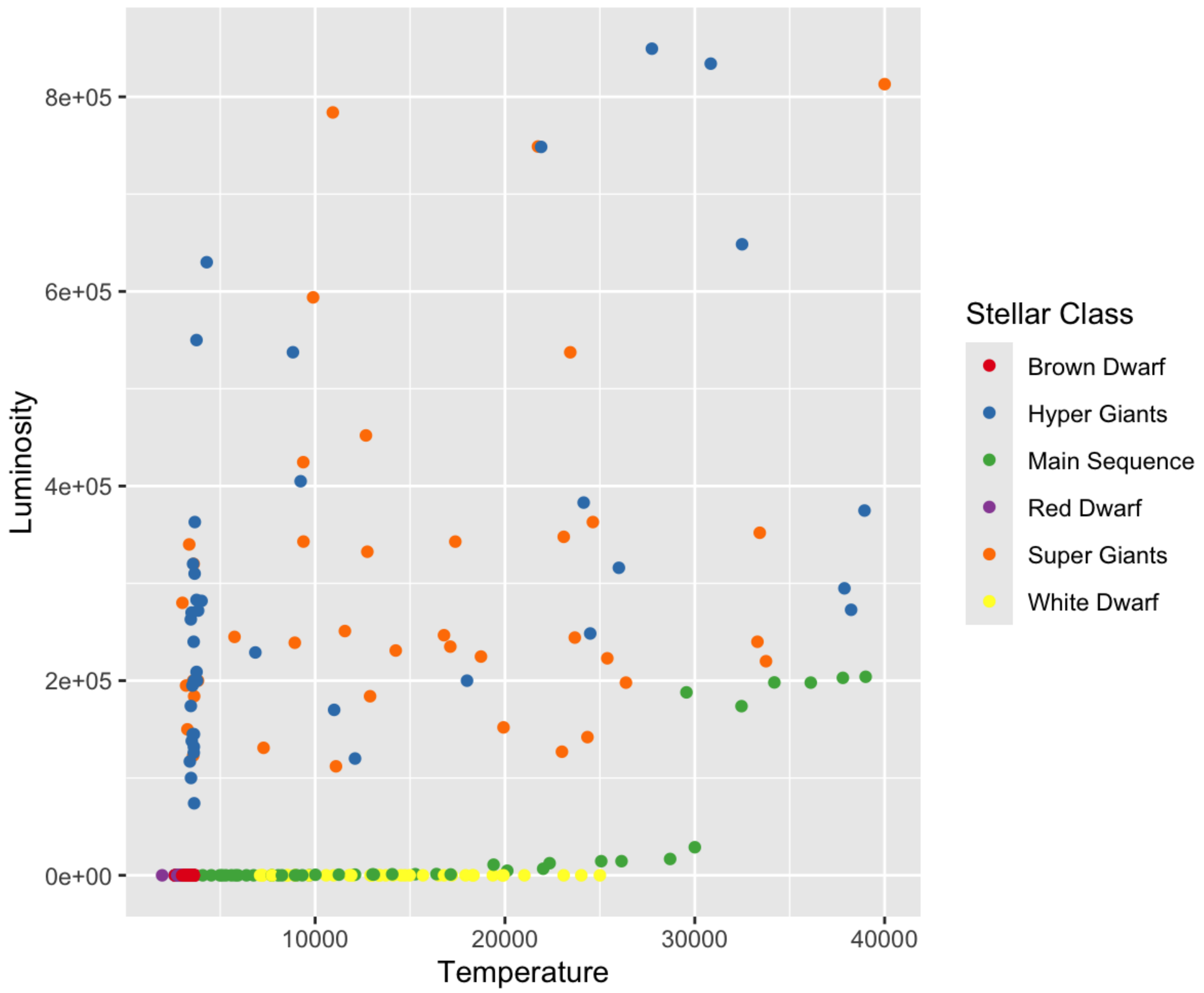
Stars Radiation Analysis

Mattia Bertoli
Matteo Campagnoli
Riccardo Girgenti
Benedetta Palmieri

Data Description

Numerical:
- Temperature
- Luminosity
- Radius
- Absolute Magnitude (AM)
Categorical:
- Stellar Class
240 observations:
- 6 classes, 40 observations per class

Hertzsprung Russell Diagram



Why should we study the luminosity?

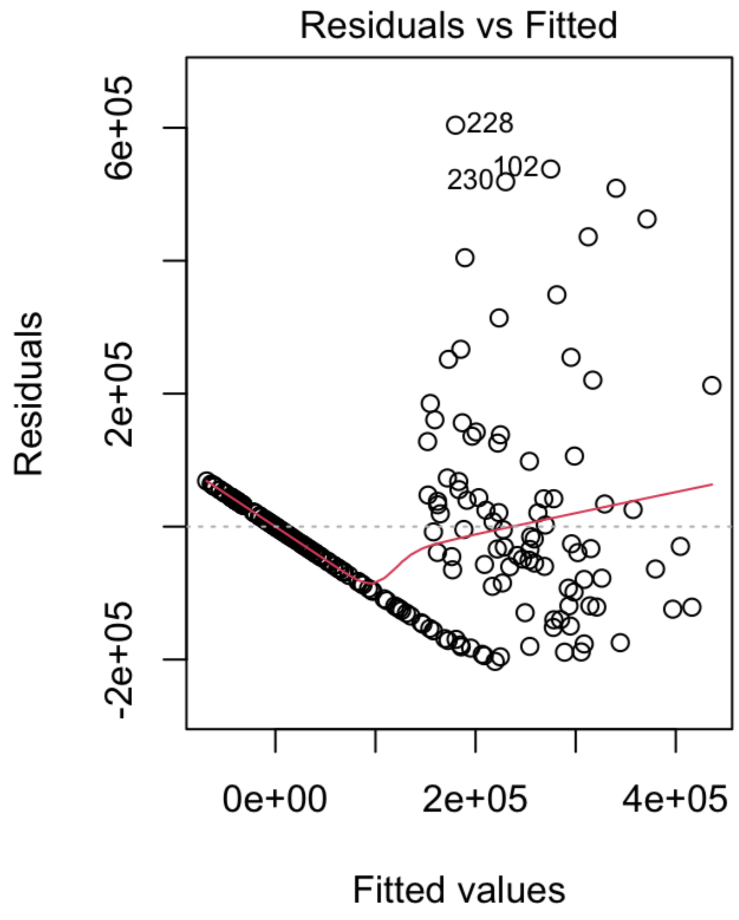
Studying a star's radiation reveals details about its chemical composition, age and other characteristics.
We are analyzing the relationship between luminosity and other stellar parameters using the following linear model:

$$\text{lm}(\ L \sim R + \text{Temperature} + AM\ )$$

Here, L is luminosity, R is radius, Temperature is surface temperature, and AM is absolute magnitude.

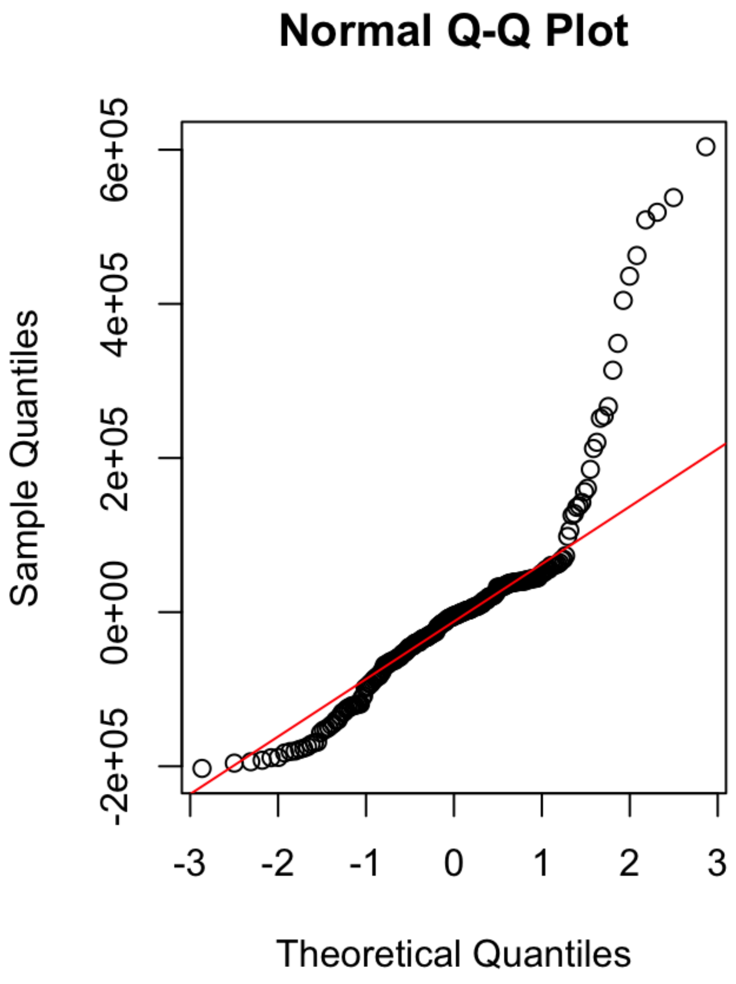
Linear Model with Numerical Variables

```
Call:
lm(formula = L ~ R + Temperature + A_M, data = stars)

Residuals:
    Min      1Q  Median      3Q     Max 
-202769  -62516   -4466   38187  603897 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 90679.7881 17857.9892   5.078 7.75e-07 ***
R              76.3555    20.4443   3.735 0.000236 ***
Temperature     3.2963     0.9677   3.406 0.000774 ***
A_M         -8260.8908  1103.9865  -7.483 1.43e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 125000 on 236 degrees of freedom
Multiple R-squared:  0.5208,    Adjusted R-squared:  0.5147 
F-statistic: 85.48 on 3 and 236 DF,  p-value: < 2.2e-16
```

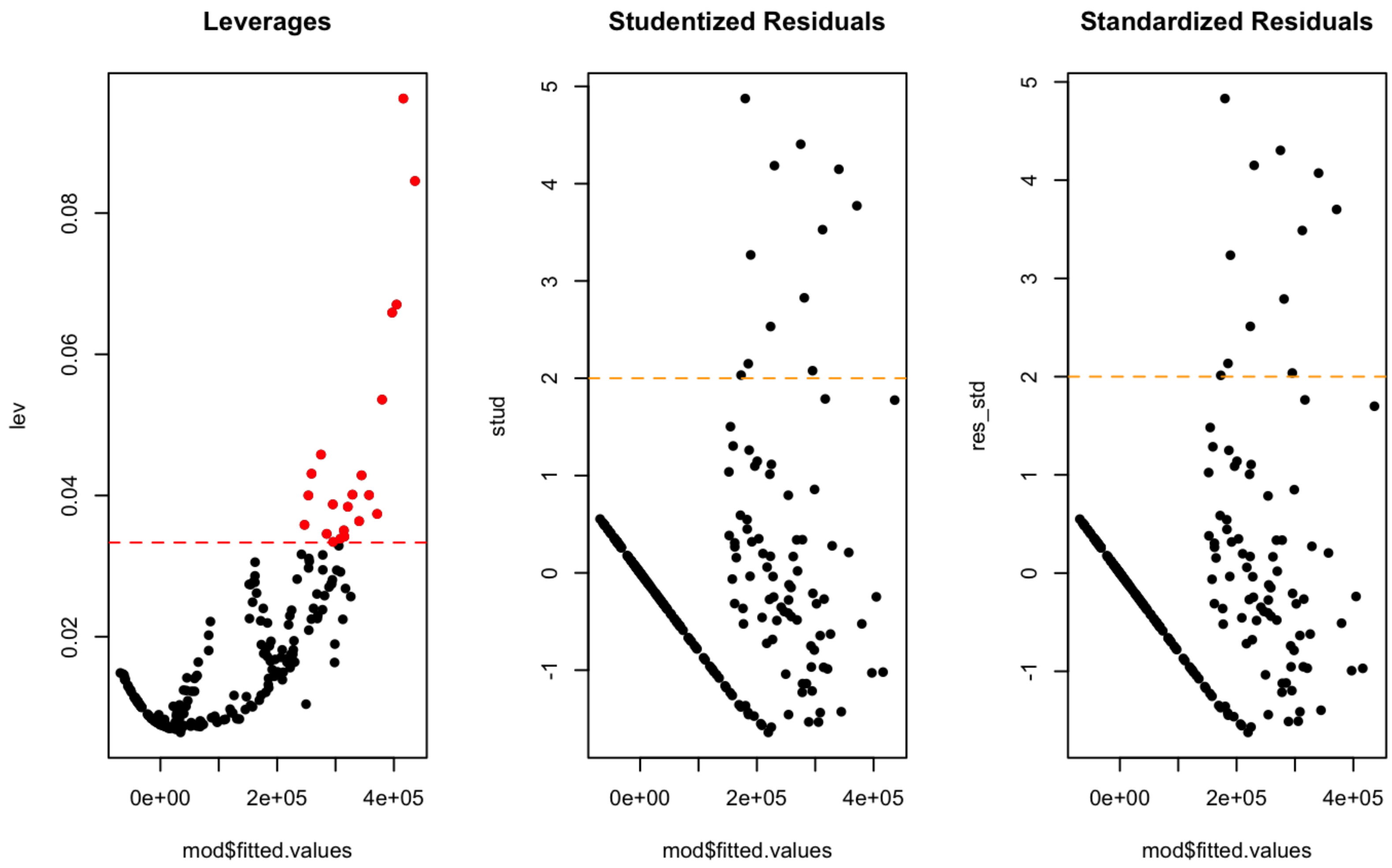


Linear Model with Numerical Variables

```
Call:
lm(formula = L ~ R + Temperature + A_M, data = stars)

Residuals:
    Min      1Q  Median      3Q     Max 
-202769  -62516   -4466   38187  603897 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 90679.7881 17857.9892   5.078 7.75e-07 ***
R              76.3555    20.4443   3.735 0.000236 ***
Temperature     3.2963     0.9677   3.406 0.000774 ***
A_M         -8260.8908  1103.9865  -7.483 1.43e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 125000 on 236 degrees of freedom
Multiple R-squared:  0.5208,	Adjusted R-squared:  0.5147 
F-statistic: 85.48 on 3 and 236 DF,  p-value: < 2.2e-16
```



Data cleaning and LM comparison



Data cleaning and LM comparison

g_post_lev <- lm(L ~ R + Temperature + A_M, data=stars, subset = (lev < 2 * p / n)):

- R-squared: 0.4939
- Adjusted R-squared: 0.4872
- AIC: 6070.846

g_post_stu <- lm(L ~ R + Temperature + A_M, data=stars, subset = (abs(res_stu) < 2)):

- R-squared: 0.6533,
- Adjusted R-squared: 0.6486
- AIC: 5765.064

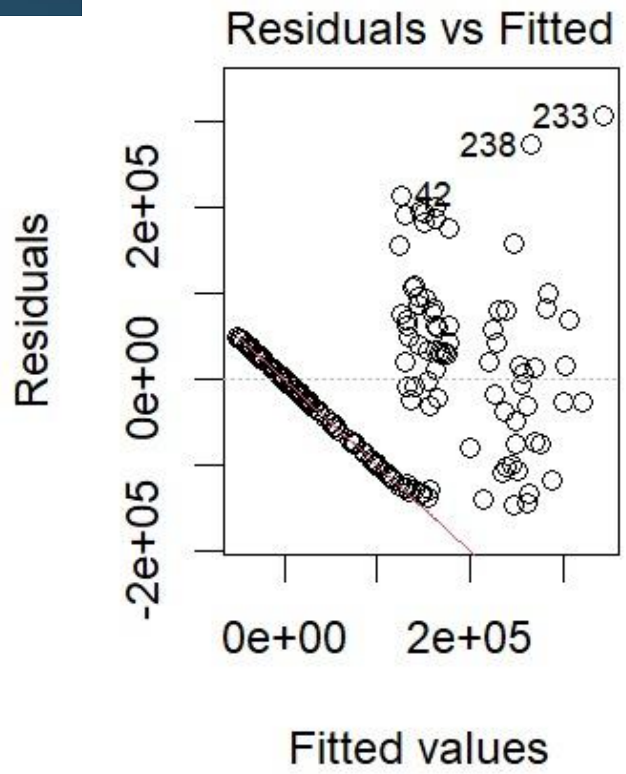
g_post_std <- lm(L ~ R + Temperature + A_M, data=stars, subset = (abs(res_std) < 2)):

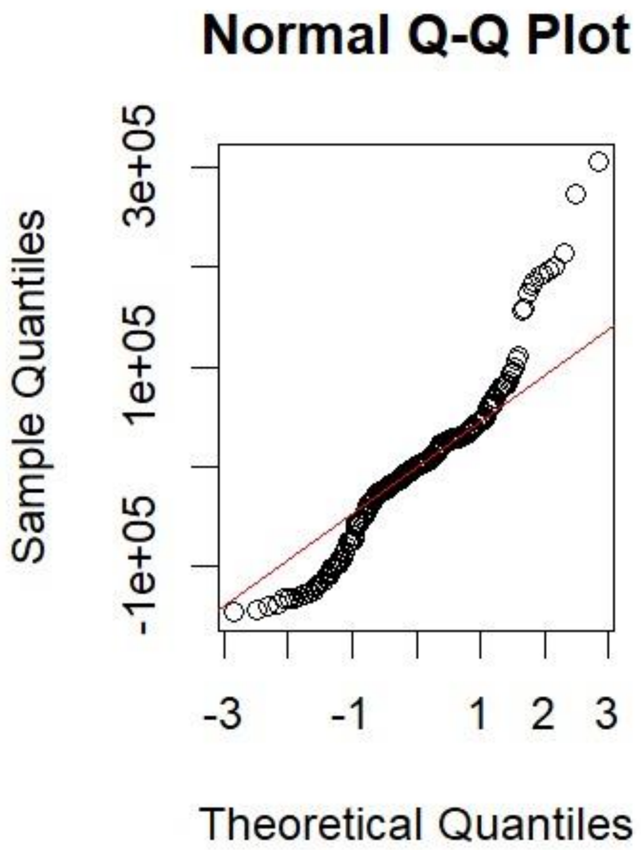
- R-squared: 0.6533
- Adjusted R-squared: 0.6486
- AIC: 5765.064

The models show varying fit levels, but none achieve good predictive accuracy. AIC values indicate some perform better, but all need significant improvement. Overall, these models require further refinement for reliable predictions.

Homoschedasticity Gaussianity

```
        Shapiro-Wilk normality test

data:  g_post_stu$residuals
W = 0.93281, p-value = 1.051e-08
```





The Shapiro-Wilk test indicates non-normality:

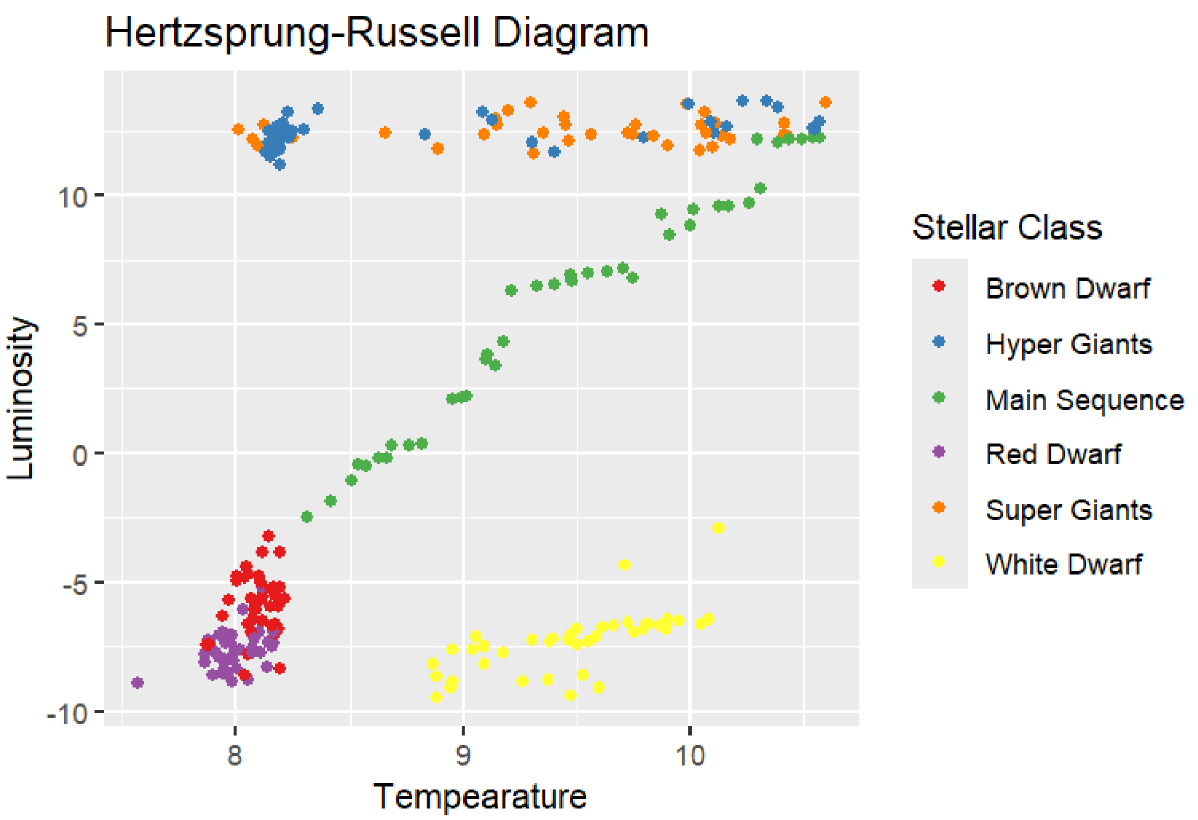
- p-value too low
- no homoschedastic patterns can be recognised

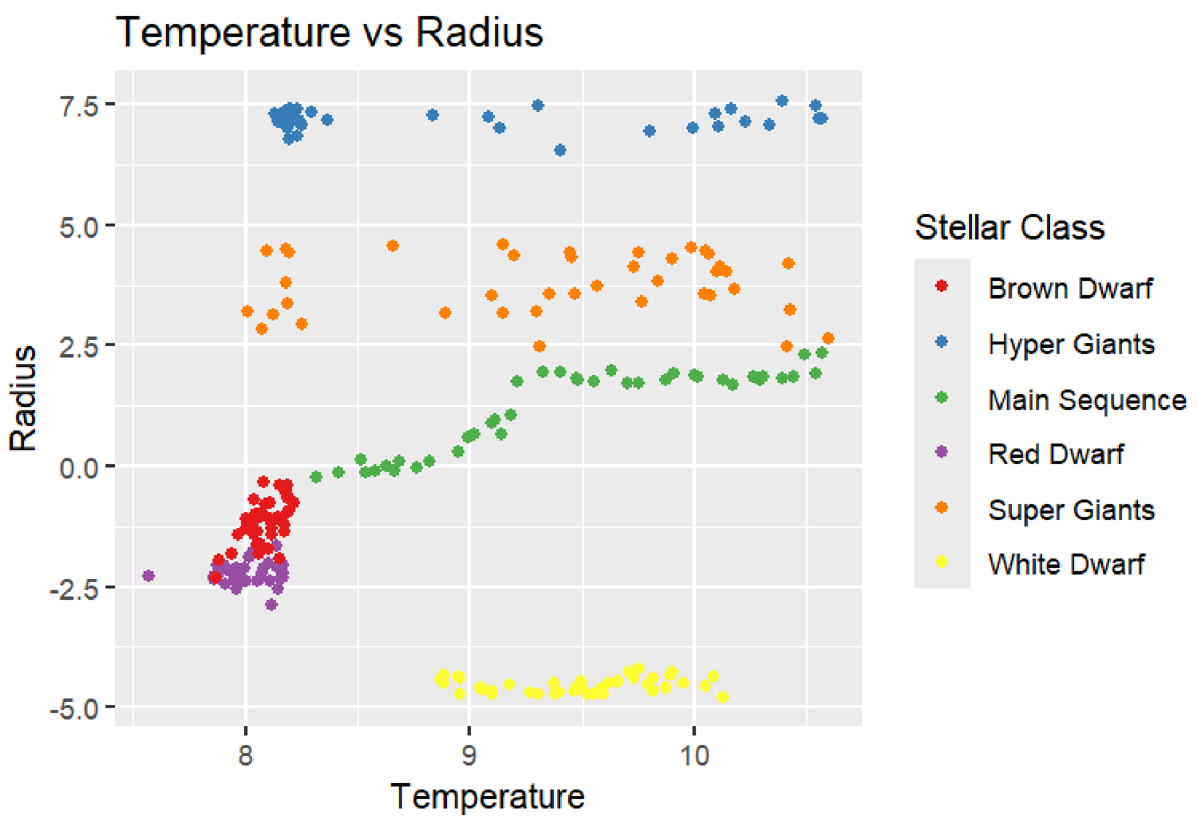
Data range

	MIN	MAX
Temperatura	1939	40000
Raggio	0.0084	1948.5
Luminosità	8x10^-5	849420
AM (log scale)	-11.92	20.06

It’s log time!

Logarithmic Plots



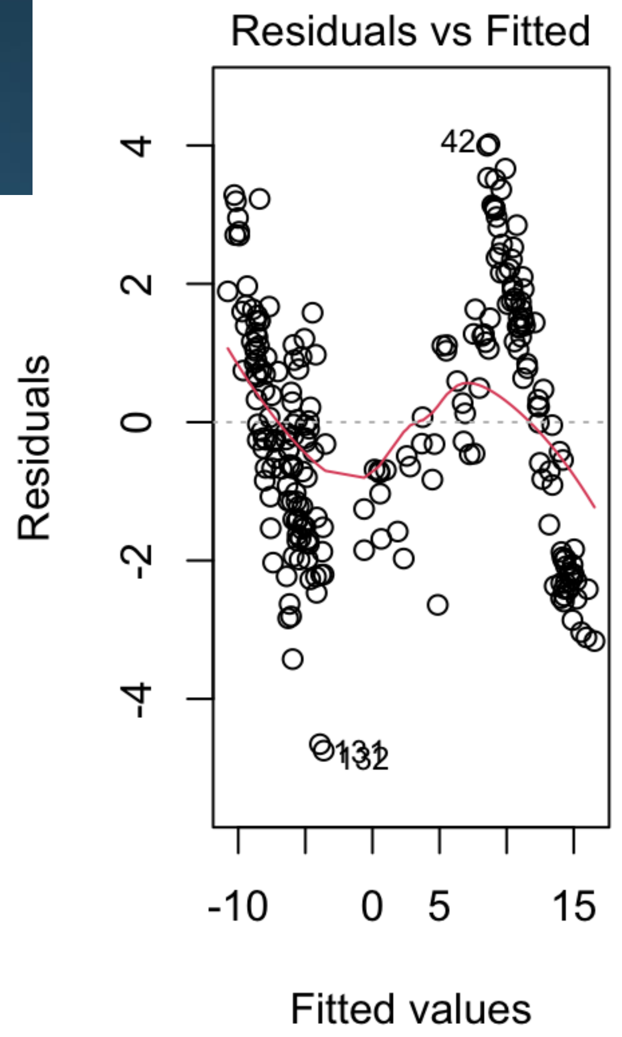


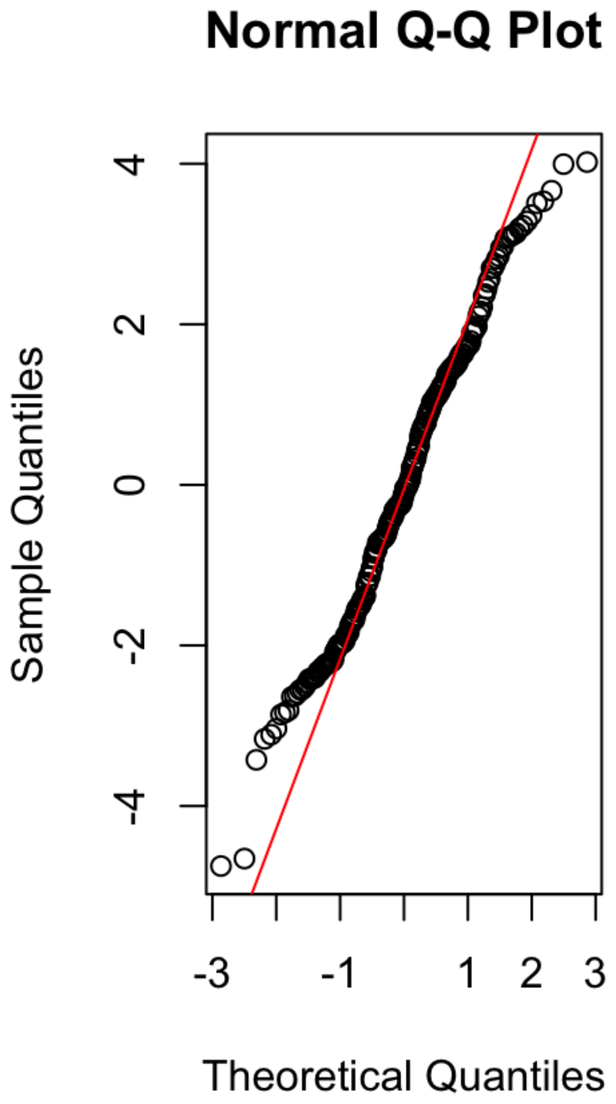
Linear Model with Numerical Variables

```
Call:
lm(formula = log_l ~ log_r + log_t + stars$A_M, data = stars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.7478 -1.4822 -0.1124  1.3709  4.0204 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.73323    1.89245  -2.501   0.0131 *  
log_r        0.66572    0.09246   7.200 7.99e-12 ***
log_t        0.95226    0.19396   4.910 1.70e-06 ***
stars$A_M   -0.58528    0.03794 -15.427  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.789 on 236 degrees of freedom
Multiple R-squared:  0.9621,	Adjusted R-squared:  0.9616 
F-statistic:  1995 on 3 and 236 DF,  p-value: < 2.2e-16
```

Homoschedasticity
Gaussianity

```
        Shapiro-Wilk normality test

data:  mod$residuals
W = 0.98523, p-value = 0.01375
```





The Shapiro-Wilk test indicates non-normality:

- p-value too low
- the residuals vs fitted plot and the Q-Q plot reveal deviations from homoscedasticity and normality, suggesting model improvements are needed

Linear Model with Categorical Variable

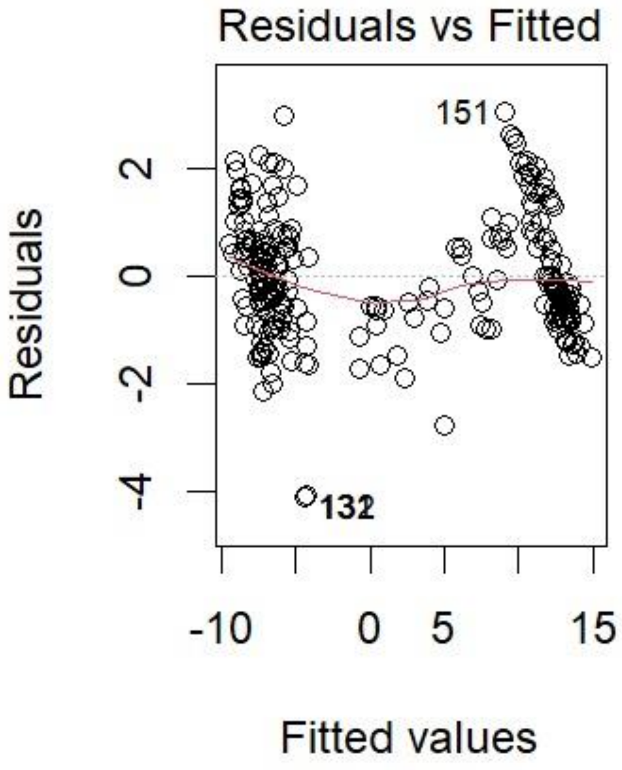
```
Call:
lm(formula = log_l ~ log_r + log_t + stars$A_M + stars$Type,
    data = stars)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.1064 -0.7755 -0.1748  0.6739  3.0283 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)              -4.48819    1.41392  -3.174  0.00171 ** 
log_r                     1.08361    0.18614   5.822 1.94e-08 ***
log_t                     0.88207    0.13880   6.355 1.10e-09 ***
stars$A_M                -0.57911    0.04986 -11.615  < 2e-16 ***
stars$TypeHyper Giants   -4.27753    1.39717  -3.062  0.00246 ** 
stars$TypeMain Sequence   0.25646    0.59521   0.431  0.66696    
stars$TypeRed Dwarf       2.41080    0.34573   6.973 3.23e-11 ***
stars$TypeSuper Giants    0.93783    0.92906   1.009  0.31382    
stars$TypeWhite Dwarf     1.01239    0.73841   1.371  0.17169    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.158 on 231 degrees of freedom
Multiple R-squared:  0.9844,    Adjusted R-squared:  0.9839 
F-statistic:  1826 on 8 and 231 DF,  p-value: < 2.2e-16
```

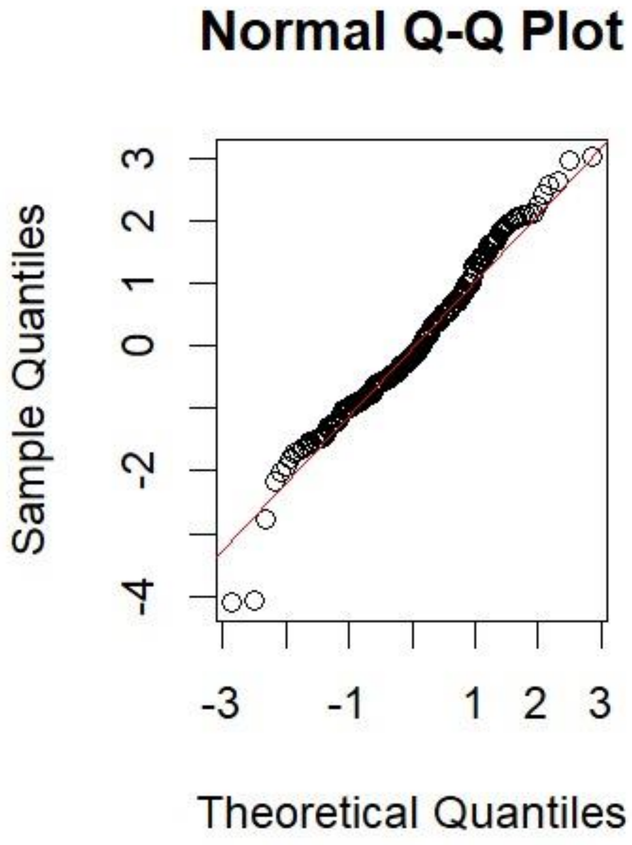
Homoschedasticity
Gaussianity

```
        Shapiro-Wilk normality test

data:  categorical$residuals
W = 0.97744, p-value = 0.0007173
```

The Shapiro-Wilk test indicates non-normality:

- p-value too low
- the residuals vs fitted plot and the Q-Q plot reveal deviations from homoscedasticity and normality, suggesting model improvements are needed





Models per Stellar Class

Let's divide the dataset into different classes of stars.
This is motivated by both:

- **Physical reasons:** stars belonging to different classes exhibit very different behaviors.
- **Statistical reasons:** within each class, the residuals are mostly normally distributed and we have enough samples per class.

Class	R^2	Shapiro Test
MAIN SEQUENCE	0.98	0.39
WHITE DWARF	0.50	0.008
BROWN DWARF	0.06	0.50
RED DWARF	0.16	0.33
SUPER GIANTS	0.16	0.40
HYPER GIANTS	0.30	0.09

Linear model Main Sequence

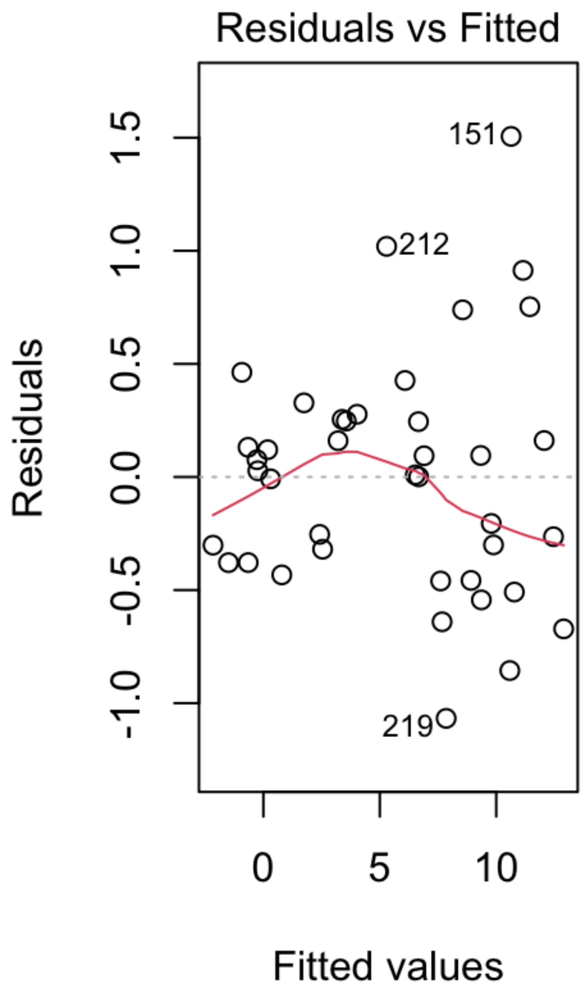
```
Call:
lm(formula = lum ~ radius + temp + AM, data = m_sequence)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.06723 -0.37858  0.01833  0.24918  1.50583 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -43.64838    4.12661 -10.577 1.36e-12 ***
radius        1.57606    0.28165   5.596 2.41e-06 ***
temp          5.01070    0.43923  11.408 1.64e-13 ***
AM            0.03079    0.10237   0.301    0.765    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5469 on 36 degrees of freedom
Multiple R-squared:  0.987,	Adjusted R-squared:  0.9859 
F-statistic: 912.1 on 3 and 36 DF,  p-value: < 2.2e-16
```

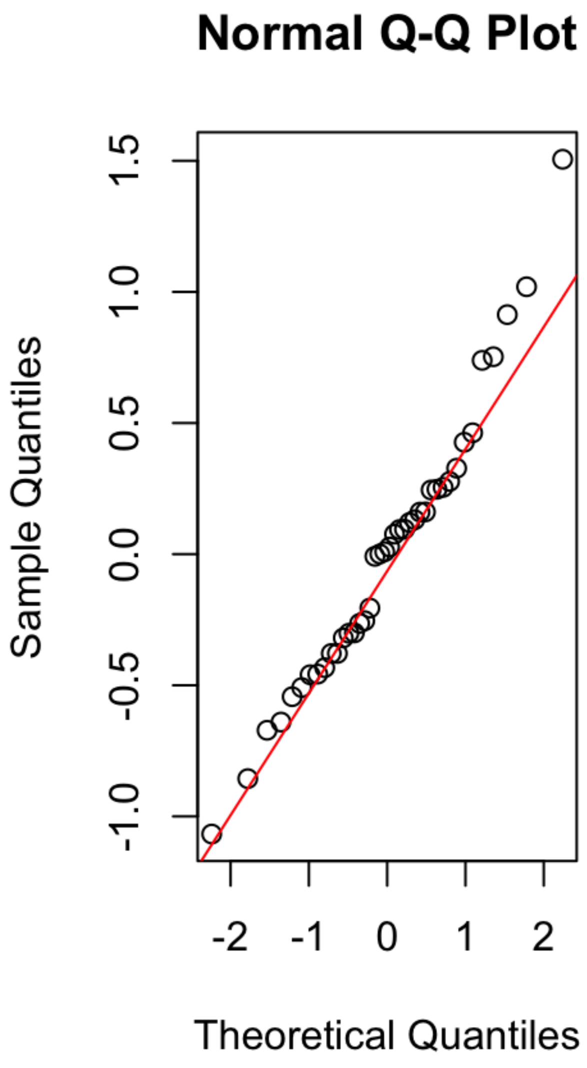
Homoschedasticity Gaussianity

```
Shapiro-Wilk normality test

data:  mod_m$residuals
W = 0.97107, p-value = 0.3889
```

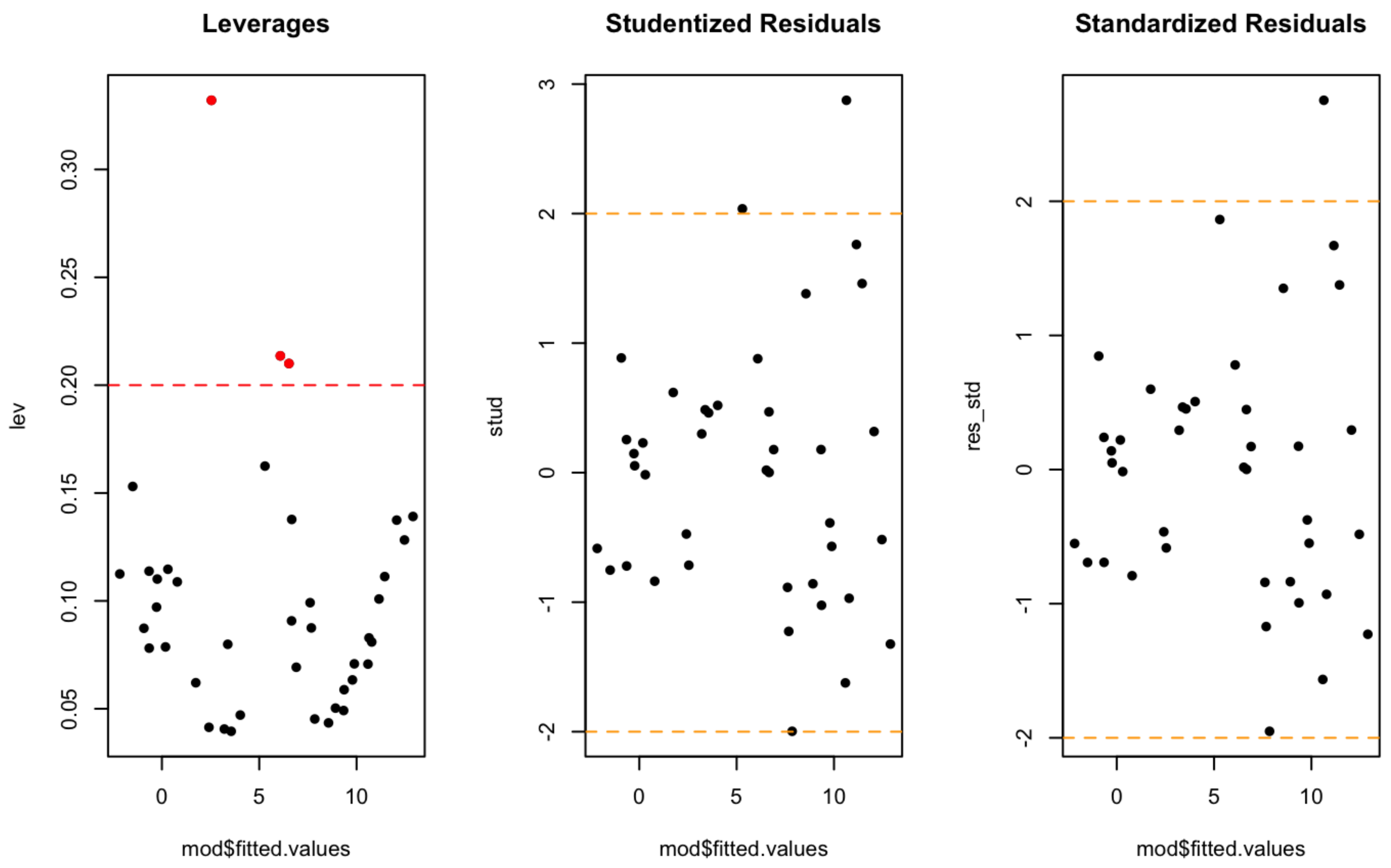
The Shapiro-Wilk test indicates normality:

- p-value good
- almost homoschedastic patterns can be recognised





Data cleaning and LM comparison



Data cleaning and LM comparison

```
mod_m_post_lev <- lm(lum ~ radius+temp+AM,data=m_seq, subset = (lev < 2 * p / n))
```

- R-squared: 0.9870
- Adjusted R-squared: 0.9859
- AIC: 70.4882

```
mod_m_post_stu <- lm(lum ~ radius+temp+AM,data=m_seq, subset = (abs(res_stu) < 2))
```

- R-squared: 0.9908
- Adjusted R-squared: 0.9900
- AIC: 54.5868

```
mod_m_post_std <- lm(lum ~ radius+temp+AM,data=m_seq, subset = (abs(res_std) < 2))
```

- R-squared: 0.9894
- Adjusted R-squared: 0.9885
- AIC: 60.31741

Each of the result is really good, but we will stick on the studentized residuals model because it has the lowest AIC

Main Sequence Model Adjustment

Removing A_M from the Model

```
Call:
lm(formula = lum ~ radius + temp, data = m_sequence, subset = (abs(res_stu) <
    2))

Residuals:
     Min       1Q   Median       3Q      Max
-0.98022 -0.35807  0.07085  0.27495  1.00496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -42.9083     2.2357 -19.192  < 2e-16 ***
radius        1.4495     0.2063   7.025 3.55e-08 ***
temp          4.9400     0.2596  19.031  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4534 on 35 degrees of freedom
Multiple R-squared:  0.9908,    Adjusted R-squared:  0.9903
F-statistic:  1892 on 2 and 35 DF,  p-value: < 2.2e-16
```

Shapiro-Wilk: W = 0.97509, p-value = 0.5457

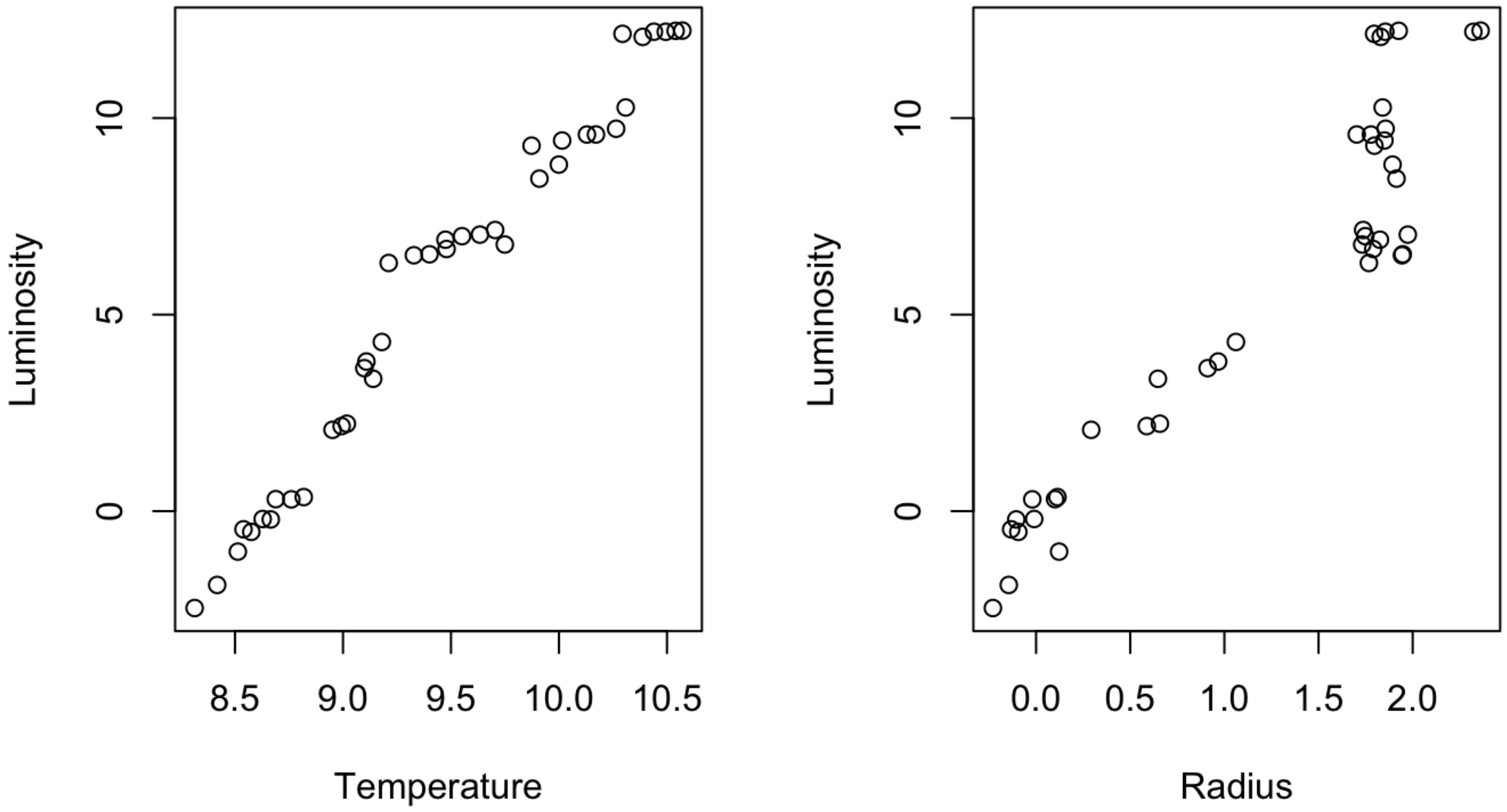
Stefan-Boltzmann Law

- $E = \sigma T^4$
- $E = \frac{L}{A}$
- $L = 4\pi R^2 \sigma T^4$
- $\log(L) = \log(4\pi R^2 \sigma T^4)$
- $\log(L) = \log(4\pi\sigma) + 2\log(R) + 4\log(T)$

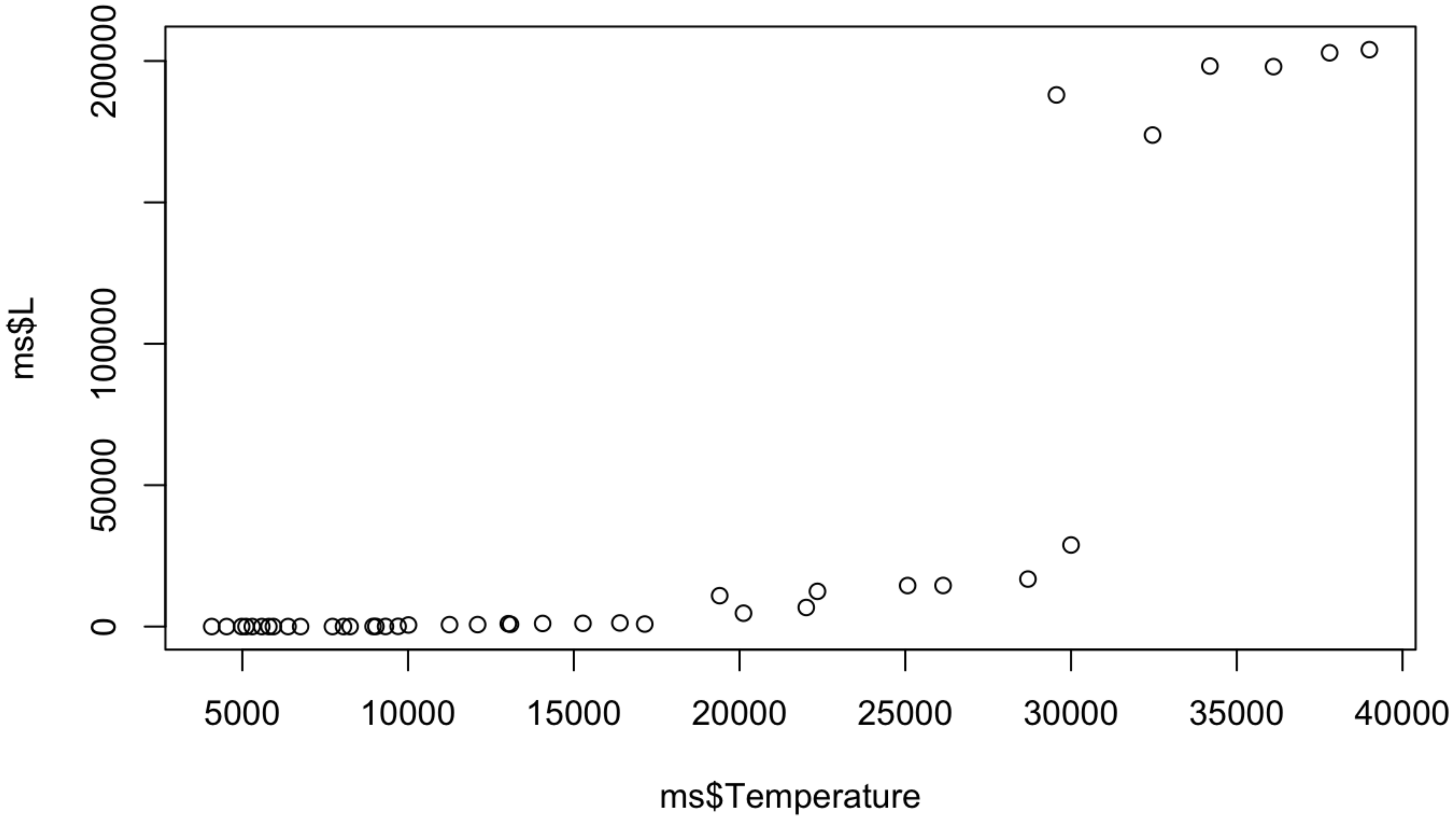
- E is the radiant energy emitted per unit area
- σ is the Stefan-Boltzmann constant ($5.67\text{x}10^{-8}\ \text{WK}^4/m^2$)
- T is the absolute temperature in Kelvin (K)
- A is the area of a sphere
- L is the luminosity
- R is the radius

LM Coefficients	Real	Fitted
Intercept*	-35.6	-42.9
Beta of log(R)	2	1.449
Beta of log(T)	4	4.940

Main Sequence data plots in log scale

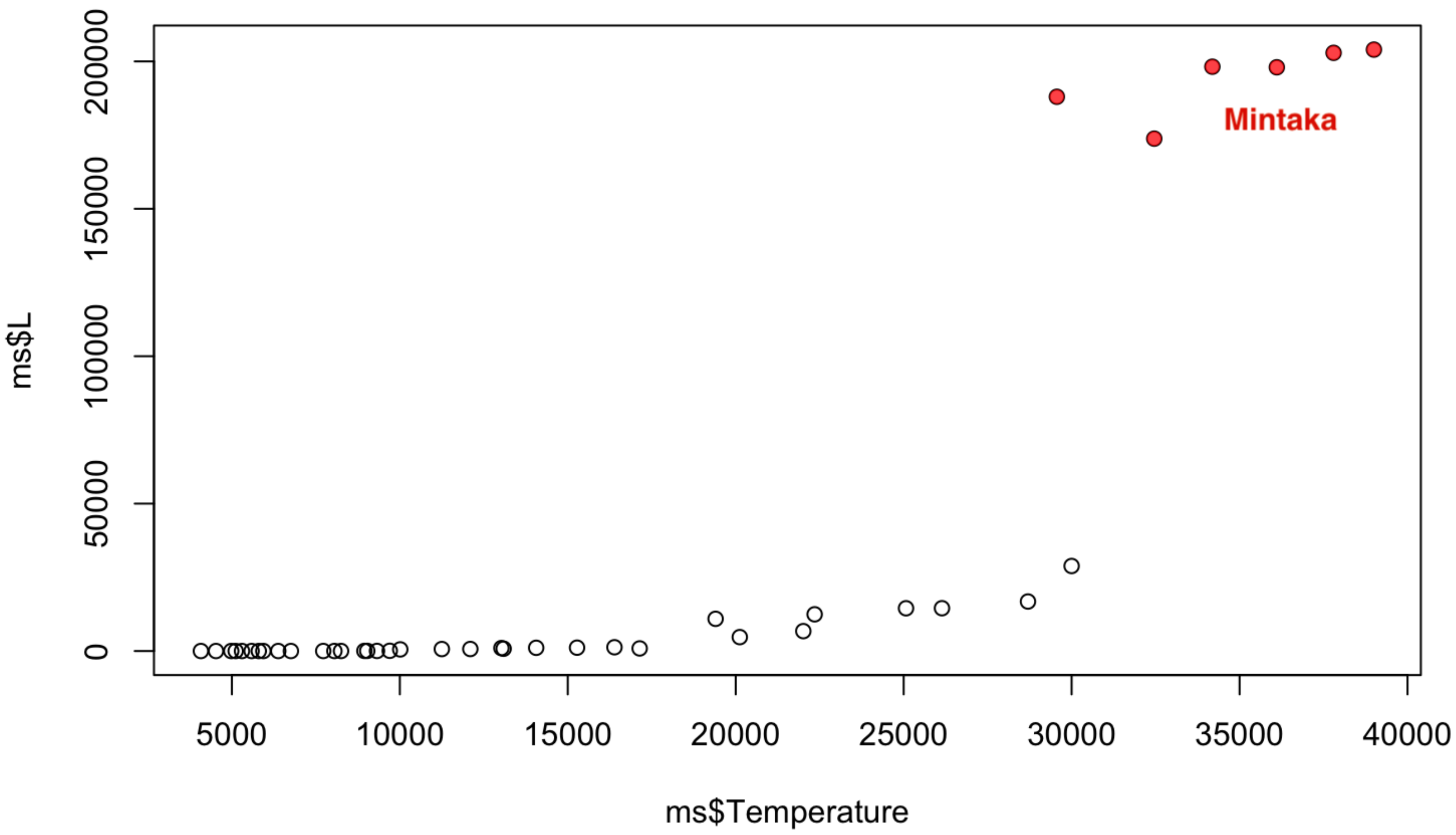


Main Sequence data plots in real scale



The star Mintaka (also called Delta Orionis) belongs to the Orion's Belt. Astronomers classify this star as a supergiant, contrary to what our dataset indicates. The same applies to the other stars marked in red on the graph. Let's try excluding Mintaka & friends from the main sequence and recalculate the model coefficients.

Main Sequence data plots in real scale



The star Mintaka (also called Delta Orionis) belongs to the Orion's Belt. Astronomers classify this star as a supergiant, contrary to what our dataset indicates. The same applies to the other stars marked in red on the graph. Let's try excluding Mintaka & friends from the main sequence and recalculate the model coefficients.

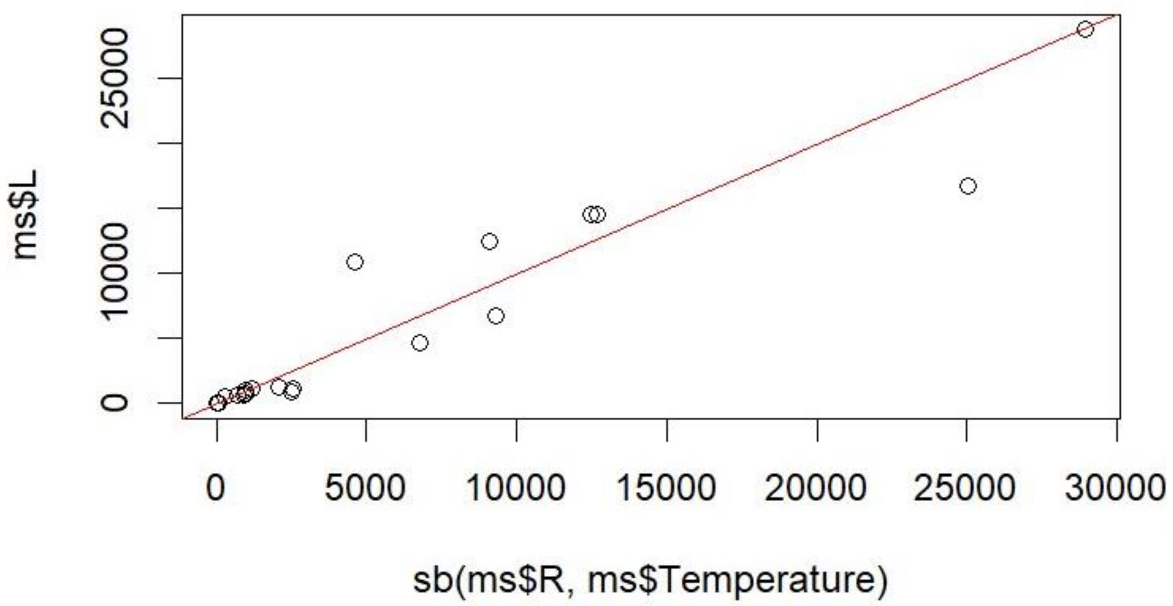
Excluding Mintaka & friends

Real
- 35.6
2
4

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.2294     2.3901  -15.16 6.96e-16
keep$radius   1.9300     0.1920   10.05 2.82e-11
keep$temp     4.1649     0.2768   15.05 8.51e-16
```

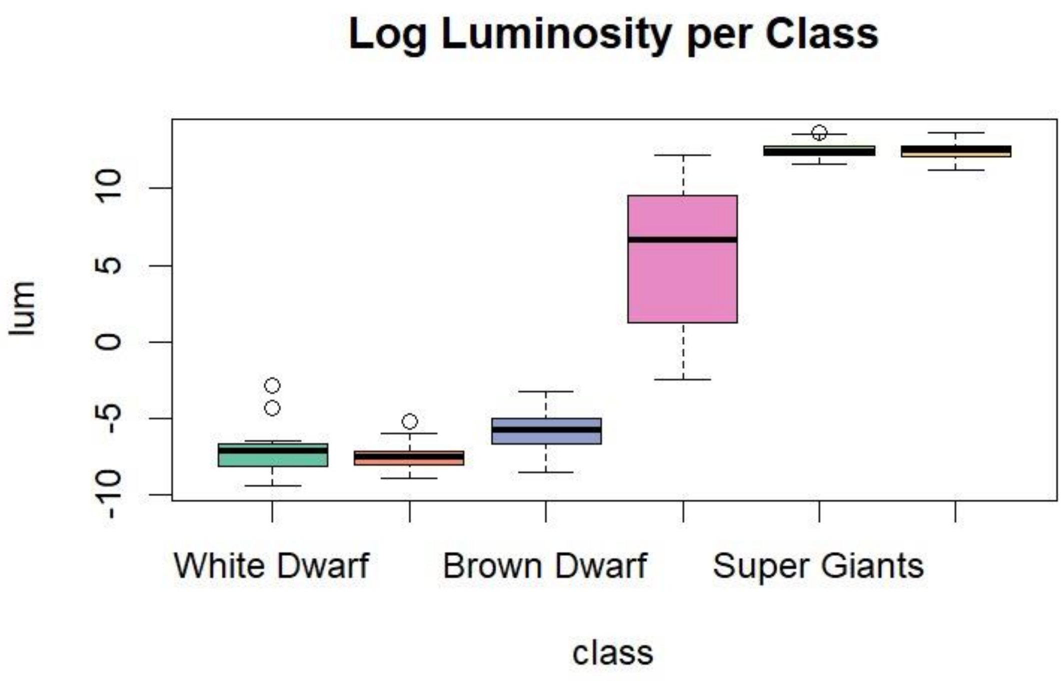
```
                   2.5 %     97.5 %
(Intercept) -41.104020 -31.354859
keep$radius   1.538468   2.321547
keep$temp     3.600382   4.729380
```

Theoretical Model vs Empirical Data



Why did we not ANOVA?

```
> tapply(df$lum, df$class, function(x) (shapiro.test(x)$p))
  White Dwarf     Red Dwarf   Brown Dwarf Main Sequence  Super Giants  Hyper Giants 
 0.0005946521  0.0464309491  0.9450750219  0.0166557821  0.1313175791  0.6141917272 
```




```
> leveneTest(df$radius, df$class)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)    
group   5  17.427 1.164e-14 ***
      234                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> bartlett.test(df$radius, df$class)

	Bartlett test of homogeneity of variances

data:  df$radius and df$class
Bartlett's K-squared = 162.26, df = 5, p-value < 2.2e-16
```