Analyzing Stellar Metallicity Using Regression Models

Jordan Stout

Abstract

This study leverages a subset of data from the ESA Gaia archive, which includes 300,000 astronomical observations to explore the predictive relationship between stellar properties and metallicity. Through exploratory data analysis, key features such as stellar temperature, luminosity, surface gravity, and color indices were analyzed with the goal of uncovering potential correlations with metallicity. A range of regression techniques to model the metallicity of stars were applied based on these features. The findings of this analysis aim to contribute to the understanding of stellar evolution and provide insights into the factors influencing metallicity within the Milky Way.

Introduction to Gaia

The European Space Agency's Gaia space telescope has been in operation since 2013, mapping the positions and motions of over a billion stars within the Milky Way and other astronomical phenomenon such as quasars, supernovae, asteroids, globular clusters and more. Gaia's lifespan was originally set for 5 years, due to the subsequent success of the mission Gaia is still collecting and transmitting data 1.5 million miles from Earth.

Exploratory Data Analysis & Data Refining

About the Data

Gaia data has had three major public releases. In this analysis, 2022 data from the third release (DR3) was used because it contains expanded photometric and spectroscopic data. This means DR3 data contains pre-calculated values that usually need to be transformed to extract metrics such as temperature, gravity and luminosity. An Astronomical Data Query Language (ADQL) query was written to extract astrophysical elements of stars with their associated upper and lower error bounds. The elements used in this project were luminosity, apparent magnitude, temperature, surface gravity, distance, radial velocity, and blue, red, and G-band magnitudes. While an initial sample of 300,000 stars was extracted from the population of billions of stars in the Gaia archive, eventually the data was refined down to a sample for analysis of 710 observations.

NA Values

Nearly 80% of observations contained some form of NA value which were promptly omitted. There are a few reasons why so many NAs are present in the dataset.

Lack of Data for Specific Stars: Not all stars will have complete information for every parameter. For example, if a star was not observed in certain photometric bands (like BP or RP), it may not have a reliable estimate for surface gravity or metallicity.

The GSP-Phot Method: The method for calculating certain astrophysical values relies on stellar model libraries (e.g., MARCS, PHOENIX), and not all stars will have a perfect match with the models.

Data Quality: The Gaia mission uses sophisticated models to derive stellar parameters, but if the models cannot converge or if the data are noisy, the parameters may be assigned as missing.

Parsing Galactic Addresses

Unlike everyday 3D plots, stellar data is given in the form of Right Ascension (RA), Declination (Dec), and Parallax (p) with respect to the sun. RA and Dec can be thought of as latitude, longitude while parallax is inversely proportional to distance. After converting RA and Dec to radians and parallax to distance in parsecs a simple polar to Cartesian coordinate transformation is done to plot the stars in 3D as shown in Figure 1. In this figure, (0,0,0) represents the location of the telescope.

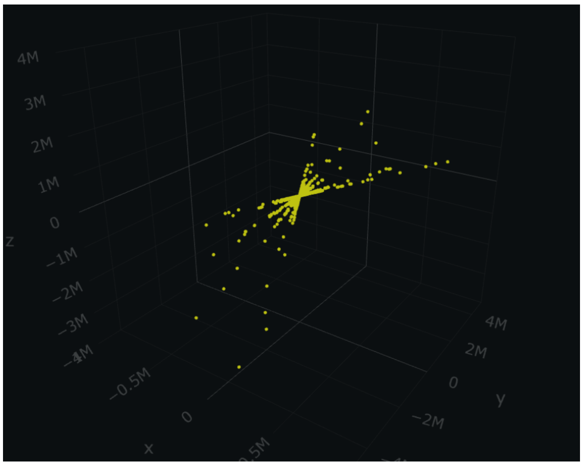


Figure 1: Initial Stellar Location Data

Two major issues with the data were made apparent with this visualization. While it could have been a result of the random nature of data selection, it seemed peculiar that the telescope appeared to have "eyes in the back of it's head". It was determined that this was a result of error in parallax measurements, making the value negative consequently flipping the signs of each Cartesian coordinate value. This was resolved after restricting parallax values to within the first quartile of it's error intervals.

The second major issue was that the scale of the distance values are on the order of millions of parsecs while Earth is only 25,000 parsecs from the edge of the Milky Way. A density plot was made to see where each value landed in terms of distance from the sun shown in Figure 2.

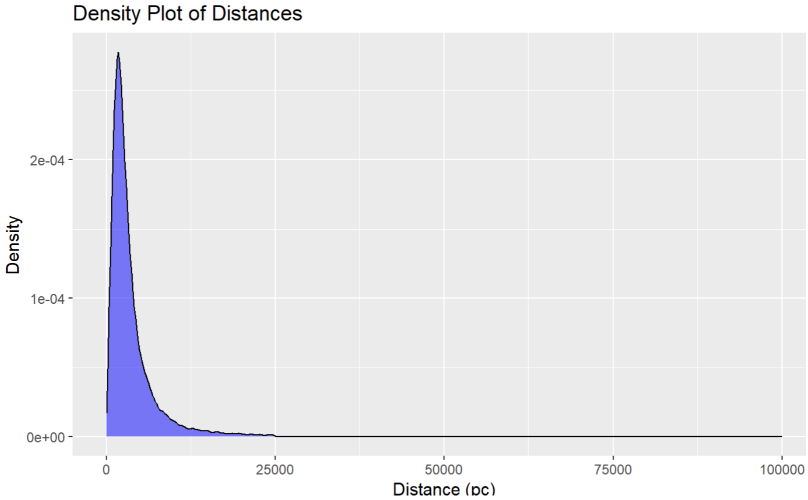


Figure 2: Density of stars' distance from Earth

As shown, while there is limited values outside of 25,000pc, the vast majority of values lie within the range of the Milky Way. It is insinuated that the few values that lie so far out are astronomical phenomenon which Gaia detects due to their extreme brightness such as quasars. A simple limit of 25,000pc was put on the data to ensure only stars within the Milky Way were present.

Error Handling

It was made apparent that error values should not be ignored in the case of this dataset. Luckily Gaia offers upper and lower error bounds for each of their astrophysical measurements. In this project, the error values of temperature, surface gravity, and metallicity were taken into account. An error magnitude was calculated using Equation 1 and the data was then limited to error values in the first quartile for each of these parameters.

$$\text{Error} = |\text{Upper Bound} - \text{Lower Bound}| \tag{1}$$

After omitting NA values, restricting distance values to within the Milky Way, and implementing error constraints, the original dataset of 300,000 observations was reduced to 710 stars which are visualized in Figure 3.

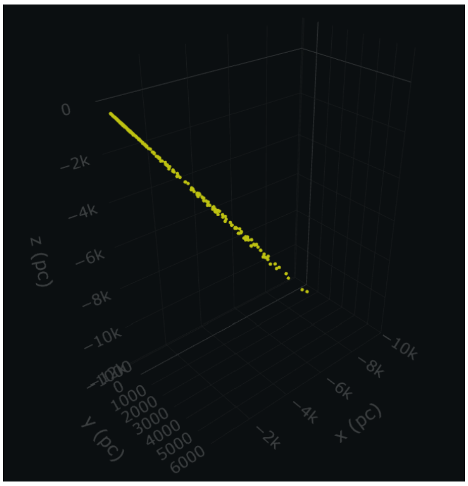


Figure 3: Density of stars' distance from Earth

Luminosity

Part of EDA is learning more about your data. It was decided to do a deep dive into the luminosity values of our observations. While luminosity was not natively available in the Gaia archive, it was calculated using the equations below. Where (L) is Luminosity of the star, $(L\odot)$ is the luminosity of the Sun, (d) is distance in parsecs, (M) is Magnitude, and $(M\odot)$ is the Magnitude of the Sun.

$$M_{\text{absolute}} = M_{\text{apparent}} - 5\log_{10}(d) + 5$$

$$L = L_{\odot} \cdot 10^{0.4(M_{\text{absolute}} - M_{\odot})}$$

A histogram was then plotted of luminosity values and it was extremely right-skewed so a log-luminosity graph instead is shown below.

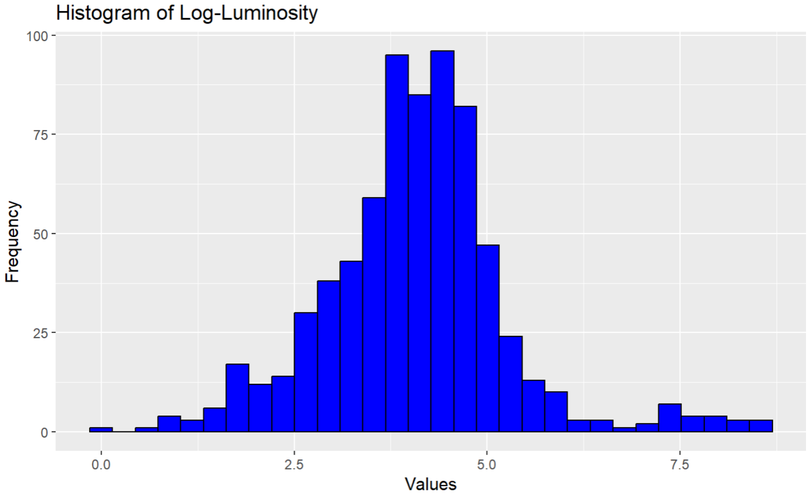


Figure 4: Log-Luminosity of Final Stars

The right-skewness of the histogram makes sense physically as the majority of stars in the universe are low-luminosity stars, such as red dwarfs. These stars make up about 70-80% of all stars in our galaxy and have low luminosity compared to larger stars like the Sun or massive blue giants.

Regression Techniques

Handling Multicollinearity

Beginning running regression models began nearly immediately after the completion of EDA. It was later discovered that many predictors being used were nearly linearly dependent if not highly correlated making all models ran untrustworthy. Figure 5 shows a correlation heat-map. A correlation score of above 0.8 is considered concerning.

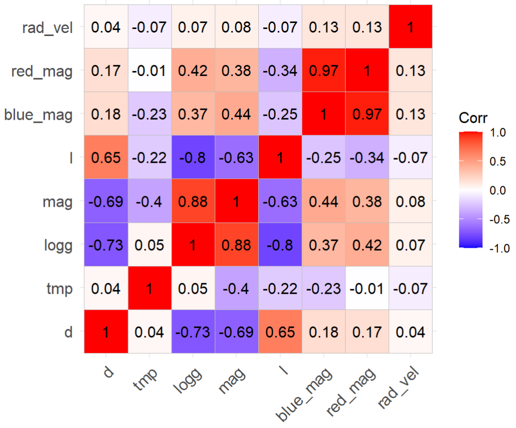


Figure 5: Correlation matrix visualized

VIF scores were also calculated for each predictor to further confirm that severe multicollinearity existed among predictors. A VIF score greater than 5 is considered high correlation and should be investigated further.

Predictor	VIF Score
Distance	10.07
Temperature	12.30
Log Gravity	31.51
Magnitude	38.26
Luminosity	3.04
Blue-Mag	759.73
Red-Mag	701.58
Radial Velocity	1.05

Table 1: VIF scores of predictors

To resolve this issue, VIF scores were continually calculated while removing predictors until the score of every predictor was below 3. For fun, using luminosity as the link the positive linear correlation between log gravity and magnitude can be proven.

$$L = 4\pi R^2 \sigma T^4, \quad g = \frac{Gm}{R^2}, \quad L = L_0 \cdot 10^{\frac{M_0 - M}{2.5}} \tag{2}$$

$$L \sim R^2, \quad g \sim \frac{1}{R^2}, \quad L \sim 10^M \tag{3}$$

$$M \sim \log_{10}(g) \tag{4}$$

The fact that the correlation score on the heat map does not show a one-to-one correlation is most likely due to errors in the data as the errors of magnitude were not taken into account. The near perfect correlation between red and blue magnitudes is due to what they represent. Both measure the star's luminosity but in slightly different wavelengths, meaning the brighter the star, the higher both values go.

Linear Regression

Linear regression has long been utilized in astronomical data analysis. It was famously used by the Edwin Hubble when he correlated the radial velocities of 46 extra-galactic nebulae with their distances from Earth (Hubble, 1929). On the coattails of Hubble, I ran a linear regression model to look for correlations between astrophysical properties of stars and their metallicities.

The standard linear regression formula is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \epsilon_i, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

- y_i: The response variable for observation i.
- β_0: The intercept, representing the expected value of y when all predictors are zero.
- β_n: The coefficient for the n-th predictor, representing the change in y for a one-unit increase in x_i, holding all other predictors constant.

- x_i: The value of the i-th predictor.
- ϵ_i: The error term for observation i, assumed to be independently and normally distributed with mean 0 and variance σ^2.

Assumptions for Linear Regression

Before analyzing the model's fit, there are 3 key assumptions of linear regression that have to be met.

1. **Linearity:** The relationship between the independent and dependent variables is linear. This can be proven by plotting the residuals versus fitted values and seeing if the resulting points appear to hover around $y = 0$ (see Figure 6).

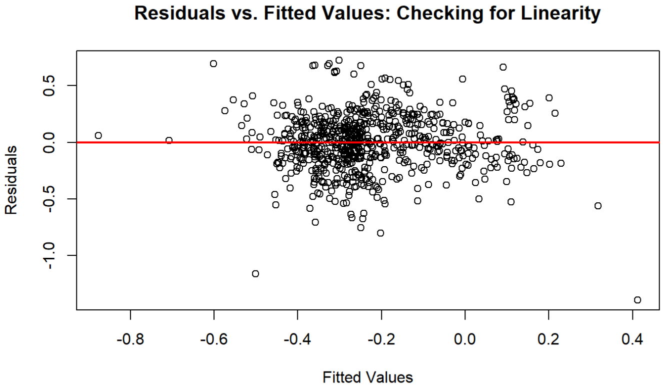
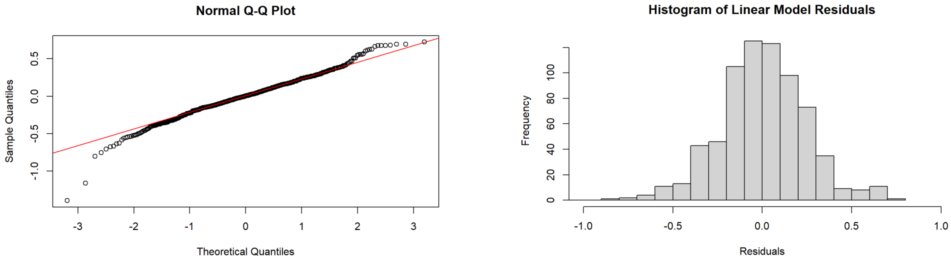


Figure 6: Linearity test

2. **Normality of Residuals:** Residuals are normally distributed. This can be shown by plotting a histogram of the model's residuals alongside a QQ plot (see Figure 7).



(a) (b)

Figure 7: Linear regression results

3. **Lack of Multicollinearity:** The independent variables are not highly correlated with each other. This was taken care of already in the Handling Multicollinearity section.

Basic Linear Regression Model

Our null hypothesis H_0 states that there is no relationship between the predictors and metallicity, i.e. all of the regression coefficients will be zero.

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

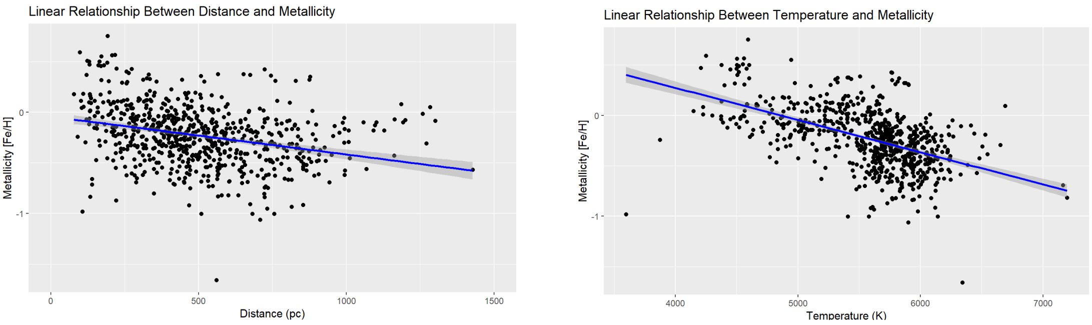
The alternative hypothesis H_1 asserts that at least one of the regression coefficients is not equal to zero, implying that at least one independent variable has a statistically significant relationship with metallicity.

$$H_1 : \text{At least one } \beta_i \neq 0 \quad \text{for some } i \in \{1, 2, \ldots, k\}$$

Variable	**Estimate**	**Std. Error**	**t value**	**p-value**	**Sig.**
(Intercept)	1.82	2.22×10^{-1}	8.18	1.29×10^{-15}	***
distance	-1.17×10^{-4}	3.65×10^{-5}	-3.21	0.0014	**
temperature	-3.40×10^{-4}	2.93×10^{-5}	-11.60	$<2 \times 10^{-16}$	***
magnitude	-2.16×10^{-2}	1.48×10^{-2}	-1.46	0.1449	
luminosity	-3.08×10^{-31}	3.22×10^{-30}	-0.096	0.9239	
radial velocity	1.81×10^{-4}	3.64×10^{-4}	0.50	0.6179	
$R^2 = 0.28$					

Table 2: Linear Regression ANOVA Results

Although temperature and distance are considered statistically significant ($p \ll 0.05$), the value of our R^2 shows that only 28% of the variance in the dependent variable can be explained by the independent variables in the model. This value traditionally indicates a poor level of explanatory power and that there is more of the story to tell. When looking at the relationship between temperature and metallicity in Figure 8(b) it can be seen that there is a clear relationship between the two. The t-value $|t| = 11.6$ shows that the ratio between the estimate and its standard error is sufficiently high. This result of a negative correlation has been observed in alternate studies (Mansfield & Kroupa, 2021).



(a) Distance vs. Metallicity

(b) Temperature vs. Metallicity

Figure 8: Linear regression results

ANOVA values aside, predictions were made of metallicity using the model and calculated the MSE using Formula 5. In the formula, n is the number of observations, $\hat{y}_i$ is the predicted value, and y_i is the actual value.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0.062 \tag{6}$$

The null MSE was also calculated using the same formula as regular MSE but instead of predicted values the mean of our dependent variable is used. The error reduction can be calculated using Equation 6 to see how effective our model is.

$$\text{Error Reduction} = 1 - \frac{\text{MSE}}{\text{Null MSE}} \tag{7}$$

$$\text{Error Reduction} = 1 - \frac{0.062}{0.086}$$

$$\text{Error Reduction} = 1 - 0.7209 = 0.2791(27.91\%)$$

Given the results of the ANOVA table and our error reduction calculation, the null hypothesis is rejected and the alternate hypothesis accepted.

Multilevel Linear Models

In the previous section, linear regression with complete pooling was used to estimate the relationship between the predictors and metallicity. While this approach assumes a single global model for all observations, other pooling methods allow us to account for variations across different groups within the data. To experiment with other pooling techniques, categorizing each star into some sort of group is needed. It was chosen that stars would be categorized into four evenly populated groups based on their distance from Earth in parsecs.

$$groups = \begin{cases} 0, & \text{if } d \leq 331 \\ 1, & \text{if } 331 < d \leq 471 \\ 2, & \text{if } 471 < d \leq 654 \\ 3, & \text{if } d > 653 \end{cases}$$

While the regression formula for complete pooling was described earlier in Equation 5, partial and no pooling follow Equation 8.

$$y = \alpha_{ji} + \beta x_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_y^2) \tag{8}$$

The difference between partial pooling and no pooling lies within the intercept α. In partial pooling, α_j is drawn from a shared normal distribution with mean μ_α and variance σ_α^2 which allows sharing of information across groups. In this case our group (j) represents how far from the Earth each star is and each individual star is i. In models with no pooling, each group is assumed to have its own unique intercept.

Method	MSE	AIC	BIC
Linear w/ Complete	0.062	55.35	87.31
Linear w/ None	0.067	112.68	144.63
Linear w/ Partial	0.061	258.34	294.86
Null Model	0.086	280.47	289.60

Table 3: Pooling method comparison

All of the pooling methods outperformed the null model on all fronts. Looking at MSE, partial pooling performed the best with the lowest MSE. The reason the AIC and BIC are higher than the other methods is because both of those metrics take model complexity into account.

Discussion & Validation

It proved difficult to find modern papers validating the use of simple linear regression on astronomical data. There are a few papers found from the late 1900's that discuss proposed improvements to linear regression in astrostatistics. The vast majority of papers found discuss the increased use of Bayesian Generalized Linear Models (Souza, 2015). While linear regression is an outdated technique, the use of Gaia DR3 data to analyze and predict metallicity values is on the cutting edge.

My findings on predicting stellar metallicity using regression methods do provide an interesting contribution to the field. Recent work leveraging Gaia's BP/RP spectra has derived metallicities for over 10 million stars, using calibrated dust maps and extinction models to address reddening and extinction challenges (Xylakis-Dornbusch, 2024). Machine learning techniques applied to Gaia DR3's RR Lyrae catalog have estimated metallicities for over 250,000 stars through analyses of pulsation periods and amplitudes (Muraveval, 2024). These advancements showcase the innovative computational methods available today, against which my regression-based methodology provides a simpler, yet still valuable, framework for comparison.

References

1. R.S. de Souza a, E. Cameron b, M. Killedar c, J. Hilbe d e, R. Vilalta f, U. Maio g h, V. Biffi i, B. Ciardi j, J.D. Riggs k (2015). The overlooked potential of Generalized Linear Models in astronomy. *Elsevier*. 12(21-32). https://doi.org/10.1016/j.ascom.2015.04.002

2. S. Mansfield, P. Kroupa (2021). A discontinuity in the luminosity–mass relation and fluctuations in the evolutionary tracks of low-mass and low-metallicity stars at the Gaia M-dwarf gap. *EDP Sciences*. 650. https://doi.org/10.1051/0004-6361/202140536

3. E. Hubble (1929). A relation between distance and radial velocity among extra-galactic nebulae. *Astrophysical Journal*. 15(3) 168-173. https://doi.org/10.1073/pnas.15.3.168

4. M. Akritas, M. Bershady (1996). Linear Regression for Astronomical Data with Measurement Errors and Intrinsic Scatter. *The Astrophysical Journal*. 470(23). arXiv:astro-ph/9605002

5. T. Xylakis-Dornbusch, N. Christlieb, T.T. Hansen, T.Nordlander, K. B. Webber, J. Marshall (2024). Metallicities for more than 10 million stars derived from Gaia BP/RP spectra. *The Astrophysical Journal*. https://doi.org/10.1051/0004-6361/202348885

6. T. Muraveva1, A. Giannetti2, G. Clementini1, A. Garofalo1, L. Monti1 (2024). Metallicity of RR Lyrae stars from the Gaia Data Release 3 catalogue computed with Machine Learning algorithms. *The Astrophysical Journal*. 93(3). https://doi.org/10.1093/mnras/stae2679