Summer 2023

# A Linear Regression Model to Predict the Critical Temperature of a Superconductor

Amir Alipour Yengejeh
*University of Central Florida*, amir.alipouryengejeh@ucf.edu

# A linear regression model to predict the critical temperature of a superconductor

Amir Alipour Yengejeh
*dept. Statistics and Data Science*
*University of Central Florida*
Orlando, United States
amir.alipouryengejeh@ucf.edu

***Abstract*—Since the superconductivity has been introduced, almost all studies in this area have been striving to predict the critical temperature ($T_c$) through the features extracted from the superconductor's chemical formula. In this study, thus, we are interested in exploring the linear association between $T_c$ and the related features.**

***Index Terms*—Superconductor, Linear Regression, Critical Temperature**

## I. Introduction

Superconductor's materials enjoy the superconductivity state where there is not any electronic resistance and can store or preserve an electronic current for an unlimited time. These materials have a wide range of applications. For instance, it is used in MRI systems by the health care specialists to see the details of the internal status of the patients. However, the superconductivity in the superconductors can be achieved in the cold temperature. In other word, they need to be cooled at the below of their critical temperature ($T_c$) to let the superconductivity happen. Thus, the problem of predicting $T_c$ has been investigating for a long time. In this regard, it has been interested in building the statistical models to anticipate the critical temperature through the features derived from the chemical formula of superconducting materials. To address this problem, therefore, a comprehensive database has been introduced to let create verity of the statistical algorithms. The pre-processed dataset contains 21,263 superconductors and 82 features [1]. In this study, we eager to fit a linear regression analysis to predict $T_c$ via these dataset, that is, estimate a statistical linear relation between $T_c$ and 81 superconductors. Some of the related works predicting critical temperature of a superconductor are [2] and [3].
Therefore, the main contribution of this paper is building a multiple linear regression model to predict the critical temperature. The framework of this study begin with glancing on the dataset and then follows by presenting the analysis the obtained results. Finally, it will end with discussion and conclusion of the results.

## II. Data set

In this section, we like to glance on the dataset. As mentioned, our dataset is divided to the critical temperature as a target variable and the rest features as predictors. Here, our main plan is fitting a multiple linear regression to predict $T_c$. However, this algorithm requires the dataset to meet some critical assumptions. The main assumption defined on the target variable in which the distribution is from the normal. However, the distribution of the predictors has less effect on the regression models. On the other hand, the number of predictors fed in the linear algorithm is the main concern. This is because that as the number of predictors goes up, the probability of co-linearity among some predictors might increase . Hence, these concerns have to be explored individually in our dataset before diving into building the model.

### A. Critical Temperature Distribution

As mentioned, the normality assumption for the response variable is essential for the linear regression analysis. To check this assumption in our case, figure 1 displays the density (histogram) plot of the critical temperature vs the normal distribution generated by mean and standard deviation of the variable.

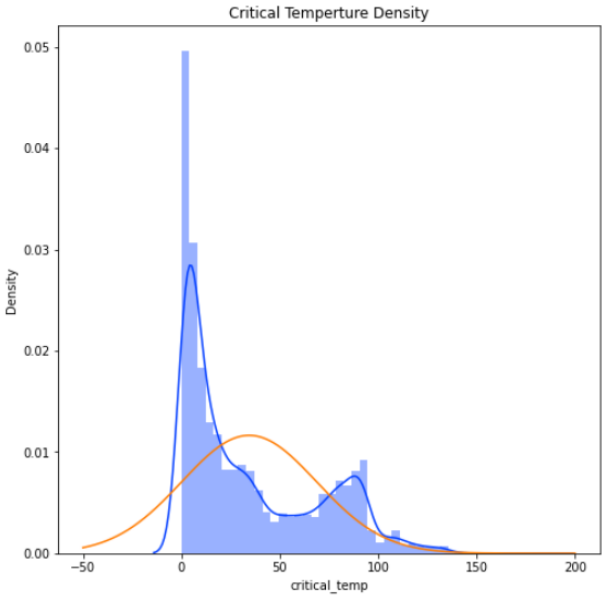


Fig. 1. The histogram plot of the observed critical temperature

According to the figure 1, the distribution of the original critical temperature is approximately right-skewed. Thus, it needs to be transformed before applying any linear regression

model to let be from the normal or at least to close to this distribution. In this regard, we applied the BoX-Cox transformation $T_c$. The results of the transformaion graphed in the figure 2.

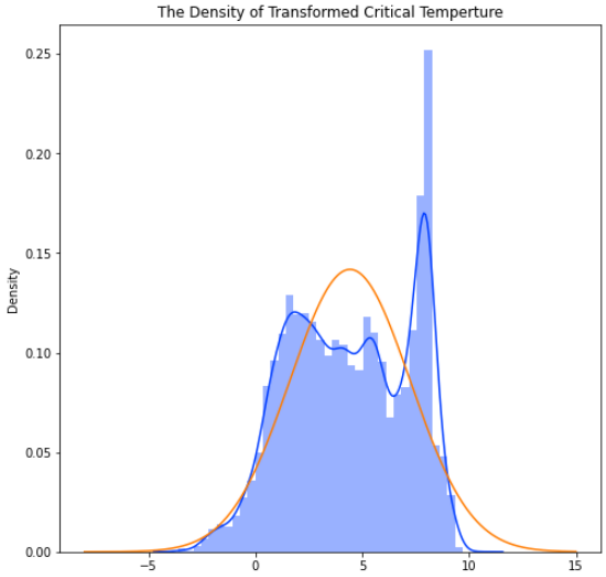


Fig. 2. The histogram plot of the transformed critical temperature

The density of transformed could improve $T_c$ distribution and make it close to normal.

### B. The Proprieties of Predictors

This section explores the properties of 81 features extracted from the superconductors. Regarding to the linear regression models, the distributions of the predictors or independent variables are not main concern, but the any association among the predictors can adversely affect on the performance of the models. This type of correlation in the predictors' matrix can lead to adverse phenomena like collinearity or multicollinearity.

There are some informal and formal methods such as correlation matrix and variance inflation factor (VIF) to identify these phenomena among explanatory variables.

The figure 3 visualize the correlation matrix of 81 superconductors.

According to the above figure3, the superconductors in our study are highly correlated and can make redundant variables. To address this issue, some remedies has been suggest. The popular ones are Principal Components Analysis (PCA) or Partial Least Square (PLS). In this study, however, our scope is fitting a full linear regression and analysis the results of this model.

## III. Results

This section presents and analysis of fitting a multiple regression model on 81 superconductors to predict $T_c$.

Before applying the model, however, the dataset was split into training and testing sets in which 70% of the data assigned

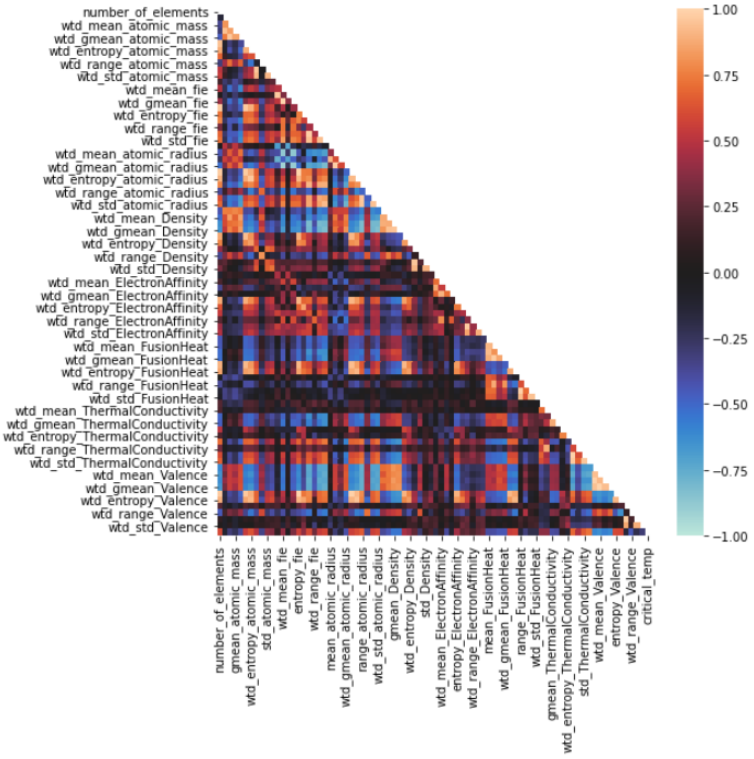


Fig. 3. The correlation matrix of the superconductors

to train set. The plan is performing the linear model on the training set and evaluate the performance of the model via validation dataset based on some measurements like Mean Square Error (MSE), Mean Absolute Error (MAE) and R-square.

So, we we can begin with applying the below model on the training set to predict the Box-Cox transformed critical temperature,

$$T_{c_i}(\lambda) = \beta_0 + \beta_1 X_{i1} + ... + \beta_{81} X_{i81} + \epsilon_i \quad (1)$$

where $i = 1, ..., 14884$ and $T_{c_i}(\lambda) = \left(T_{c_i}^{\lambda} - 1\right)/\lambda; \lambda = 0.24$

The results of fitting the model in (1) on 14884 superconductors of the training set show that 11 features such as *mean_atomic_radius*, *mean_ThermalConductivity*, *mean_Valence*, and to name but a few are not statistically significant in 5 % level. In addition, the *R square* and *R square adjusted* are reported 0.934. It means that the 93.4 % of the variation in the critical temperature of superconductors can be explained by the model. Due to the p-value of the model is zero, we have enough evidences to reject $H_0 : \beta_1 = \beta_2 = ... = \beta_{81} = 0$. So, the model (1) is statistically significant.

The figure 4 shows the observed $T_c$ versus its predicted values with the full model in (1).

Note that the red line fitted with the zero intercept and slop one.

According to the figure 4, many of points are under line. It indicates that the multiple linear regression is suffering

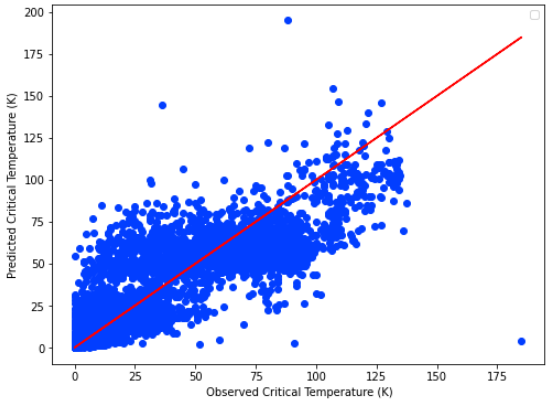


Fig. 4. The Scatter plot of the observed vs Predicted critical temperatures (K)

from under and over prediction and fails to provide a proper prediction of the critical temperature. This is because that prediction of many superconductors with higher $T_c$ are in the below line while the predicted for lower $T_c$ are above the line.

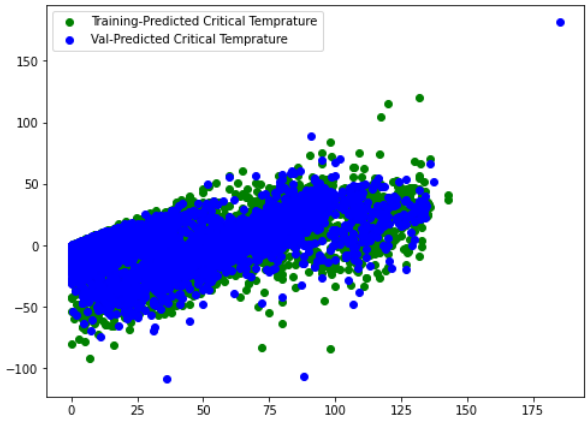


Fig. 5. The Scatter plot of the training residuals vs validation residuals of the critical temperature (K)

Figure 5 shows the scatter points of the residuals of predicted and observed $T_c$ for both training and validation set. It indicates that both errors are overall overlapped approximately.

The table 1 also shows the Mean Square Error, Mean absolute Error, and R-square, for both training and validation sets.

TABLE I
THE EVALUATION'S CRITERIA

| Measure | Train | Validation |
|---|---|---|
| MSE | 296.355 | 303.491 |
| MAE | 11.94 | 12.033 |
| R Square | 0.749 | 0.738 |

According to the above table, the measurements are approximately close to each other. For instance, MSE for validation is slightly higher that of training set.

Our observations from both table 1 and figure6 indicates that the model is not overfitted.

## IV. DIAGNOSTIC TESTS

The multiple linear regression is conducted based on some main assumptions such as linearity, homoscedasticity, independence, and normality. The models can be also adversely affected by the presence of data points called outliers.

If the fitted model violates one of these defined assumptions, it is usually tried to remedy or address this problem. If this try fails, the model cannot be considered as a good model. In this section, we try to check the above assumption in our fitted linear regression model.

1) **Normality:** it is suggested that the residuals is from the normal distribution with zero mean. To check this assumption we can use the graphical methods like *Q-Q plot* and formal tests like *Kolmogorov–Smirnov*. Figure 7 and 8 display the histogram (density) and Q-Q plots of the residuals respectively.

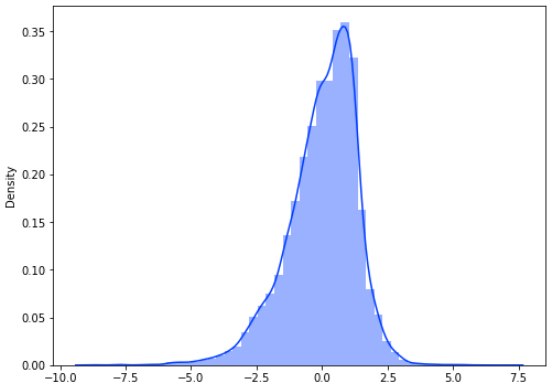


Fig. 6. The density plot of the estimated training residuals

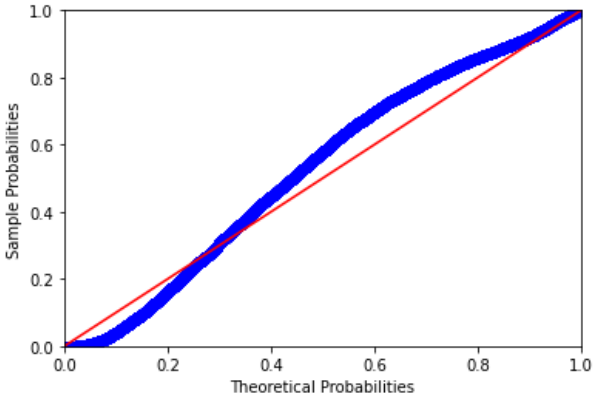


Fig. 7. The Q-Q plot of the estimated training residuals

The observation from two plots in figure 6 and 7 confirm that the residuals are pretty normal. Figure 7 also indicates that the mean of residuals are zero.

2) **Homoscedasticity:** This assumption suggests that the variance of residuals should be constant. To check this assumption, we can plot the predicted critical temperature values versus the residuals. If any defined patterns (linear, quadratic, or funnel shaped) on the plot is observed, we can conclude the presence of heteroscedasticity.
The figure 8 shows the residuals against the predicted $T_c$ values.

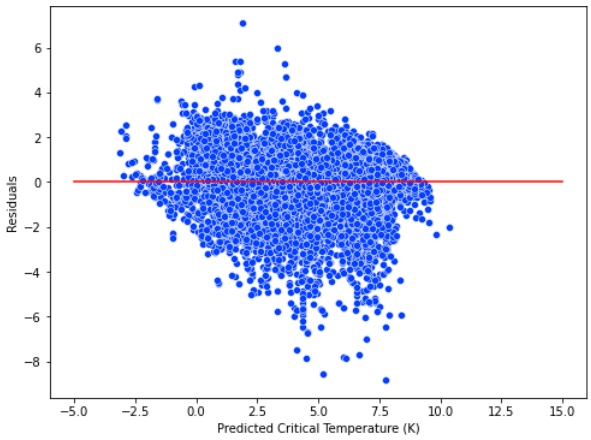


Fig. 8. The residuals of the estimated training residuals vs the predicted critical temperature (K)

Even though many points on the figure 8 are under line, we cannot see any special pattern. So, we conclude that the variance of residuals are almost constant.
However, there is some formal tests like Brown_Forsythe and Goldfeld Quandt to examine this assumption in which $H_0$ assumes that the errors are not heteroscedastic.
The Goldfeld Quandt conducted and the obtained p-value=0.977 is greater than 0.05. Therefore, we can conclude that the variance of error in our study is Homoscedastic.

3) **Independancy of residuals:** It is assumed that residuals should be independent. To check this assumption, we can calculate Dorbin-watson statistic. The range for this test statistic is between 0 and 4. The general rule of this test is that if the statistic value is close 0, there is a strong positive correlation between the residual values. In contrast, the close value to 4 indicates a strong negative correlation between the residuals.
Here, the Dorbin-watson statistic is 1.98. It means the residuals are almost independent.

4) **outliers:**
The observations with the large residual are defined as outliers.
Some statistics were defined to detect the outliers. One popular is call Cook's Distance. It measures how much excluding one observation from the dataset can change the predicted value of the response feature.
The figure 9 shows the results of calculated cook's distance for each observation.

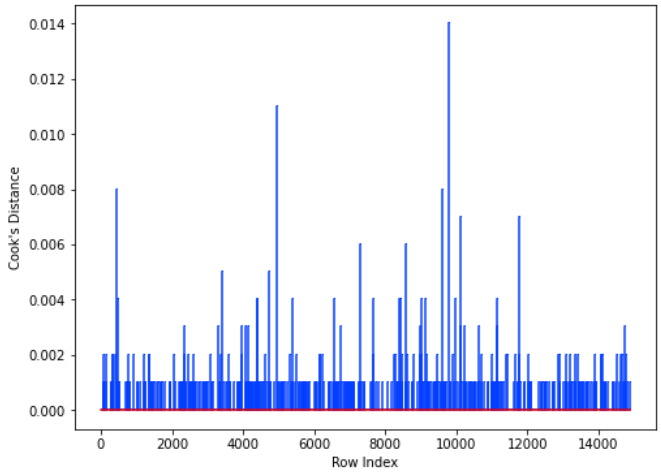


Fig. 9. The cook's distance

The common rule is that if the obtained cook's distance greater than $4/n$, it can be considered as an outlier. In our case, the threshold is around $4/14884 = 0.0002$. Thus, the figure 10 indicates that there are some potential outliers.

## CONCLUSION

The foundation of superconductors is based on the combination of many chemical elements such that the function (superconductivity) of which relies on the specific range of temperature known as critical temperature. Thus, the value of critical temperature is related on these chemical elements. In other words, the critical temperature is a function of features extracted from superconductors. Therefore, it is worthwhile that the value of critical temperature must be predicted by a model based on these attributes before producing superconductors .To do so, we examined the multiple linear regression in this study. Even though the model is not overfitted, it was failed to provide a proper prediction. This is because that it was failed to meet some predefined assumptions. For instance, the distribution of the critical temperature is not following the normal distribution, even though it was transformed. It also shows that it is suffering from over and under prediction. In Addition, using this model results in some potential outliers. Some predictors are also highly correlated and is suffering from the collinearity. In conclusion, the multiple linear regression is not be able to address this problem individually. Thus, to the future study, we can examine the combination of PCA and multiple linear regression or needs to use some other complex models like Xgboot or Lasso.

## REFERENCES

[1] Hamidieh, Kam, A data-driven statistical model for predicting the critical temperature of a superconductor, Computational Materials Science, Volume 154, November 2018, Pages 346354.

[2] Dhakal, Pradip, "Machine Learning-based Approaches for Predicting the Critical Temperature of Superconductor" (2023). Data Science and Data Mining. 9. https://stars.library.ucf.edu/data-science-mining/9

[3] Agbemade, Emil, "Developing a Data-Driven Statistical Model for Accurately Predicting the Superconducting Critical Temperature of Materials using Multiple Regression and Gradient-Boosted Methods" (2023). Data Science and Data Mining. 2. https://stars.library.ucf.edu/data-science-mining/2