages of missing data affect our models accuracy.

Without prejudice to generality, we systematically introduce missing values in our dataset. Specifically we will introduce $25\%$, $50\%$ and $75\%$ missing values.

### Analysis of Missing Data on Regression Models

An investigation into the impact of missing data on regression models was conducted by systematically introducing missing values into the dataset. The performance of both Lasso and Least Squares regression models was evaluated, yielding the following insights:

- **Error Amplification with Increased Missing Data:** There is a noticeable escalation in the Root Mean Squared Error (RMSE) for both models as the missing data percentage augmented from 25% to $75\%$. This trend indicates a decrement in model accuracy correlating with the volume of missing data.
- **Model Robustness Comparison:** The Least Squares model consistently exhibited a marginally lower RMSE across various levels of missing data, suggesting a potential higher resilience compared to the Lasso model, which may be influenced by the dataset's characteristics and the missing data pattern.
- **Influence of Lasso Regularization:** Lasso's intrinsic regularization, aimed at mitigating overfitting, may inadvertently lead to the exclusion of pertinent predictors amid data scarcity, particularly if the missingness is not entirely random.
- **Interpretability Impairment:** As missing data proliferates, the interpretability of models is compromised. Lasso regression, in particular, might erroneously eliminate significant predictors, diminishing its interpretability.
- **Implications of Imputation Methodology:** The approach to data imputation significantly affects model efficacy. The simplistic mean imputation strategy employed here may necessitate enhancement, especially with increasing levels of missing data, to preclude bias induction.
- **Practical Considerations:** These observations highlight the criticality of advanced imputation techniques or models proficient in coping with substantial missing data. Comprehending the underlying mechanism of missingness is vital to ensure the appropriateness of the imputation method.

Conclusively, the findings accentuate the importance of meticulous missing data management and the selection of fitting imputation techniques, which bear heightened significance as missing data accumulates.

## Exercise 4: Matrix Completion

The approach of the matrix completion algorithm we created is based on the lecture notes. In this section we will explain how the function in the $JupyterNotebook$ is implemented, in order to achieve good results in matrix completion after the insertion of missing values in our dataset.

### Matrix Completion via Nuclear Norm Minimization

Given a matrix $Y \in \mathbb{R}^{m\times n}$, and only observed entries $Y_{ij}, (i,j) \in \Omega$. The goal of matrix completion is to fill in the missing entries. This can be formulated as the following optimization problem:

$$\min_B \frac{1}{2} \sum_{(i,j)\in\Omega} (Y_{ij} - B_{ij})^2 + \lambda \|B\|_{tr}$$

where $\|B\|_{tr}$ is the trace (or nuclear) norm of $B$, which is the sum of its singular values:

$$\|B\|_{tr} = \sum_{i=1}^{r} \sigma_i(B)$$

with $r = \text{rank}(B)$ and $\sigma_1(X) \geq \ldots \geq \sigma_r(X) \geq 0$ as the singular values.

### Proximal Gradient for Matrix Completion

The projection operator onto the observed set is defined by:

$$[P_\Omega(B)]_{ij} = \begin{cases} B_{ij} & \text{if } (i,j) \in \Omega \\ 0 & \text{if } (i,j) \notin \Omega \end{cases}$$

The optimization criterion is then:

$$f(B) = \frac{1}{2}\|P_\Omega(Y) - P_\Omega(B)\|_F^2 + \lambda\|B\|_{tr}$$

To solve this optimization problem, the following algorithm is used, which employs iterative applications of Singular Value Thresholding (SVT) to approximate the missing values. The soft-thresholding function $S_\lambda(x)$ and SVT operation are defined as follows:

$$S_\lambda(x) = \text{sign}(x) \cdot \max(|x| - \lambda, 0)$$

$$\text{SVT}(X) = U \cdot \text{diag}(\text{soft\_threshold}(S)) \cdot V^T$$

after performing Singular Value Decomposition of $X$. This process is iterated until convergence is achieved or the maximum number of iterations is reached.

**Algorithm 1** Matrix Completion via SVT

```
function MATRIXCOMPLETION($X$, mask, $\lambda_{\text{reg}}$, max_iter, tol)
    Initialize $B$ as a copy of $X$
    for $i \leftarrow 1$ to max_iter do
        Create a combined matrix with current estimates from $B$
        $B \leftarrow \text{SVT}(\text{combined\_matrix})$
        if $\frac{\|X-B\|_F}{\|X\|_F} < \text{tol}$ then
            break
        end if
    end for
    return $B$
end function
```

## Exercise 5: Re-evaluation Using the Completed Dataset

We evaluated the performance of two regression models, Lasso and Least Squares, on our dataset under different conditions of missing data, and then reassessed their performance after applying a matrix completion process. The key metric for evaluation was the Root Mean Square Error (RMSE), which quantifies the difference between the predicted and actual values.

### RMSE Values With Different Levels of Missing Data

Initially, we observed the following RMSE values with different proportions of missing data:

- **With 50% Missing Data:**
    - Lasso Regression RMSE: 6.8429
    - Least Squares Regression RMSE: 6.8011
- **With 25% Missing Data:**
    - Lasso Regression RMSE: 5.5442
    - Least Squares Regression RMSE: 5.1593

These results indicate that as the proportion of missing data increases, the RMSE values also increase, reflecting a decrease in model accuracy. This outcome is expected since more missing data typically leads to a greater loss of information, making it more challenging for models to learn and predict accurately.

### RMSE Values After Matrix Completion

After applying matrix completion to impute missing values, we noted significant improvements in RMSE for both models:

- **After Matrix Completion (50% Missing Data):**
  - Least Squares Regression RMSE: 5.0139
  - Lasso Regression RMSE: 5.4401
- **After Matrix Completion (25% Missing Data):**
  - Least Squares Regression RMSE: 4.3089
  - Lasso Regression RMSE: 4.2296

These improved RMSE values post-matrix completion suggest that the imputation of missing values has a positive impact on model performance. The matrix completion process helped provide a more complete dataset, allowing the models to capture underlying patterns more effectively.

### Insights and Interpretation

- The **reduction in RMSE after matrix completion** demonstrates the effectiveness of the matrix completion approach in providing a better basis for predictive modeling. It suggests that the imputed values are reasonably accurate and help mitigate the impact of missing data.

- The **variation in RMSE improvement** between the two levels of missing data (50% vs. 25%) highlights the relationship between the amount of missing data and the effectiveness of the matrix completion. Generally, less missing data leads to a more accurate imputation and, consequently, better model performance.

- The **Lasso model shows a slightly larger improvement than Least Squares** in the 25% missing data scenario. This could be due to Lasso's inherent regularization, which might make it more adaptable to the changes introduced by the matrix completion.

### Caveats and Further Considerations

While the matrix completion method has shown promise in improving model accuracy, it's crucial to acknowledge that the imputed values are estimates. The nature of these estimates depends on the matrix completion method and the characteristics of the missing data. There might be underlying biases or inaccuracies introduced during imputation, especially if the missing data mechanism is complex. Therefore, additional validation techniques, such as cross-validation, are recommended to ensure the robustness and generalizability of the observed improvements.

## Conclusion

This assignment has delved into various regression techniques, applied to the Boston Housing dataset, to address different predictive modeling scenarios. We began with Unconstrained Least Squares Regression, establishing a baseline for predictive performance. Our implementation demonstrated comparable accuracy to standard industry tools, reinforcing the validity of classical methods in regression analysis.

Moving to Lasso Regression, we explored the benefits of regularization in feature selection and model interpretability. Lasso's ability to simplify the model by shrinking less relevant features' coefficients to zero proved valuable in enhancing model simplicity without significantly compromising accuracy.

Handling missing data presented a challenge, one that we approached by introducing systematic missingness and observing its impact on model robustness. It was evident that as missing data increased, so did the predictive error, underscoring the importance of proper data management and advanced imputation techniques.

The final piece of the assignment was Matrix Completion, where we aimed to reconstruct the incomplete matrix using Nuclear Norm Minimization.

Overall, the assignment highlighted the trade-offs between model complexity, interpretability, and predictive accuracy. It showcased the necessity for rigorous data preprocessing, the application of tailored modeling techniques, and the constant evaluation of model performance through appropriate metrics. These insights are really important in order to build reliable, understandable, and accurate predictive models in the face of real-world data challenges.