Rainfall model

Group_12

2021/6/22

Introduction

There are historic weather stations around the United Kingdom, which collects monthly data about the weather forecast in their location. We have selected data from weather station from Paisley, Glasgow and we have taken a subset of all the months between 1959 to 1999. [1]

The variables that are measured by the station are: mean daily maximum temperature (degree Celsius), mean daily minimum temperature (degree Celsius), air frost (days), total rainfall (millimetres), and total sunshine duration (hours). The values are recorded each month of the year.

We will be exploring if we can predict the total rainfall using the other measured variables. The total rainfall within the area is an important factor in range of areas, for example gives insight into Paisley's ecology or even benefitable for the council to prepare for adverse weather conditions. [2]

Additionally, we are interested in seeing if there is a significant linear relationship between the total rainfall and other potential variable. We found that there is weak linear relationship between the total rainfall and mean daily maximum(minimum) temperature. So it is an interesting question that if the temperature difference(degree Celsius) produced by mean daily maximum(minimum) temperature is a significant variable.

Using the observations recorded each month, we will apply linear regression to create a model to predict the total rainfall within Paisley. Furthermore, we shall use bootstrap method to find the linear relationship between the total rainfall and temperature difference.

[1] MetOffice UK, 2021, Historic station data, viewed 29 June 2021, https://www.metoffice.gov.uk/research/climate/maps-and-data/historic-station-data

[2] Schultz, C 2011, 'Rainfall: State of the Science', AGU Geophysical Monograph Series, vol. 92, no. 43, pp. 378

Numerical Summaries

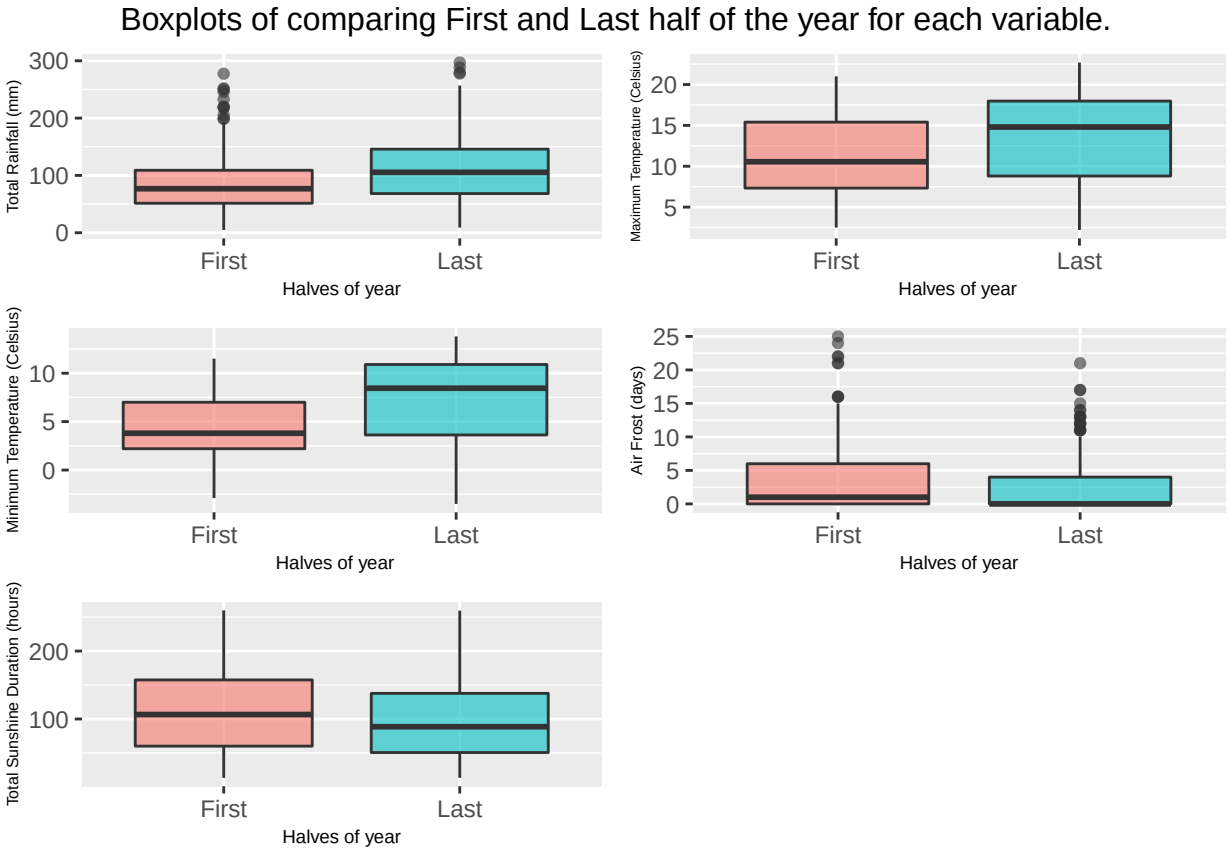


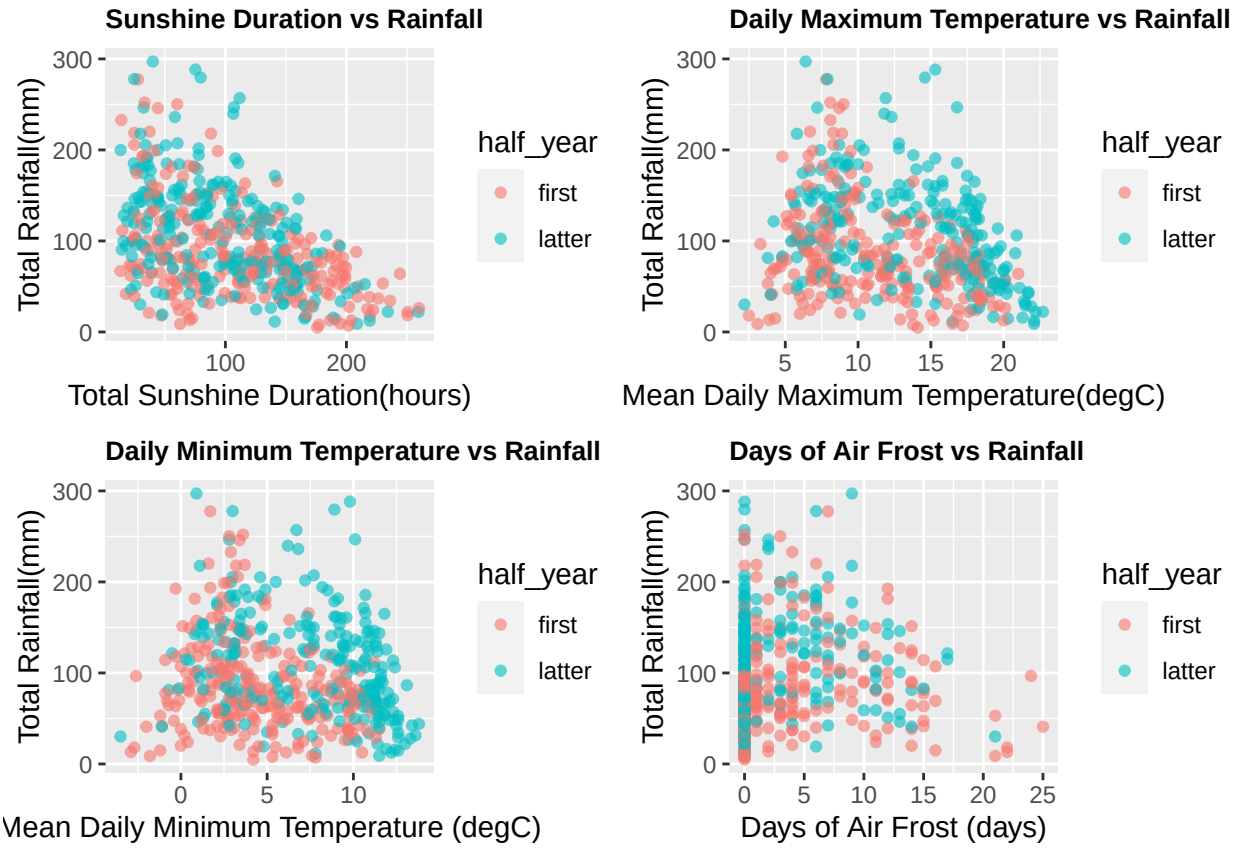
Table 1: P-values from Wilcoxon's Signed Ranks test (paired data)

Variable Name	P-Value
Max temperature	2.72504537011535e-05
Min temperature	4.54917118750366e-10
Air frost	0.00371906194298542
Total rainfall	3.48169763250871e-06
Total sunshine duration	0.0277977223990777

[Draft] what do boxplots shows and why use them? comparison and distribution. The data is paired so apply Wilcoxon signed rank test, we use mu=0 and see if sig difference between each one by calc pvalues. As we can see, all variables median have stat sig difference between first and last of the year. [make sure to lay out mathematically]

Model Selection

The data set has five continuous variables and one categorical variable, so I'll choose suitable continuous variables as explanatory variable at first as rainfall has been the response variable.



Variable Selection

Our approach is using confidence intervals.

Firstly, we fit the most general model, i.e.$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \epsilon_i$

Table 2: Estimate summaries from the MLR Model

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	191.685	16.775	11.427	0.000	158.724	224.645
tmax	-4.229	3.663	-1.155	0.249	-11.426	2.967
tmin	4.348	3.648	1.192	0.234	-2.819	11.516
af	-3.689	0.704	-5.238	0.000	-5.073	-2.305
sun	-0.537	0.108	-4.955	0.000	-0.750	-0.324

According to the **??**, all of the 95% CIs for the parameters in the model contain zero except that for af(-5.073,-2.305) and sun(-0.750,-0.324), therefore we can conclude that tmax and tmin does not contribute significantly to the model and thus remove them from the model and refit the model with af and sun.

So we choose the af and sun as our explanatory variables and estimate the model again.

Model Comparisons

Then we can check the conclusion by calculating some objective criterias such as R^2_{adj},AIC and BIC.

Table 3: Estimate summaries from the refitted MLR Model

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	176.100	5.949	29.599	0	164.410	187.790
af	-3.827	0.507	-7.547	0	-4.823	-2.830
sun	-0.643	0.044	-14.655	0	-0.729	-0.556

Table 4: Model comparison values for different models

Model	adj.r.squared	AIC	BIC
MLR(general)	0.3	5173.66	5198.85
MLR(refit)	0.3	5171.09	5187.89

In the **??**, the refitted model has the same R^2_{adj} value and lower AIC and BIC values. However, the high AIC and BIC values and very low R^2_{adj} value could suggest that these models is not a good fit to the data.

Linear Models

Table 5: linear models

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	161.39	8.87	18.20	0.00	143.97	178.81
af	-3.28	0.66	-5.00	0.00	-4.56	-1.99
sun	-0.56	0.06	-9.48	0.00	-0.68	-0.45
half_yearlatter	23.85	12.45	1.92	0.06	-0.61	48.30
af:half_yearlatter	-0.68	1.11	-0.61	0.54	-2.86	1.50
sun:half_yearlatter	-0.13	0.09	-1.43	0.15	-0.31	0.05

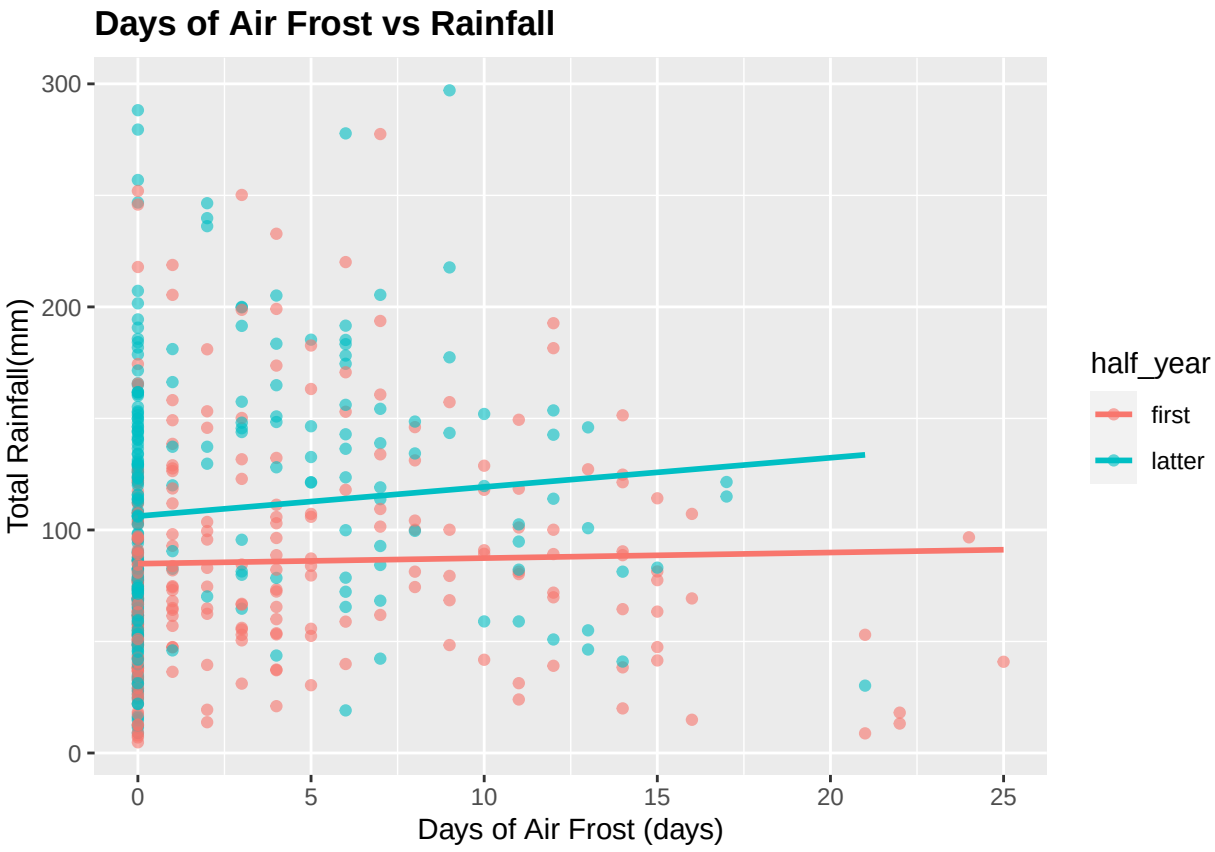
The parameter estimates are in \ref{tab:model}, hence the regression line for the month in the first half year is given by:

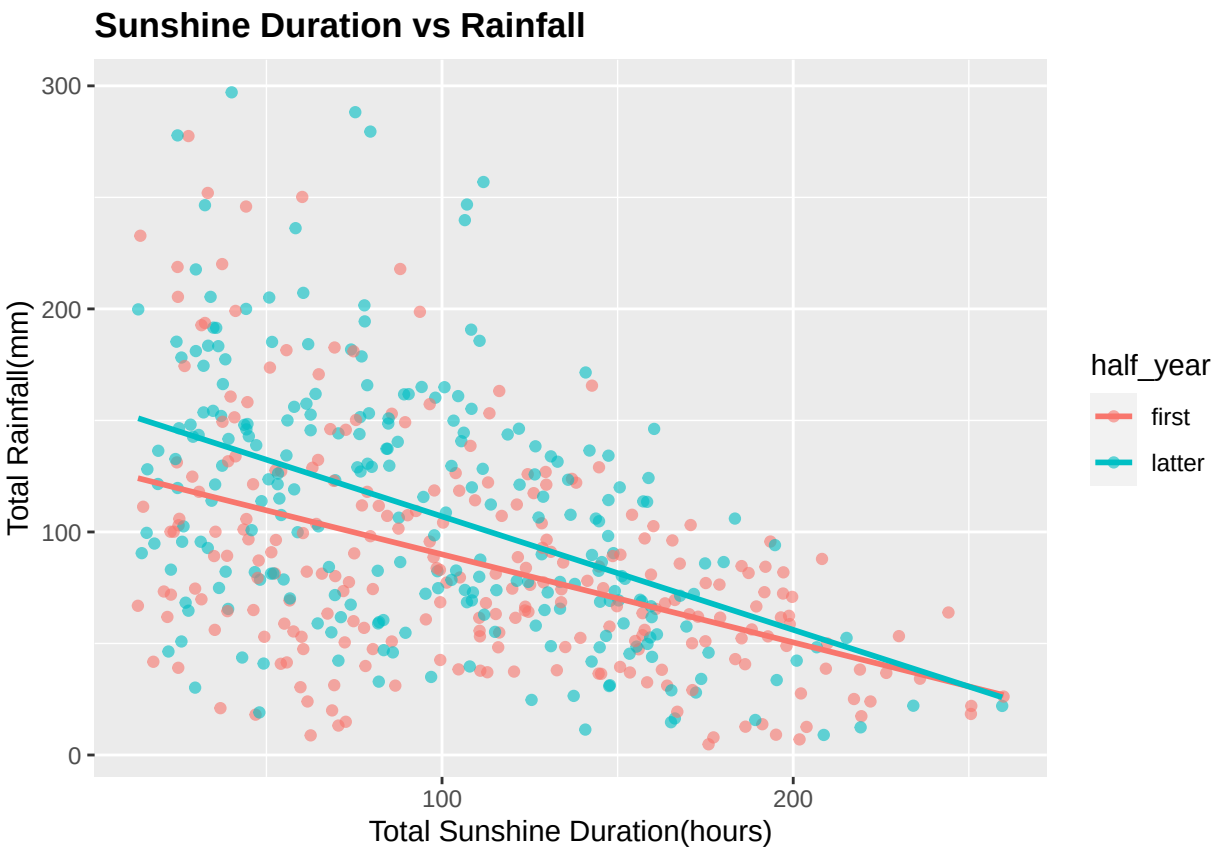
$$\widehat{Rainfall} = 161.39 - 3.28 \cdot af - 0.56 \cdot sun$$

while the regression line for the month in the latter half year is given by:

$$\widehat{Rainfall} = 185.24 - 3.96 \cdot af - 0.69 \cdot sun$$

To see the linear relationship between these variables, we plot scatterplots:

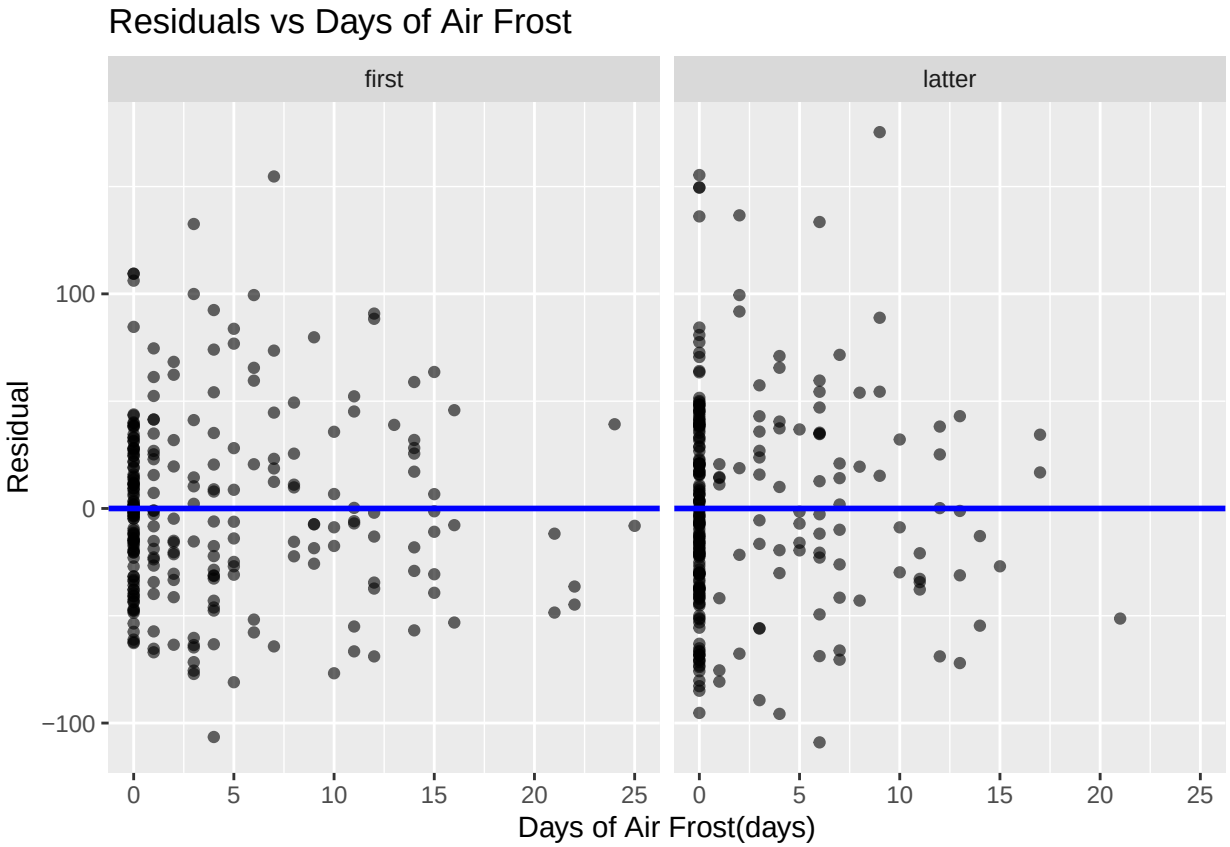

Days of Air Frost vs Rainfall
Total Rainfall(mm)
300
200
100
0
0
5
10
15
20
25
Days of Air Frost (days)
half_year
first
latter

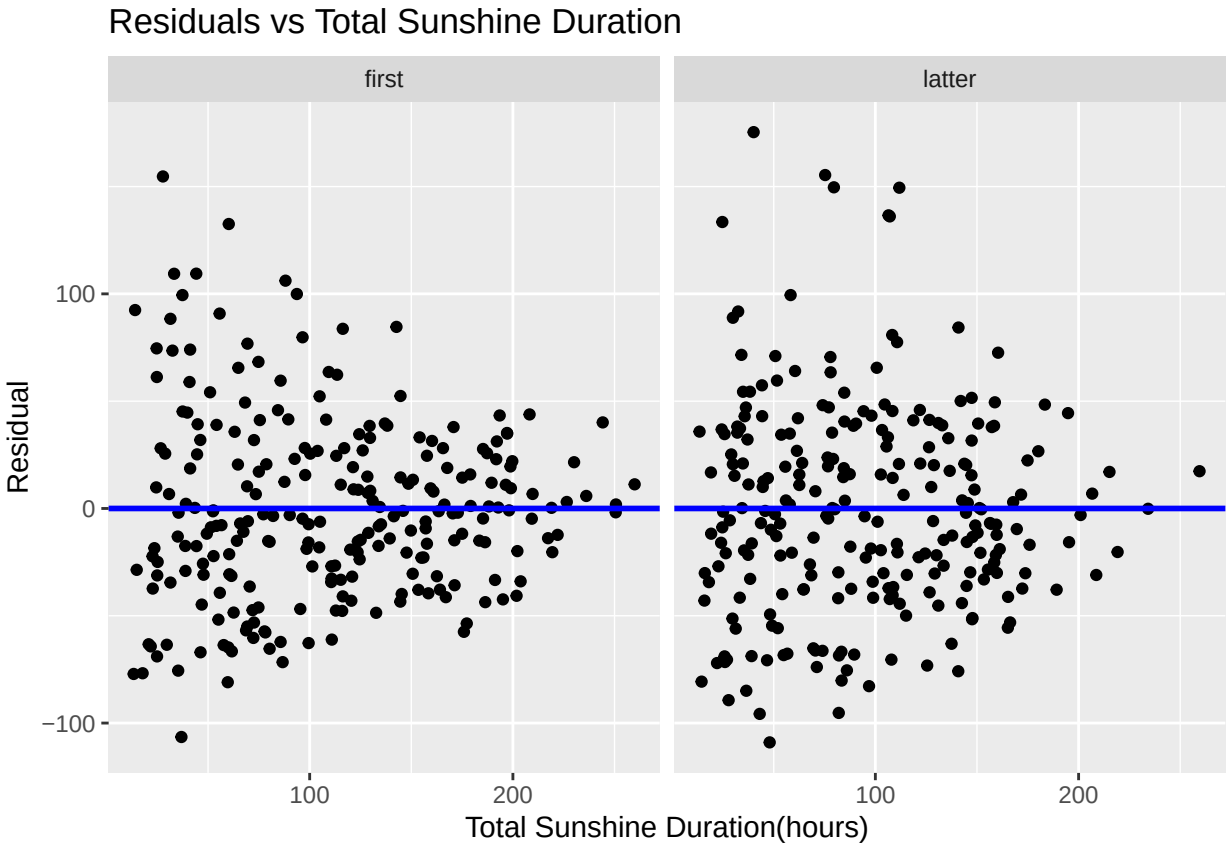


According to the plot above, the linear relationship between the explanatory variables and outcome variables seems not obvious.

Assessing Model fit

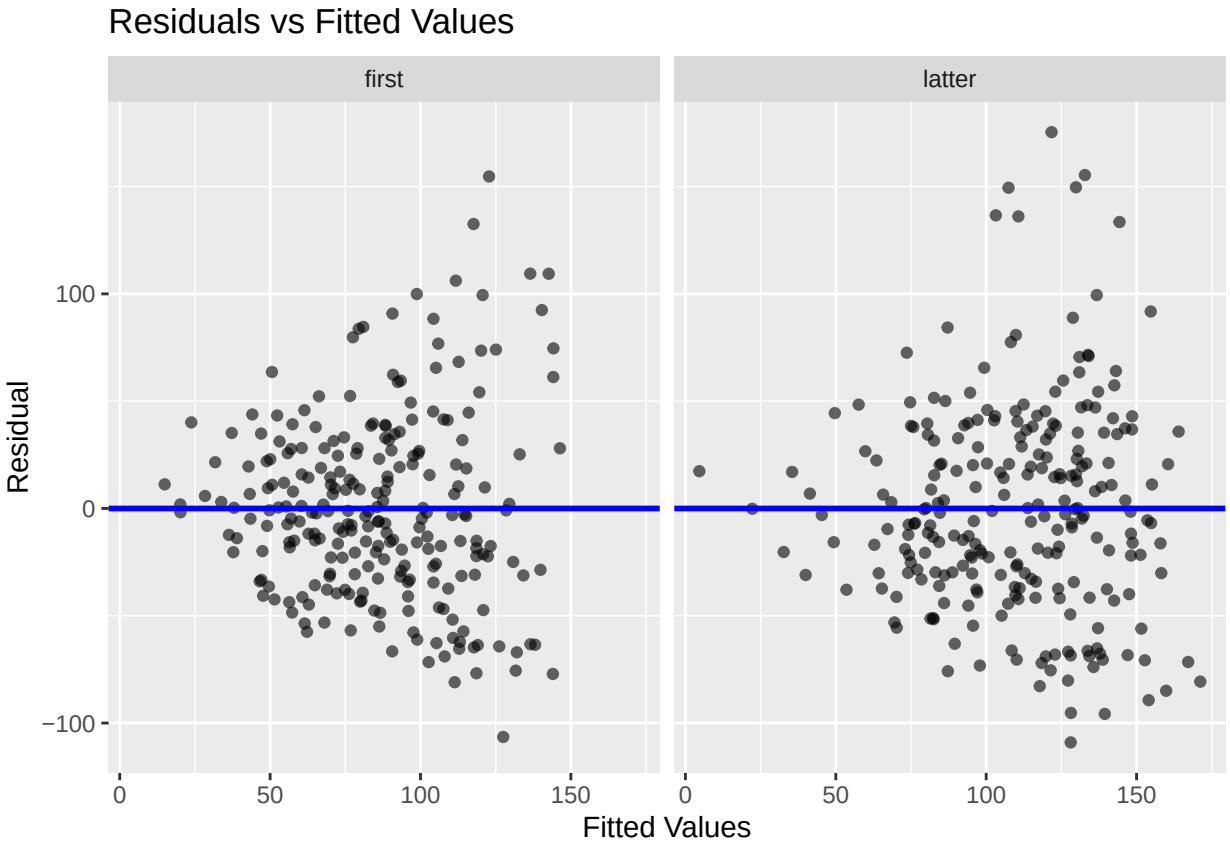
After producing the model, we shall check our model assumptions. First of all, we need to plot the scatterplot of the residuals against variables af and sun:


Residuals vs Days of Air Frost
first
latter
Residual
100
0
−100
0
5
10
15
20
25
Days of Air Frost(days)



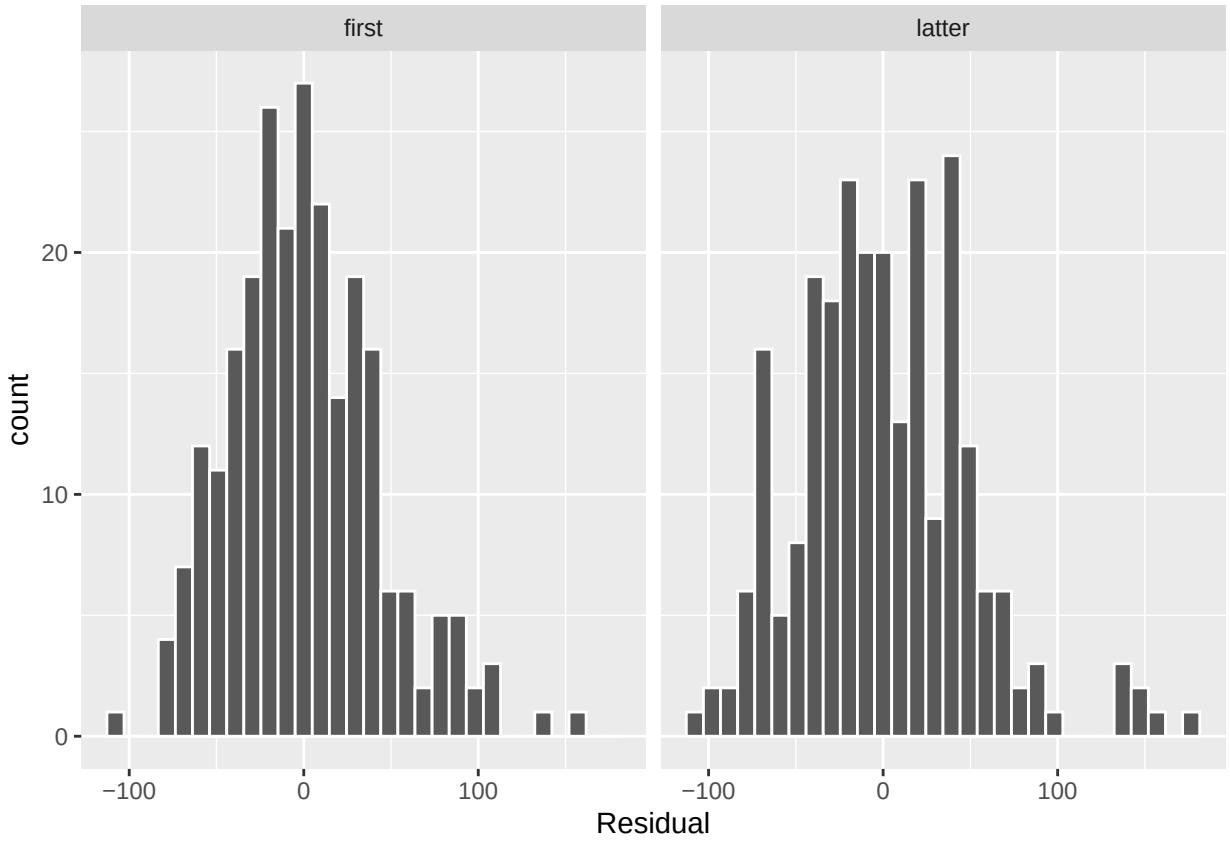
The plot shows that there is an even scatter of points above and below the zero line indicating the residuals have mean zero. And the scattering of the points is also constant across all values of the explanatory variable with no systematic pattern observed in the residuals. So this assumption is valid.

Then we plot the residuals against the fitted values:



In this case, the plot shows that there is an even scatter of points above and below the zero line indicating the residuals have mean zero. And the scattering of the points is also constant across all values of the explanatory variable with no systematic pattern observed in the residuals. So this assumption is also valid.

Then we need to check if the residuals follow the normal distribution.



Our subjective impression is that the residuals appear to be bell-shaped. It means the normality of the residuals is good.

To conclude, the assumptions is valid in this model and the model is an appropriate fit to the data.

CI for Regression Parameters

CIs have been shown in Table 5.

Conclusion

Hence we built the regression model for the month in the first half year which is :

$$\widehat{Rainfall} = 161.39 - 3.28 \cdot af - 0.56 \cdot sun$$

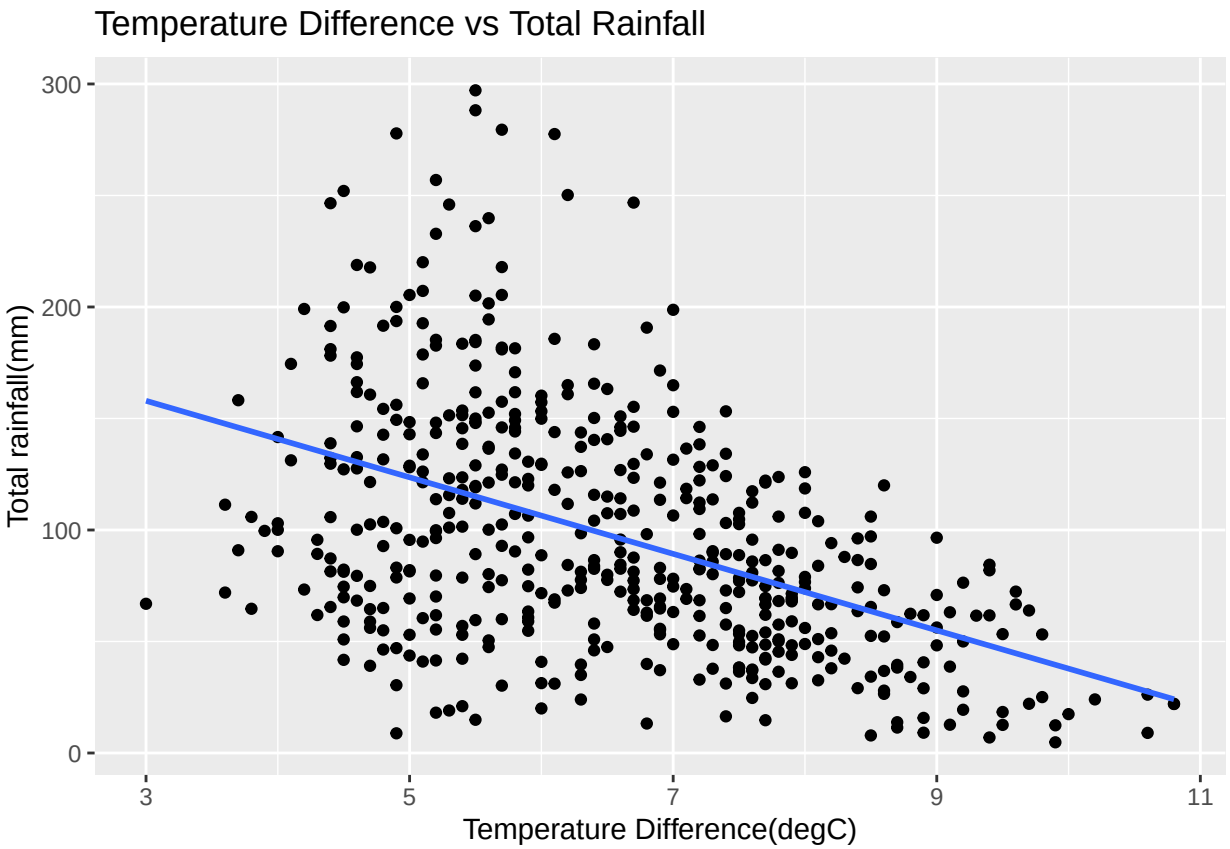
while the regression model for the month in the latter half year is :

$$\widehat{Rainfall} = 185.24 - 3.96 \cdot af - 0.69 \cdot sun$$

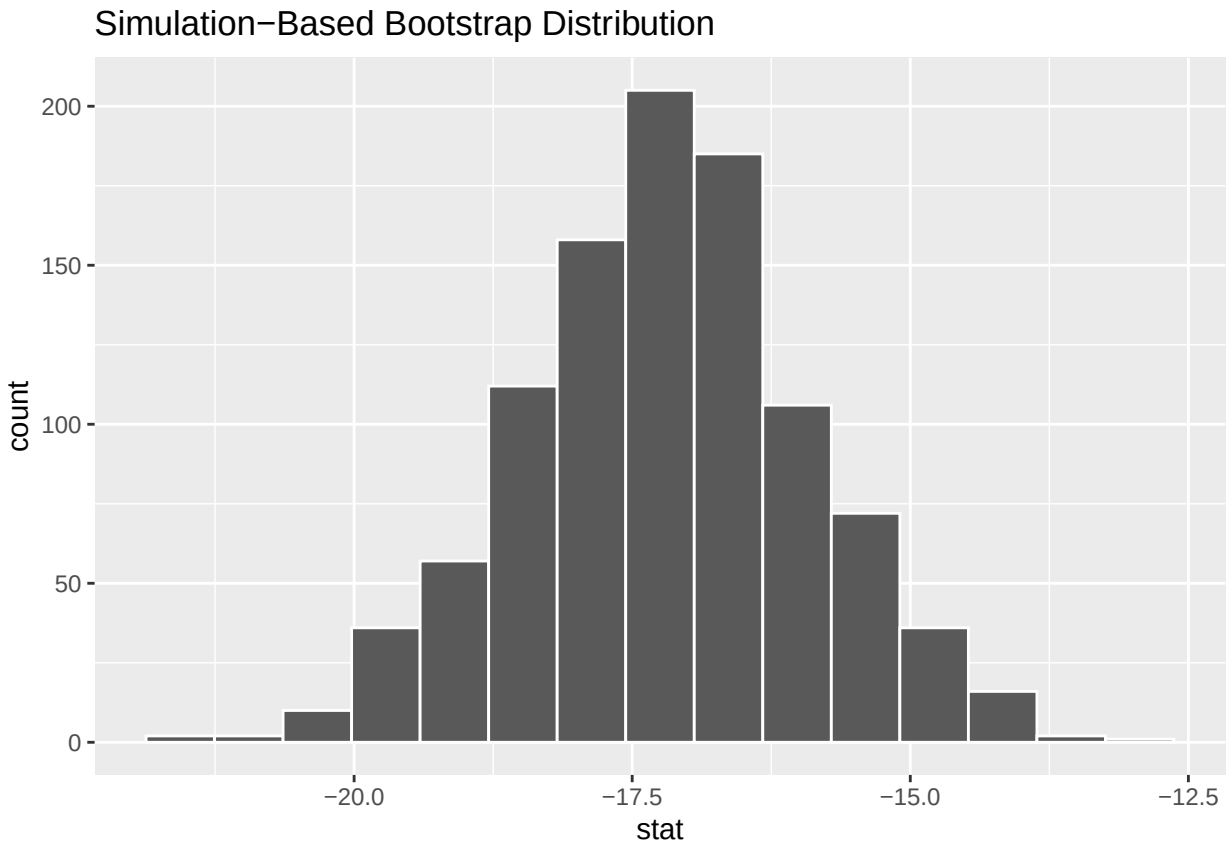
We can use the model analyse and predict the rainfall in Paisley.

Futher Research

As the variable tmax and tmin have very weak linear relationship with Rainfall, we want to find if there is a linear relationship between rainfall and the new variable td(temperature difference) produced by tmax and tmin. We can see the data and fitted model.



We estimate the sampling distribution of the slope parameter $\hat{\beta}$ via the bootstrap method. And here we shall view the bootstrap distribution as a histogram.



Just as the plot shows, the the bootstrap distribution is bell-shaped and accord with what Central Limit Theorem predicted that the sampling distribution would be a normal distribution. Also we can calculate CI for the slope by the bootstrap method.

The 95% CI for the slope parameter in our bootstrap distribution is from -19.72 to -14.62 which doesn't contain zero, hence we could conclude there is a linear relationship and that for every degC the temperature difference the average rainfall decreases between 14.62 and 19.72 units.

We can see an interesting phenomenon that although tmax and tmin has weak relationship with Rainfall, the temperature difference(tmax-tmin) has a statistically significant linear relationship with Rainfall.