Star temperature and light intensity

The case of the star cluster CYGOB1

Louis SIRUGUE

08/2021

Data presentation

The package `HSAUR` comes with a data set of **2 variables** on the **47 stars** of the CYGOB1 star cluster in the Cygnus constellation:

- Effective temperature: $\log(T_e)$
- Light intensity: $\log(L/L_0)$

```
# Load necessary packages
invisible(lapply(c("HSAUR", "tidyverse", "stargazer"), library, character.only = TRUE))
# Load data
data("CYGOB1", package = "HSAUR")
# Look at the ten first observations
kable(t(head(CYGOB1, 10)), caption = "10 first observations") %>% column_spec(1, bold = T)
```

10 first observations

logst	4.37	4.56	4.26	4.56	4.30	4.46	3.84	4.57	4.26	4.37
logli	5.23	5.74	4.93	5.74	5.19	5.46	4.65	5.27	5.57	5.12

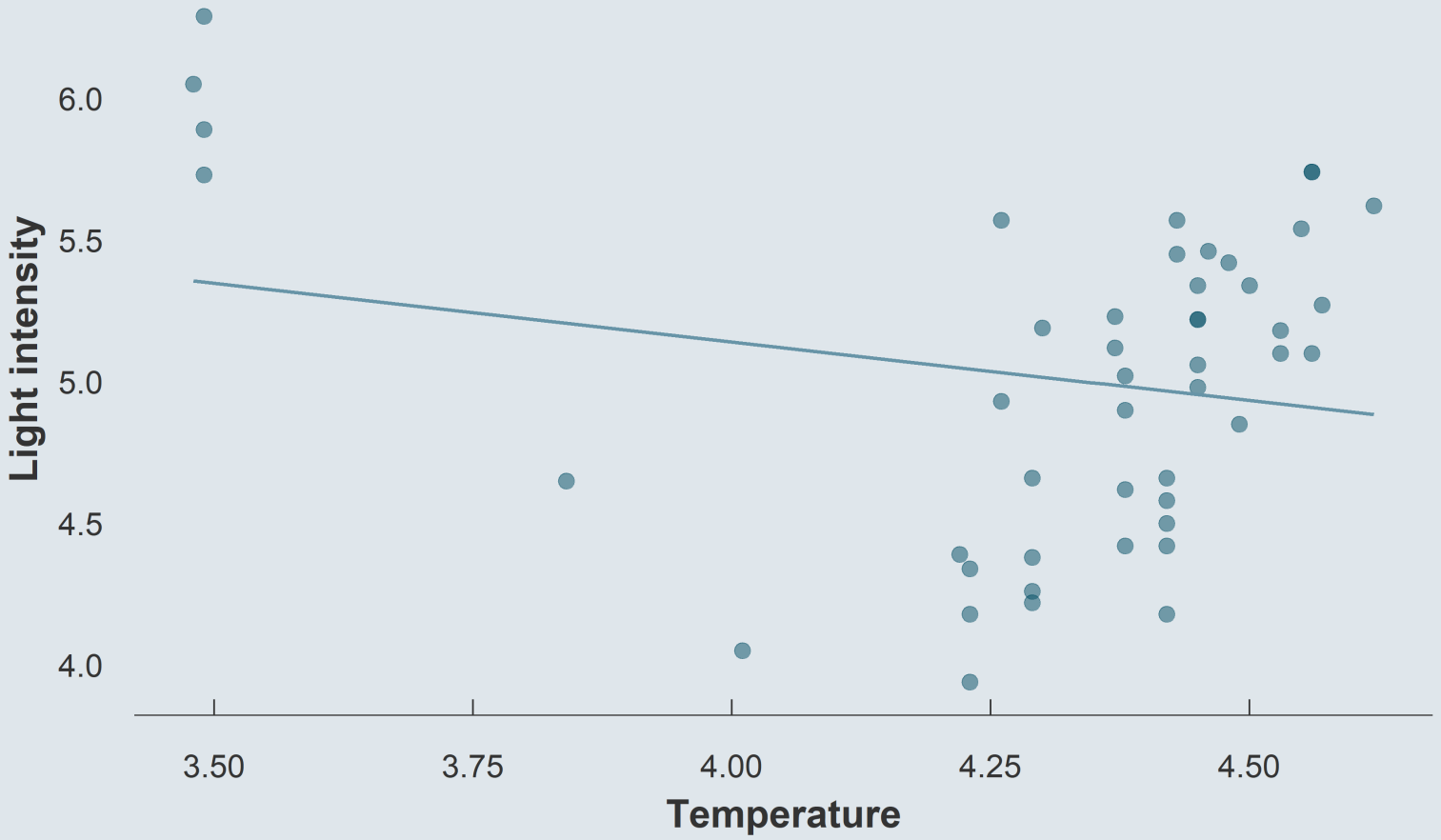
→ We would like to know the relationship between these two variables

Data visualization

Let's start with a regression line on top of a scatter plot of the light intensity against the temperature:

```
# Rename variables with convenient names
CYGOB1 <- CYGOB1 %>% rename(Temperature = logst,
                           `Light intensity` = logli)

# Specify the data to use for the plot
ggplot(data = CYGOB1,

       # Tell which variable should plotted on which axis
       aes(x = Temperature,  y = `Light intensity`)) +

  # Scatter plot
  geom_point(color = "#014D64", alpha = .5, size = 4) +

  # Fit of a linear regression model
  geom_smooth(method = "lm", se = F, color = "#6794A7") +

  # Styling of the plot
  theme_minimal(base_size = 18)
```

Data visualization



Something's wrong:

- The relationship is negative overall...
- ... but this seems to be fallaciously driven by four stars at the top left

Data manipulation

Indeed, the documentation of the data mentions that there are **two types of stars**:

- Stars that lie on the main sequence
- Giants stars

There's no variable in the data to distinguish between these two groups, but the documentation indicates that Giants are located at the following rows:

- The 11th;
- The 20th;
- The 30th;
- The 34th.

Based on these information, a variable indicating the type of star can easily be created as follows.

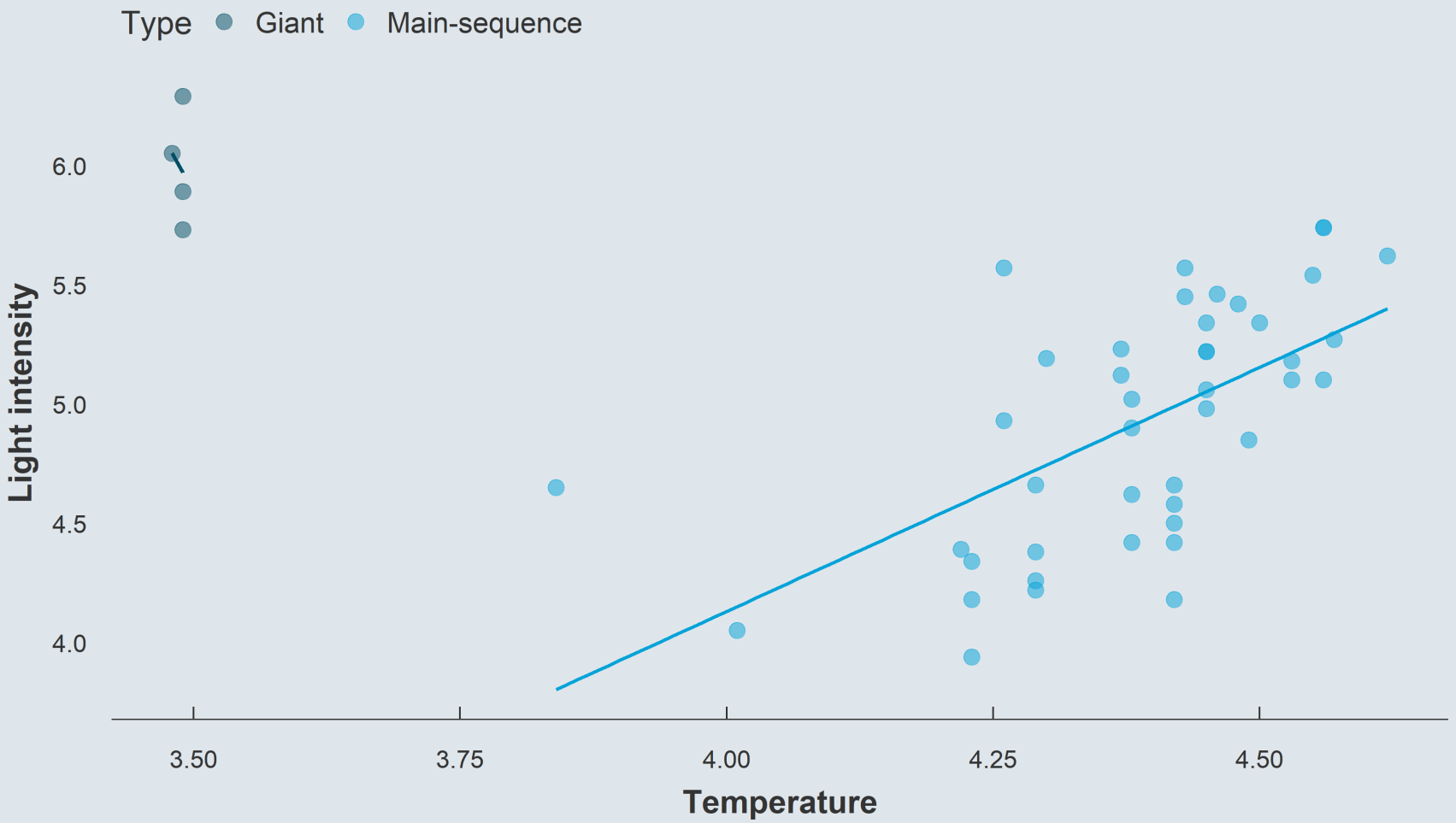
```
CYGOB1 <- CYGOB1 %>%
  mutate(Type = ifelse(row_number() %in% c(11, 20, 30, 34), # If this condition is met:
                       "Giant",                             # Type = Giant,
                       "Main-sequence"))                    # else: Type = Main-sequence
```

Data visualization

We can now distinguish the two types of stars in the plot by attributing them separate colors and regression lines:

```
# Same code as previous plot
ggplot(data = CYGOB1,
       aes(x = Temperature,
           y = `Light intensity`,
           color = Type)) + # But specifying to group by star type
  geom_point(alpha = .5, size = 4) +
  geom_smooth(method = "lm", se = F, show_guide = FALSE) +
  theme_minimal(base_size = 18) +

  # And specify the location and text-alignment of the legend
  theme(legend.position = "top", legend.justification = "left")
```

Data visualization



Conclusion

	Dependent variable:		
	Light intensity		
	Whole Sample	Giant	Main-sequence
Temperature	-0.413	-8.000	2.047***
	(0.286)	(33.307)	(0.420)
Constant	6.793***	33.890	-4.057**
	(1.237)	(116.157)	(1.844)
Observations	47	4	43
Adjusted R^2	0.023	-0.458	0.351
Note:			*p<0.1;**p<0.05;***p<0.01

By naively regressing light intensity on temperature, we would have **fallaciously** concluded that the two variables have a statistically non-significant **negative relationship**.

➜ *But data visualization allowed to notice that there is actually:*

- **No relationship for the Giants** (very large standard error/coefficient ratio)
- And a highly significant (p-value <0.01) **positive relationship for** stars that lie on the **main sequence**