Project 1: Predicting the Critical Temperatures of Superconductors using a Linear Regression Model

Rahrooh, Allen
Department of Statistics and Data Science
University of Central Florida
Orlando, Florida
rahrooan@knights.ucf.edu

***Abstract* — Superconductors are materials where the electrical resistance vanishes and magnetic fields are expelled. The critical temperature of a superconductor is where the electrical resistivity of a metal drops to zero and the metal goes to a different form of matter which is known as the superconducting phase. The purpose of this study is to see if given various numerical values of the superconductors to see if there is a linear relationship between the critical temperature and the properties of each superconductor. I first applied a linear regression model and analyzed the Residual Standard Error, F-statistic, and Multiple R-squared values. I found the Residual Standard Error of 17.59 on 21181 degrees of freedom, F-statistic of 733.8 on 81 and 21181 degrees of freedom, and a Multiple R-squared of 0.7373. These values suggest that there is a positive linear relationship between the critical temperature and the properties of each superconductor.**

A. Introduction & Background

Linear regression has been a very popular method for machine learning and predicting outcomes on big data sets. Previous researchers have been able to use the linear regression model to predict outcomes of mortality [1] and water temperatures from chalk streams [2]. The linear regression works by fitting a regression line to the data and outputs how linearly the data is using (1).

$$y = \beta_0 + \beta_1 x_1 + \cdots \beta_n x_n + \varepsilon_i \quad (1)$$

The y represents the response variable, the x_1,x_2, and x_n represent the predictors, and the β_0, β_1, and β_n represent the weights of each predictor and determine which predictors are significant in terms of how linear the data is. The ε_i represents the random error.

B. Objective

The goal of the data set that is used from [3] is to predict the temperature of a superconductor based on various features also known as predictors (x_1,x_2, and x_n) using a multiple linear regression model.

II. Methods

A. Setting Up Linear Models

The data set was obtained from https://archive.ics.uci.edu/ml/datasets/Superconductivty+Data [3] and exported to R Studio using the import dataset option in the R Studio console. To fit the linear regression model I used the linear model (lm) command with the response variable (y) set as the critical temperature and the predictors as the rest of columns such as number of elements and mean atomic mass. A total of 81 predictors were used for the linear model. I decided to include all 81 predictors and not do a variable and feature selection because it might lower the model performance by removing valuable predictors.

B. Analyzing and Interpreting Linear Model Output

Once the linear model is fitted using the lm function I created a summary of the model. From this model I analyzed the Residual Standard Error (2), F-statistic (3), and Multiple R-squared (4) values. The Residual Standard Error determines the quality of the linear regression fit. The Residual Standard Error is equal to the average amount that the predictors will be from the regression line which is also equivalent to the square root of the Mean Squares for Error (MSE).

$$\text{Residual Standard Error} = \sqrt{\text{MSE}} \quad (2)$$

The lower the Residual Standard Error the better the model fit is. The F-statistic determines if there is a relationship between the predictor and response variables which is calculated by the division of the Mean Square for the Model (MSM) and the MSE.

$$F = \frac{\text{MSM}\left(\frac{\text{SSM}}{\text{DFM}}\right)(\text{explained variance})}{\text{MSE}\left(\frac{\text{SSE}}{\text{DFE}}\right)(\text{unexplained variance})} \quad (3)$$

The further the F-statistic is from the value of 1 the better the linear relationship is. The Multiple R-squared value provides a measure of how well the model fits to the actual data and is calculated by the division of the Sum of Squares Regression (SSR) and the Sum of Squares Total (SST).

$$R^2 = \frac{\text{SSR}}{\text{SST}} \quad (4)$$

The R-squared value can range from -1 to +1 with +1 being a positive linear fit and -1 being a negative linear fit.

C. Plotting Linear Models

Once the summary model were produced I generated 4 plots being: residuals vs. fitted, normal Q-Q, scale-location, and residuals vs. leverage plots. The residuals vs. fitted plot shows if the residuals have a non-linear pattern. The normal

Q-Q plot shows if the residuals are normally distributed. The scale-location plot shows if the residuals are spread equally along the ranges of the predictors (x_1,x_2, and x_n). The residuals vs. leverage plot shows if the data has any outliers using cook's distance (red dashed line).

III. Results & Discussion

A. Linear Model Results

```
Call:
lm(formula = critical_temp ~ ., data = train)

Residuals:
Min      1Q  Median      3Q     Max
-84.987  -9.370   0.595  10.976 171.246

Residual standard error: 17.59 on 21181 degrees of freedom
Multiple R-squared:  0.7373,        Adjusted R-squared:
0.7363
F-statistic: 733.8 on 81 and 21181 DF,  p-value: < 2.2e-16
```

Fig. 1 Linear Model Results

From the results of the linear model fitting as shown in Fig. 1 we see that the Residual Standard Error is 17.59 on 21181 degrees of freedom, a Multiple R-squared value of 0.7373, and a F-statistic value of 733.8 on 81 and 21181 degrees of freedom. These values indicate that the linear model fitted well with a high F-statistic, low Residual Standard Error, and Multiple R-squared value closer to 1 then 0.

B. Linear Model Plots

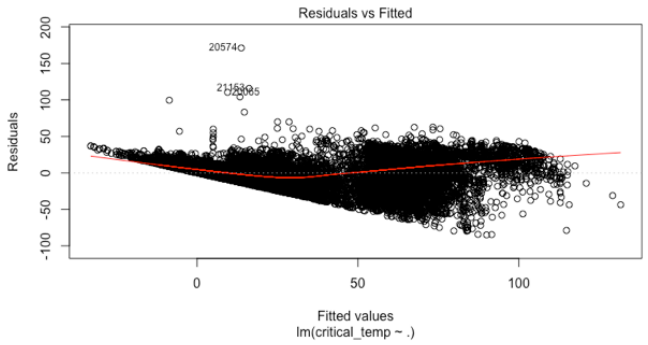

Fig.2 Residuals vs Fitted Plot

Fig. 2 shows the residuals vs fitted plot of the fitted linear model. From the figure it seems that the residuals do not have a nonlinear pattern between the predictors and the response variable since there are few outliers.

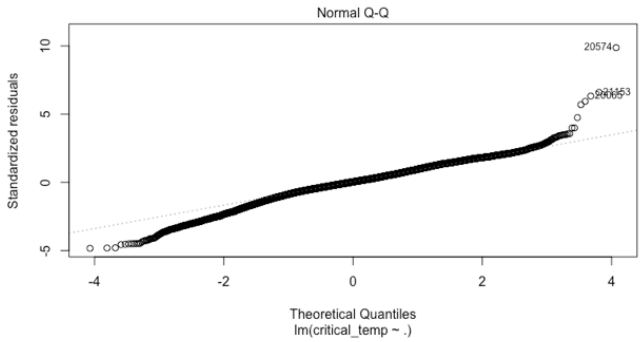

Fig.3 Normal Q-Q Plot

Fig. 3 shows the normal Q-Q plot for the fitted linear model. From the figure it shows that the residuals are normally distributed with a few outliers towards the end of the distribution.

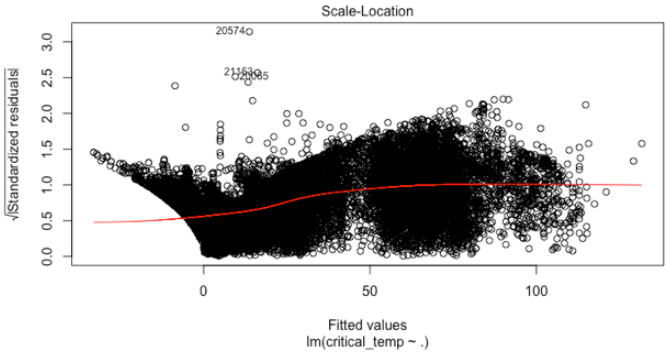

Fig. 4 Scale-Location Plot

Fig. 4 shows the scale-location plot for the fitted linear model. From the figure it shows that the residuals have a good spread across the range of predictors with the same outliers (20574 & 21153) as Fig. 3 and Fig. 2.

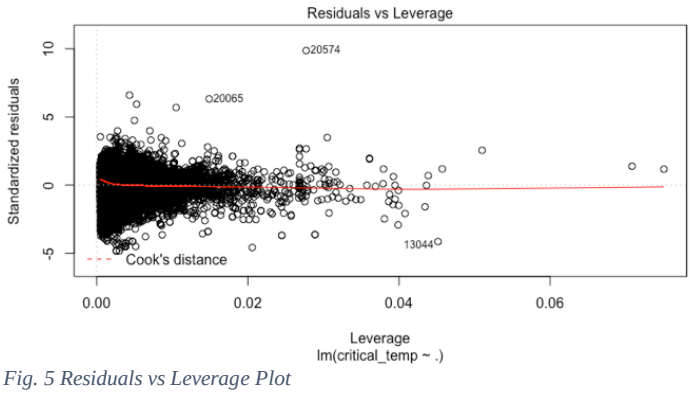

Fig. 5 Residuals vs Leverage Plot

Fig. 5 shows the residuals vs. leverage plot of the fitted linear model. As the plot shows cook's distance does not appear on the plot, which means that the outliers (20574 & 20065) do not have a strong influence to affect the R-squared value and overall fit of the linear model.

IV. Conclusions

Overall the linear regression model for predicting the critical temperature of superconductors had a good fit. The overall results of the model fitting showed that with minimal data filtering a good fit can be achieved. All the plots saw outliers at data point 20574 so for future analysis the removal of the feature and related features could increase the model fit.

Acknowledgment

I would like to thank Kam Hamideh[3] for providing the data set to the public.

References

[1] A.G. Barnett, S. Tong, and A.C.A. Clements, "What measure of temperature is the best predictor of mortality?," *Environ. Res.*, 2010.

[2] A.P. Mackey and A. D. Berrie, "The prediction of water temperatures in chalk streams from air temperatures," *Hydrobiologia*, 1991.

[3] Hamidieh, Kam, A data-driven statistical model for predicting the crticial temparture of a superconductor, Computational Materials Science, Volume 154, November 2018, Pages 346 – 354.