Assignment 1
A mathematical essay on linear regression.

Anik Bhowmick
Inter Disciplinary Dual Degree Data-Science
Indian Institute of Technology Madras
ae20b102@smail.iitm.ac.in

***Abstract*—This assignment is meant for Linear Regression on a dataset that contains information about cancer incidence and death cases across different states of the US for various population groups. This article demonstrates the feature selection technique, data visualizations linear model fitting and its evaluation.**

***Index Terms*—Feature selection, Visualization, Statistical Models**

I. Introduction

- This present assignment is on the prediction of cancer incidence and its mortality among various socio-economic groups such as poor, middle-class and rich, privileged, and unprivileged people from various states of the US. In this context, the application of Linear Regression is demonstrated along with various data-handling techniques
- Linear regression is a statistical modelling technique that tries to find a linear relationship between independent variables and a dependent variable. For the case of 2 variables, it finds a best-fit straight line that minimizes the sum of squared errors between the predicted value and the actual value. With this line, any value for unseen data points can be predicted with a good degree of confidence. For multiple variables case it is more like finding a hyperplane such that predicted values lying on it bears smaller error with true value. Several Machine Learning-based optimization techniques can minimize this error function such as gradient descent. Machine Learning makes these numerical computations feasible and faster. Several evaluation metrics can tell us whether a Regressor model is good or bad, one best example is the Coefficient of Determination or R^2. These will be discussed in a great deal in the next section.
- The main task in this assignment is to build a regressor model that finds the relation between socioeconomic features and cancer incidence and its mortality in the United States using a dataset named "merged_data.xlxs" given to us. For this, we have to use various feature selection and visualisation techniques and finally fit and validate the Regressor model.
- This paper will demonstrate the data analysis technique relevant to linear regression with the help of visual plots, and mathematical equations. In this process, various data handling processes such as handling non-numeric quantities, missing values, feature imputations etc. will be discussed.

II. Linear Regression

Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. When the number of the independent feature, is 1 then it is known as Univariate linear regression. In the case of more than one feature, it is known as multivariate linear regression. The goal of the algorithm is to find the best linear equation that can predict the value of the dependent variable based on the independent variables. The equation provides a straight line that represents the relationship between the dependent and independent variables.

A. Assumptions

- The independent and dependent variables have a linear relationship with one another.
- The observations in the dataset are independent of each other. This means that the value of the dependent variable for one observation does not depend on the value of the dependent variable for another observation.
- Across all levels of the independent variable(s), the variance of the errors is constant. This indicates that the amount of the independent variable(s) has no impact on the variance of the errors. This is called Homoscedasticity.
- The errors in the model are normally distributed.
- There is no high correlation between the independent variables. This indicates that there is little or no correlation between the independent variables.

B. Hypothesis function

As already mentioned the input features and target output has a linear relationship the hypothesis function for ith data point with n features can be given as:

$$\hat{y}^{(i)} = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + ... + \theta_n x_n^{(i)}$$

Where θs are the parameters that are determined by the linear regression minimizing the cost function, the term θ_0 is called bias. So basically linear regression is all about finding these parameters.

C. Cost Function

As mentioned earlier it is a supervised learning so we already have the truth value of the target variable available. This cost function is just a mean squared error between the true target variable and the predicted target variable. The squared error for the ith training example is given as

$$e_i^2 = (\hat{y}^{(i)} - y^{(i)})^2$$

Now for all N sample points, it is :

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{i=N}(\hat{y}^{(i)} - y^{(i)})^2$$

D. Gradient Descent

A linear regression model can be trained using the optimization algorithm gradient descent by iteratively modifying the model's parameters to reduce the mean squared error (MSE) of the model on a training dataset. Partially differentiate concerning θ

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{N}\sum_{i=1}^{i=N}(\hat{y}^{(i)} - y^{(i)})$$

$$\frac{\partial J}{\partial \theta_k} = \frac{1}{N}\sum_{i=1}^{i=N}(\hat{y}^{(i)} - y^{(i)})x_k \;\; k \neq 0$$

There is a non-trainable parameter called learning rate is used to update the parameters as follows

$$\theta_0 := \theta_0 - \alpha\frac{\partial J}{\partial \theta_0}$$

$$\theta_k := \theta_k - \alpha\frac{\partial J}{\partial \theta_k}$$

The choice of learning rate is what decides the convergence of the cost function. A larger learning rate helps faster convergence and a smaller one slows the convergence rate. But keeping very high learning causes parameters to blow out sometimes. So a typical value of this hyperparameter is around 0.01-0.0001. Again this depends on problems to problems and the choice of cost function as well.

E. Evaluation Metrics

After training a model its evaluation and validation are very important. For that, we usually use a validation/ test set. We apply the model to the input features and check how close the predicted value is to the actual value. In the context of linear regression, the best metric is the Coefficient of Determination or R^2. The formula for this function is given as

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where RSS is Residual Sum Squares and TSS is Total Sum Squares.

$$RSS = \sum_{i=1}^{i=N}(\hat{y}^{(i)} - y^{(i)})^2$$

$$TSS = \sum_{i=1}^{i=N}(y^{(i)} - \bar{y})^2$$

bar denotes the mean. The higher the R^2 score better the model. But there is no such fixed benchmark for it. It varies from problem to problem. Typically score above 0.8 suggests a good performing reliable model. But the R^2 score alone can't give complete reliability of the model. We also need to check the distribution of the error. As mentioned in the assumption it should have near-normal distribution. We have provided all these insights in our implementation part of the Problem section.

III. The Problem

As already mentioned our goal in this assignment is to predict death cases and mortality due to cancer for various population groups by linear regression with a primary focus on finding a correlation between incidence rate and mortality with socio-economic status.

A. Cleaning and preparing the data

The dataset had some missing as well as categorical entries. In either case, the linear regression model can not work. Access the dataset here

1) Handling missing values and categorical columns: Some of the columns of the given dataset had categorical features such as "State", and "AreaName". The "Incidence_Rate" had missing, categorical and numerical values all present. We first dropped the first two categorical features. The "_" and "__" of "Incidence_Rate" were dropped as we had no reference with what we should substitute them. These rows belong to certain states like "KS", "MN" and "NV". The same strategy is followed for the other few columns. The missing values of pure numerical columns as "Med_Income", "Med_Income_White", and "Med_Income_Asian" etc. were replaced with the respective column's mean values. Because they are undoubtedly continuous variables.

B. Exploratory analysis

1) Visualizing the correlation among features: To have the overview we obtained the correlation plot first. The formula for correlation between two features x_i and y_i is :

$$r_{xy} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_i x_i^2 - n\bar{x}^2}\,\sqrt{\sum_i y_i^2 - n\bar{y}^2}}.$$

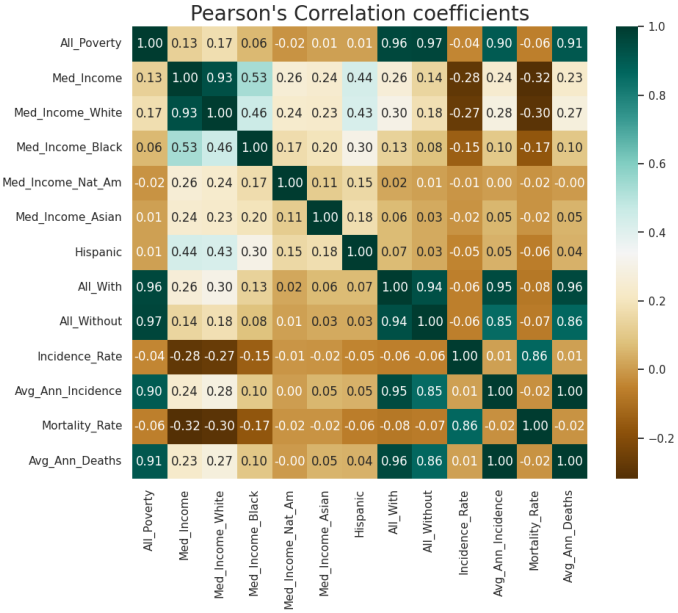

Fig. 1. Correlation plot for features

2) Extracting target columns: The above plot gives a strong hint that the ”Avg_Ann_Deaths” and ”Avg_Ann_Incidence” have a strong dependence on socioeconomic features. The other columns ”Incidence_Rate” and ”Mortality_Rate” do not strongly depend on the input features. So it’s useless to make predictions for them. Because the linear assumption might not hold here. We have checked them in our code.

C. Outliers detection

Below are box and distribution plots of a few columns

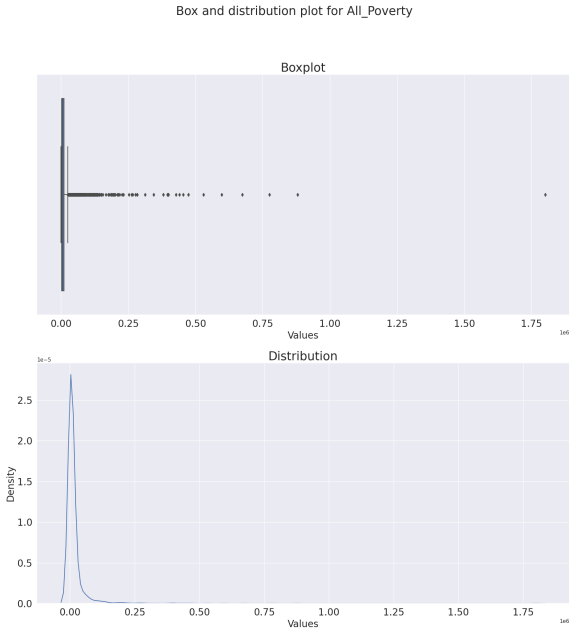

Fig. 2. Box and density plot of ”All_Poverty”

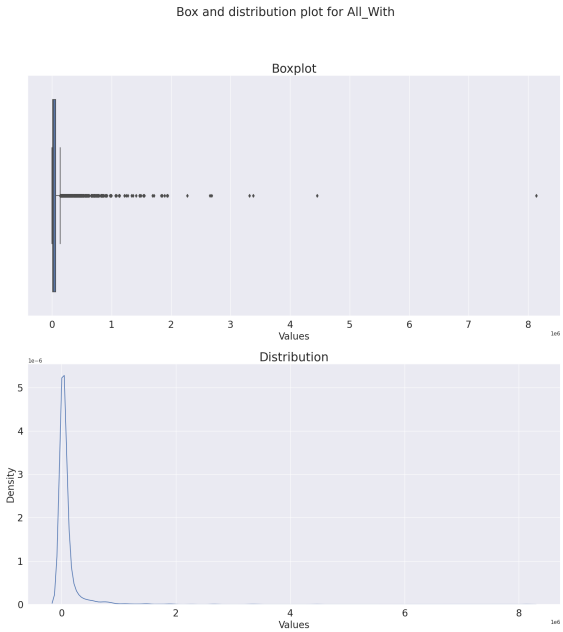

Fig. 3. Box and density plot of ”All_With”

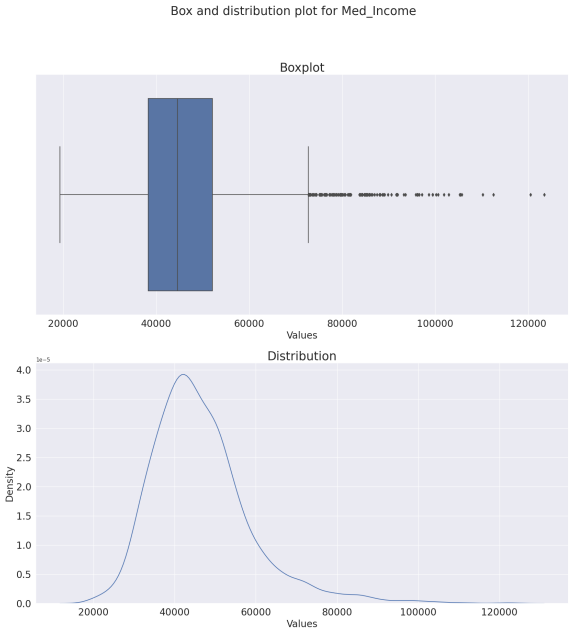

Fig. 4. Box and density plot of ”Med_Income”

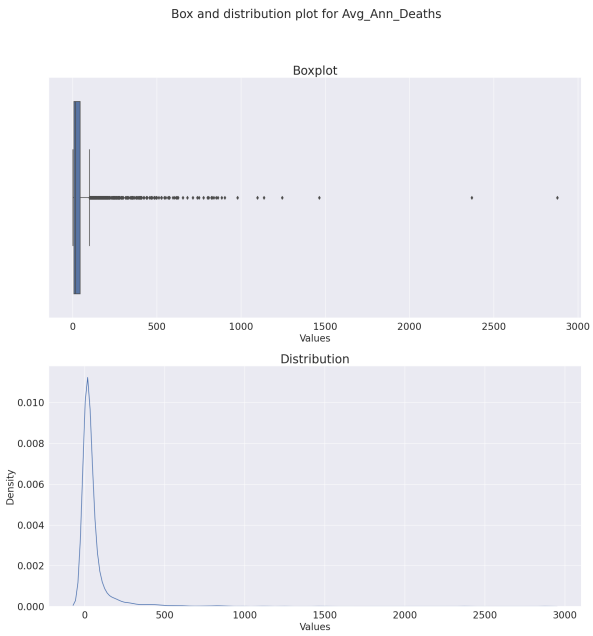

Fig. 5. Box and density plot of "Avg_Ann_Deaths"

In the code, we have seen almost all features have very high outliers. So we need to drop them based on some statistical method. In my case, I chose to calculate the Z-score and drop the samples with a Z-score higher than 1.5 or lying away from their mean above 1.5σ.

D. Final Correlation plot before training

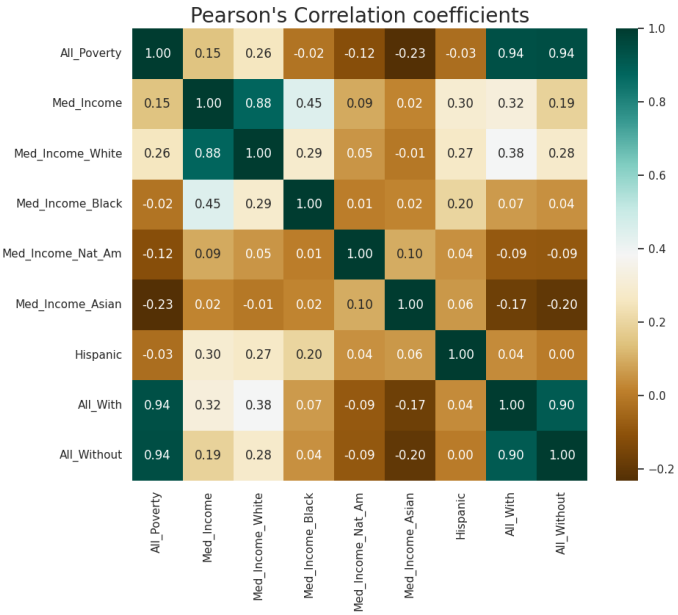

Fig. 6. Correlation plot for features

Now the correlation among the input features has sufficiently reduced. We will further drop a few more features after checking the feature importance.

E. Statistical model

As we guessed there are four target variables possible so we decided to make 4 models and check their effectiveness. Further, we used the sklearn library for linear regression to keep the codes simple.

1. Model 1 for Avg_Ann_Deaths

- The initial model took 9 input features and resulted in an R^2 score of 0.897.
- The final model was developed to take only one input feature "All_With" because using the feature importance method we found this feature among all socio-economic features is most important. The model resulted in obtaining the functional form as

$$\hat{y} = 35.41x + 31.61$$

2. Model 2 for Incidence_Rate

- This model performed ill as the given target doesn't carry a linear relation with the inputs. This gave a R^2 score of 0.1. Which is extremely low. This confirms that "Incidence_Rate" does not correlate with the socio-economic features.

3. Model 3 for Avg_Ann_Incidence

- This model performed well giving a final R^2 score of 0.882. Feature importance test here also confirms "All_With" is the most important feature. The functional form of this model is

$$\hat{y} = 50.29x + 42.59$$

4. Model 4 for Mortality_Rate

- This model gave R^2 of 0.19. Confirming this also has no relation with the inputs.

F. Visualization and validation

Model validation is the process that is carried out after Model Training where the trained model is evaluated with a testing data set. The ultimate goal for any machine learning model is to learn from examples in such a manner that the model is capable of generalizing the learning to new instances which it has not yet seen. Two different models using similar data can predict different results with different degrees of accuracy and hence model validation is required. For validation, we are using R^2 score and residual plots.

1. Model 1

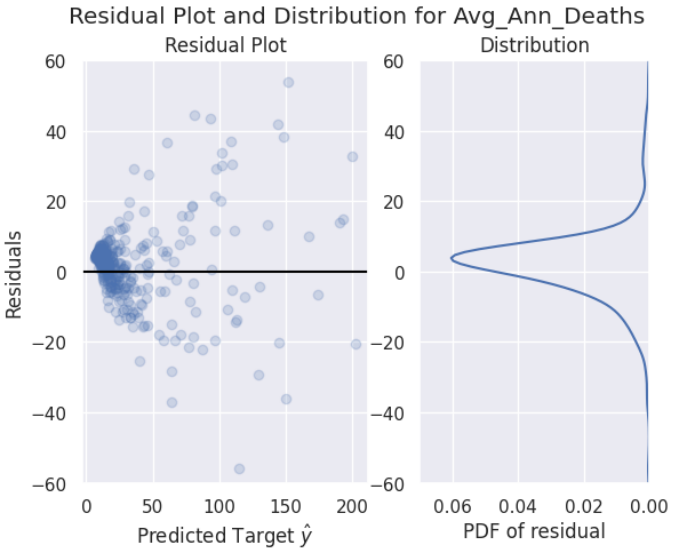

Fig. 7. Residual plot and its distribution

So this suggests the error between actual and predicted variables is normally distributed, which is essential for model reliability. Below is our best-fit predictor plot.

2. Model 3

This also clearly suggests the error between actual and predicted variables is normally distributed, making the model quite reliable.

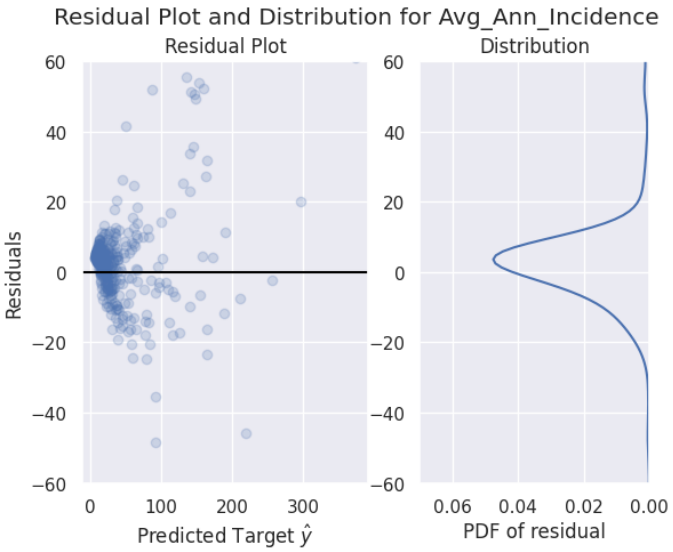

Fig. 8. Residual plot and its distribution

G. Answers to the primary goal

- The socioeconomic features primarily include income, education, employment, social status, social privileges etc. All the social features used for model training are mentioned in the correlation plot above Fig. 6. The dependence of the targets on these features is illustrated in correlation plot Fig. 1. From this correlation plot, this is evident that "Incidence_Rate" and "Mortality_rate" don't depend much on socioeconomic factors. Average Annual Cancer Incidence and Average Death cases due to cancer are strong functions of socioeconomic factors. So the Annual Cancer Incidence and Annual Mortality do have high correlations on socioeconomic features.
- Regarding the applicability of the model for the mission below are model plots with best fit lines for both output variables. Note we have only shown the "All_With" variable as input for best-fit visualization. Even without a line, the increasing trend is perceptible.

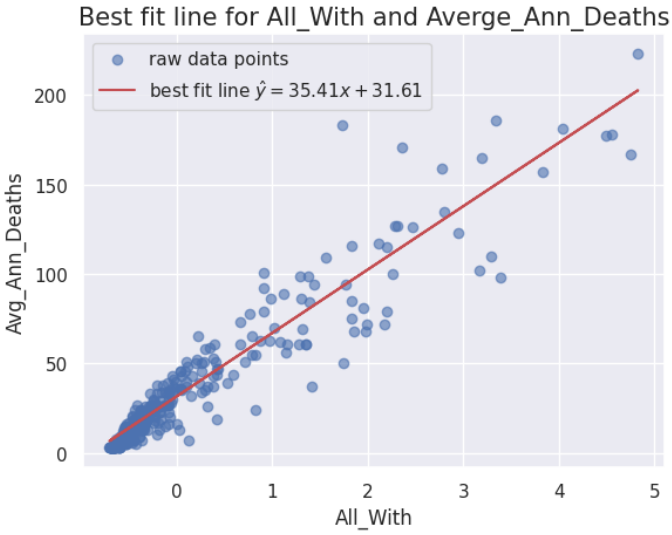

Fig. 9. Scattered plot with best-fit line (Model 1 output with single input)

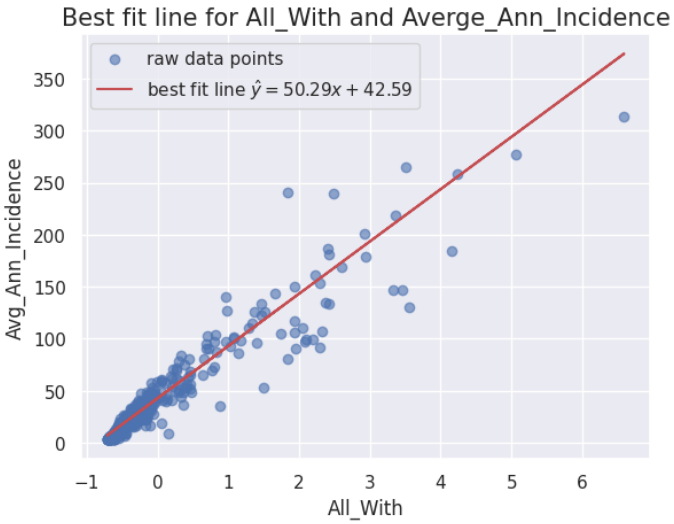

Fig. 10. Scattered plot with best-fit line (Model 3 output with single input)

As the number of individuals with medical insurance increases the incidence rate also increases. This may be attributed to the fact that the individuals with medical

insurance are privileged classes, so upon manifestation of symptoms they are quickly diagnosed and detected with cancer. For the death case, it might be because deaths with cancer are recorded for those individuals who came in contact with health care teams. So in this case these deaths are officially recorded and have been put in the dataset. (Note this is our intuition drawn from the model output.)

- In both models the line perfectly fits the data. Also comparatively higher R^2 score (0.897 for model 1 and 0.882 for model 2) suggests us the models are reliable even considering all the features. So undoubtedly the models can be released for productions and applications.

IV. Conclusions

Linear Regression is a powerful statistical method for predictive modelling when variables are continuous. In this assignment, we demonstrated the techniques of handling tabular data and fitting a linear model using linear regression. This assignment allowed us to explore the mathematical details of regression and other statistical techniques. But we have to keep in mind that linear regression assumes a linear nature of the relation between input and target. This may not be the case for all datasets. Even in our case, we got the two target columns "Incidence_Rate", and "Mortality_Rate" bearing no linear relationship with inputs. In these cases when we don't understand nature it's better to use Neural Networks because they are trained to capture the non-linear relations. In neural networks, we don't make any such functional assumption beforehand. But linear regression is no less in terms of its importance especially when the dataset is small enough and contains continuous variables.

References

[1] Linear Regression: https://www.geeksforgeeks.org/ml-linear-regression/
[2] Gradient Descent: https://builtin.com/data-science/gradient-descent
[3] Evaluation Metrics: https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/
[4] Residual plot: https://towardsdatascience.com/how-to-use-residual-plots-for-regression-model-validation-c3c70e8ab378
[5] Feature engineering: https://medium.com/analytics-vidhya/feature-engineering-part-1-imputation-techniques-eafce8f341bc
[6] scikit-learn: https://scikit-learn.org/stable/modules/linear_model.html

Access the original code here