Capstone Project: Exploring Red Giant Stars using the APOGEE Dataset, with Linear Regression models and Hypothesis Testing.

Mary Zhao, Rebecca Li, William Li

Abstract

Properties about stars can help us learn more about where they came from, what they experienced, and how they behave within our universe. To learn more about their properties, we looked at the SDSS APOGEE dataset (Leung, n.d.) and investigated observations about the temperature, surface gravity, wavelengths, and metallicity of stellar spectra of red giant stars.

We will be creating a linear regression model to analyze the relationship between surface gravity and temperature of a star if any, a hypothesis test to see how well measured wavelengths of stars compare to that of our Sun, and another hypothesis test to if the median amount of iron in stars are similar to the Sun or not. We found that the wavelengths of the observed stars have a similar difference to the Sun while the median amount of iron of the same stars is significantly less than the Sun. We also found that the surface gravity and temperature of a star are correlated. These results spike interest into the reasoning behind why we obtained results that we did, which could be explored in the future and tested against other, similar, datasets.

Introduction

Stars can be thought of as living fossil records. Many of their properties can tell us about their past. For instance, the amount of a certain element in a star can tell us its historical birth rate. Because these properties can teach us a lot about stars, we wanted to analyze more into them. More specifically, we investigated these three questions:

Question 1: **Can we use a linear regression model to fit the relationship between the logarithm of surface gravity and effective temperature of a star?** A linear regression model describes how correlated two quantitative variables are.

Question 2: **How do well-measured wavelengths of stars compare to that of our sun?** This will be done using a hypothesis test, which determines how significant the findings are by examining how unusual the sample mean is with a p-value.

Question 3: **Is the median amount of iron on the surface of the star similar to our Sun or not?** Again, this will use a hypothesis test.

Data

The dataset we will be using is the SDSS APOGEE dataset, collected by Henry Leung.

This dataset has 99,705 observations and 21 variables. Our variables of interest include:

- **logg**, which measures the surface gravity of the star
- **teff**, which is the temperature of the star
- **wavelength**, which measures the wavelengths of stars in Angstroms (10^-10 m = 0.1 nm) for the individual stellar spectra.
- **snr**, The signal-to-noise ratio (SNR) for the spectrum. This tells us roughly how well each spectrum has been measured (with higher=better).
- **fe_h**, which measures [FE/H] in base-10 logarithm units relative to the Sun. A positive value indicates more iron than our Sun, while a negative value indicates less iron.

The dataset was stored in a file called “STA130_APOGEE.h5.” To open it, we need to ensure the library rhdf5 was available in RStudio. From there, we can extract each variable and its values and store that into a tibble and/or dataframe. The data was already cleaned, so there were no missing values when we obtained the dataset. The tidyverse package is also needed to be able to use some analysis functions described later on.

Question 1

Is using a linear regression model a suitable method to describe the relationship between the logarithm of surface gravity and effective temperature of a star?

Methods/Analysis

For question 1, we started by plotting the logg and teff variables into a scatterplot to look at the distribution between them. Next, we ran a linear regression model between the teff and logg variables from the dataset to find out if a linear line was an appropriate model for the graph, which it did end up being. Finally, we found the correlation coefficient of the data to determine if the relationship between the 2 variables is strong or weak.

Results

Based on the general shape of the scatterplot, we can see that the linear model is a good fit. We can also see that these two variables are positively correlated. After calculating the correlation coefficient, we obtain a value of 0.8275 which demonstrates a strong relationship between the 2 variables.

```
> cor(logg$value, teff$value)
[1] 0.8274559
```

Data Visualizations

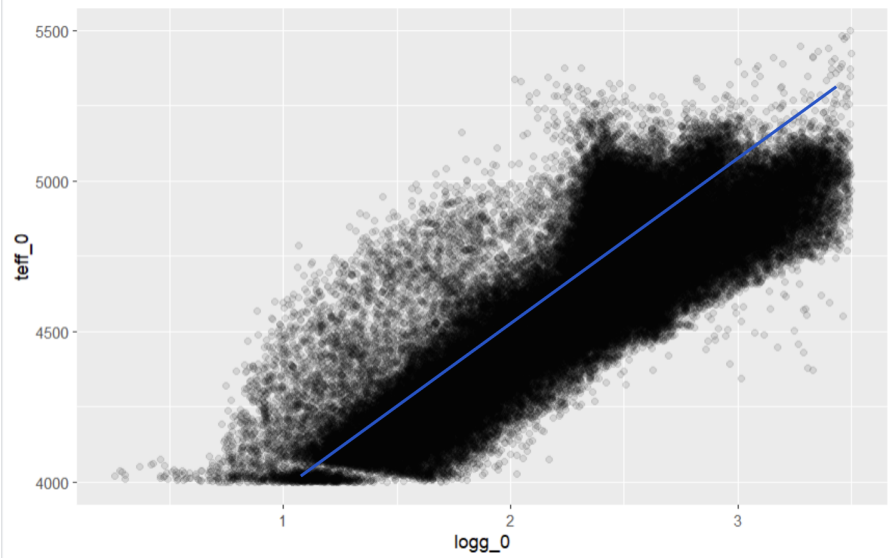

Figure 1.1

Discussion

While we were able to find that the two variables are strongly correlated, it's important to note that this is not a sign of causation, which we cannot conclude based on the results that we obtained. This means we cannot say that a larger surface gravity from a red giant star causes a hotter temperature. One explanation for this correlation could be that bigger stars tend to be hotter, and the larger a star is the greater gravitational pull it has.

Question 2

How do well-measured wavelengths of stars compare to that of our sun?

Methods/Analysis

For question 2, we first extracted the wavelength variable and the SNR variable from the dataset, then we showed a bar graph to show how well the data is measured based on the SNR variable. From figure 2a, we can see that there are more well-measured data than poor-measured data, so we can proceed (In fact, I could not do anything even if I needed to filter some of the data since the wavelength variable has only 7514 entries while the SNR variable has 99705 entries. This means that not every measured wavelength has a corresponding snr variable, thus I cannot say which wavelengths are well-measured and which are not - I can only check the overall performance).

Next, we did a hypothesis test with alpha-level 0.05, sample size 10, and 1000 repetitions. The null hypothesis is that the wavelengths of the suns measured is the same as the sun, while the alternating hypothesis is that the wavelengths are not the

same. In addition, the wavelength of the sun is determined by finding the Solar irradiance spectrum at the surface level, which is approximately 1625 nm and 16250 Angstroms.

The resulting p-value was 0.095.

```
> p_value
[1] 0.095
```

Results

The p-value found was ultimately 0.095, which is more than the alpha level 0.05 that we set before the experiment. Therefore, we have to say that we fail the reject the null hypothesis that the wavelengths of the well-measured data is similar to the sun, meaning that the wavelengths are different.

Data Visualizations

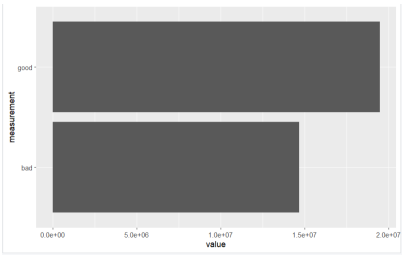

Figure 2a

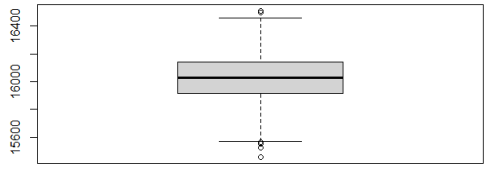

Figure 2b

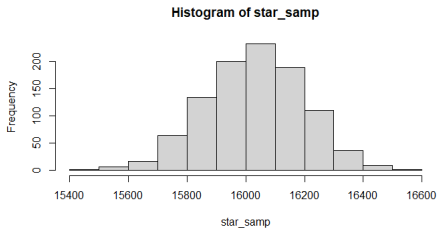

Figure 2c

Discussion

From this question, we see that the wavelengths of red giants stars are different from

the sun. One possible reason is that the mass of red giants stars are often less than the sun (only 80% of the mass of the sun), meaning they radiate less energy and have a lower spectrum (they radiate type K spectrum while the sun radiates type G spectrum). Another possible reason is that the temperature of the sun is higher than the average temperature of the red giants, leading to a less radiation, thus different wavelengths.

Question 3

Is the median amount of iron on the surface of the star similar to our Sun or not?

Methods/Analysis

After extracting the 'fe_h' values, a summary table was created using the summarize function in R. Here, the table is displayed in Figure 3.1. However, we need to use hypothesis testing to see if the median calculated is significant. Our null hypothesis is that the median is equal to zero (similar to the Sun), and our alternative is that it is not equal to zero (not similar to the Sun).

Before the test, we set the seed so that results are reproducible.Then, we set an alpha significance level to 0.05. Since we will run N=1000 tests, we create an empty vector of size N and repeatedly collect the median of N samples of n=100 stars. The two-sided p-value is calculated taking the sum of sample medians greater than or equal to zero, divided by N, and then multiplied by two (for two sides).

n	median	mean	iqr
<int>	<dbl>	<dbl>	<dbl>
99...	-0.19942	-0.2322343	0.3932...

Figure 3.1 The summary table.

Results

The calculated median amount of iron compared to the Sun is -0.19942. This means the amount of iron on red giant stars are likely to have less iron than the sun.

From our hypothesis test, the calculated p-value of 0 was smaller than our alpha level of 0.05. Thus, we can reject the null in favor of the alternative and say the median amount of iron is not similar to the Sun.

Data Visualizations

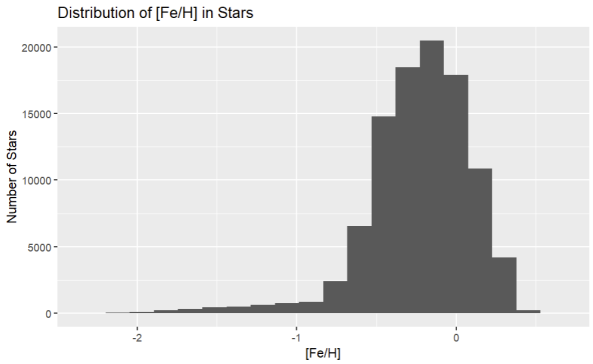

Figure 3.2 A histogram of the distribution of [Fe/H] in stars. This helps show the overall shape of the data. It is left-skewed.

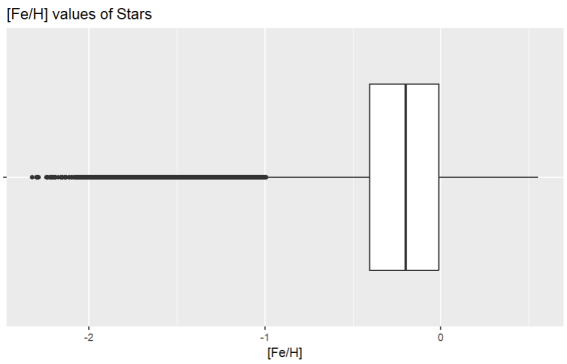

Figure 3.3 A box plot of the distribution of [Fe/H] in stars. This shows the quartiles and outliers.

Discussion

The median amount of iron being less than the Sun is significant. This implies red giant stars have a lower metallicity, or fraction of heavier elements, than our Sun. Low metallicity stars imply they were formed a longer time ago (“Elemental abundances”, n.d.), so the red giant stars are generally much older than our Sun.

Conclusion

All in all, the properties of stars tell us a lot about their past. To study these properties, we conducted a linear regression model and two hypothesis tests to see if there is a relationship between surface gravity and temperature, if the well-measured wavelengths are comparable to the Sun, and if the median abundance of iron is similar to the Sun or not.

Our results show that we were able to find a correlation between surface gravity and temperature, a comparable wavelength of red between red giant stars, and our Sun and that the median abundance of iron is not similar to that of the Sun.

In the future, we hope to be able to study deeper into why surface gravity and temperature are related and potentially study other types of star than just red giant stars, such as stars that are a similar type to our own Sun in order to continue to compare which traits it shares with other stars.

References

Elemental abundances. Center for Astrophysics. (n.d.). Retrieved April 9, 2023, from https://www.cfa.harvard.edu/research/topic/elemental-abundances

Leung, Henry. *Apogee*. SDSS. (n.d.). Retrieved April 9, 2023, from https://www.sdss4.org/dr17/irspec/

Wikimedia Foundation. (2023, April 10). Sunlight. Wikipedia. Retrieved April 11, 2023, from https://en.wikipedia.org/wiki/Sunlight