Rainfall model

Group_12

2021/6/22

Introduction

summary

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 1: 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

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.

Table 2: 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

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

Model Comparisons

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

Table 3: 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

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 4: 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

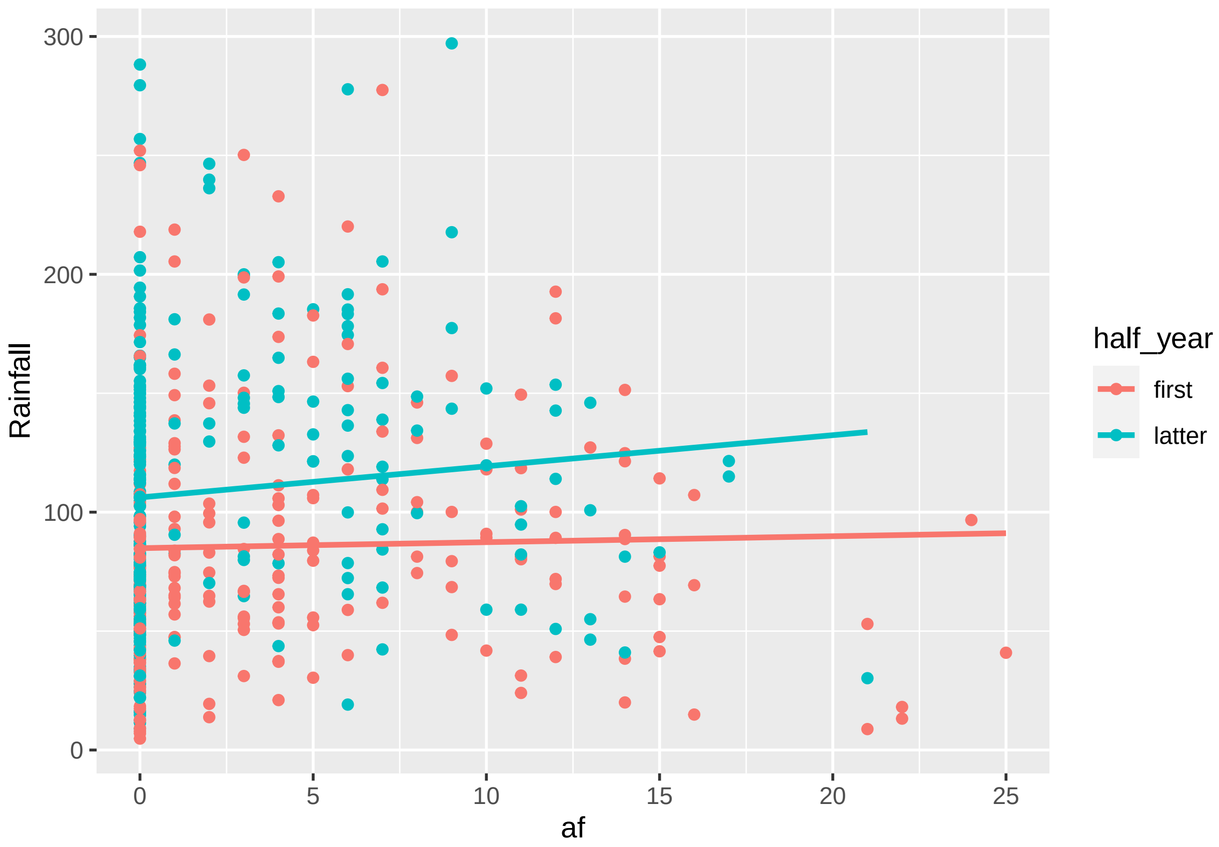
Hence the regression line for the month in the first half year is given by:

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

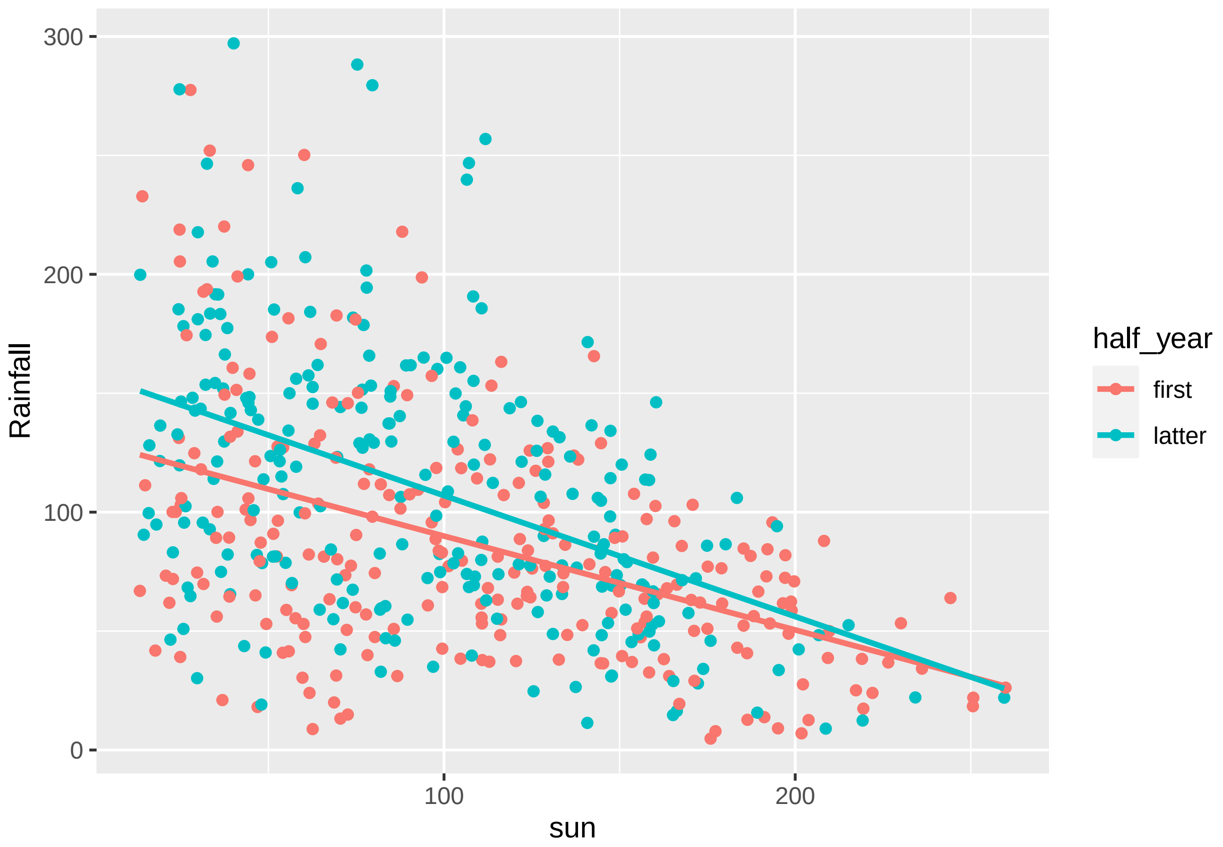
while the regression line for the month in the latter half year is given by:

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

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

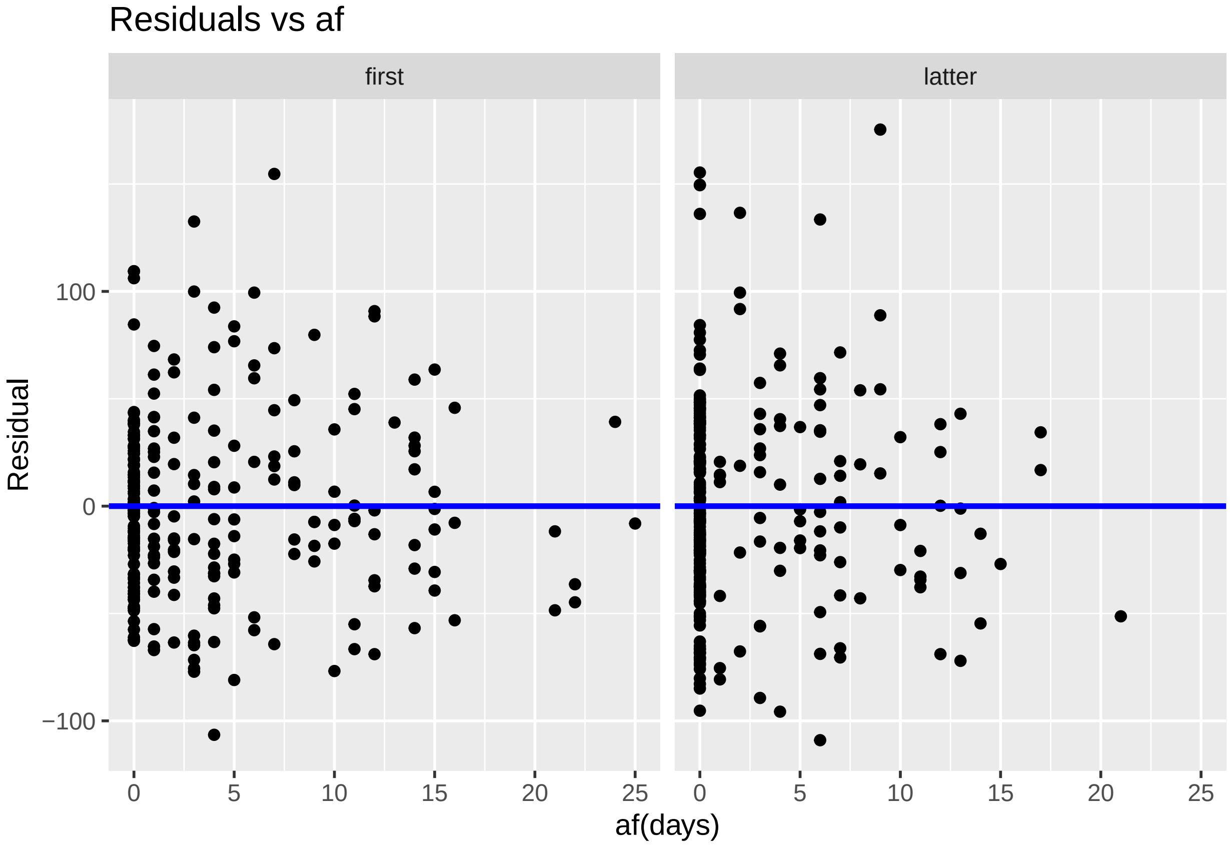


Rainfall
300
200
100
0
0
5
10
15
20
25
af
half_year
first
latter

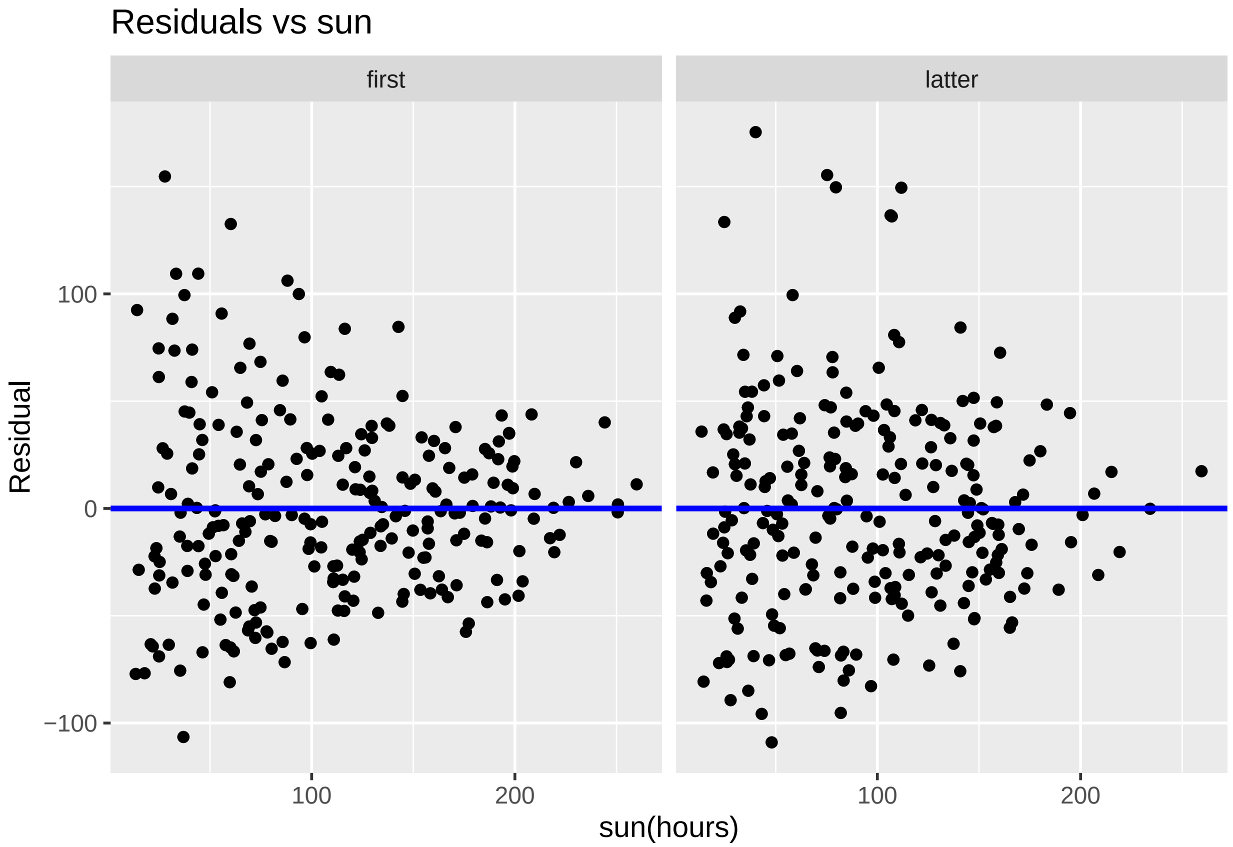

Rainfall
300
200
100
0
100
200
sun
half_year
first
latter

Assessing Model fit

We have to check our model assumptions. First, we need to plot the scatterplot of the residuals against credit

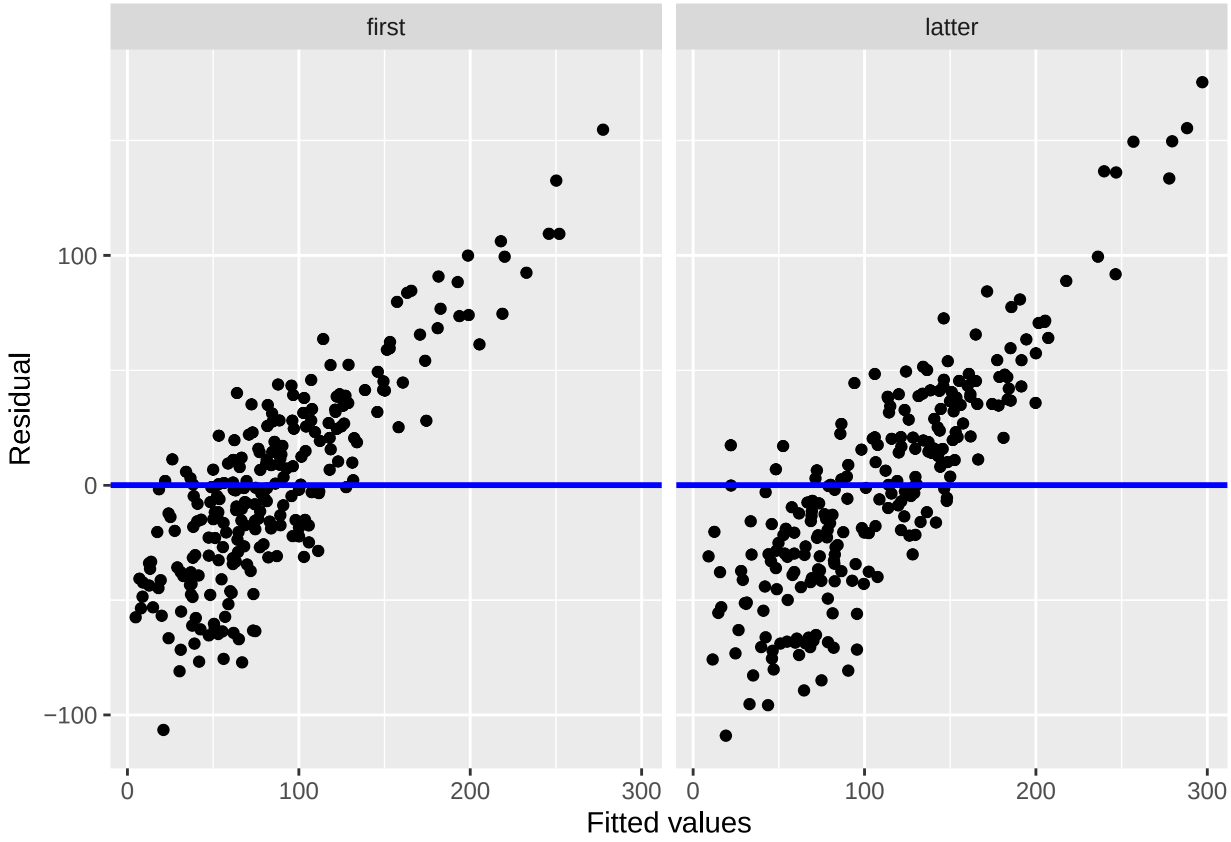


af and sun:

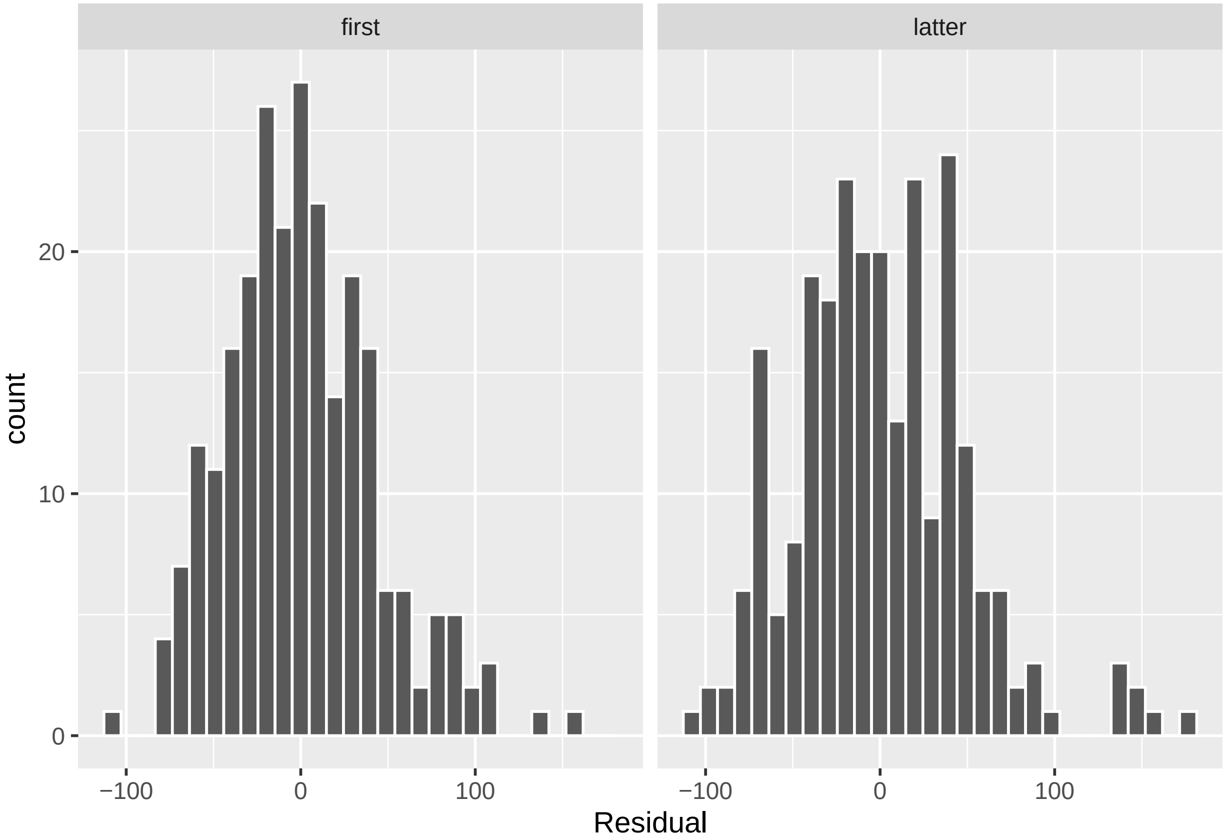


The plot shows that there is an even scatter of points above and below the zero line indicating the residuals have mean zero. And the scattering of the points is also constant across all values of the explanatory variable with no systematic pattern observed in the residuals. So this assumption is valid.

Then we plot the residuals against the fitted values:



In this case, the residuals don’t have mean 0 and there appears to a systematic pattern in the residuals with the scatter of the residuals around the zero line not consistent.The assumptions of the residuals having mean zero and constant variability across all values of the explanatory variable do not appear to be valid.



Our subjective impression is that the residuals appear to be bell-shaped.

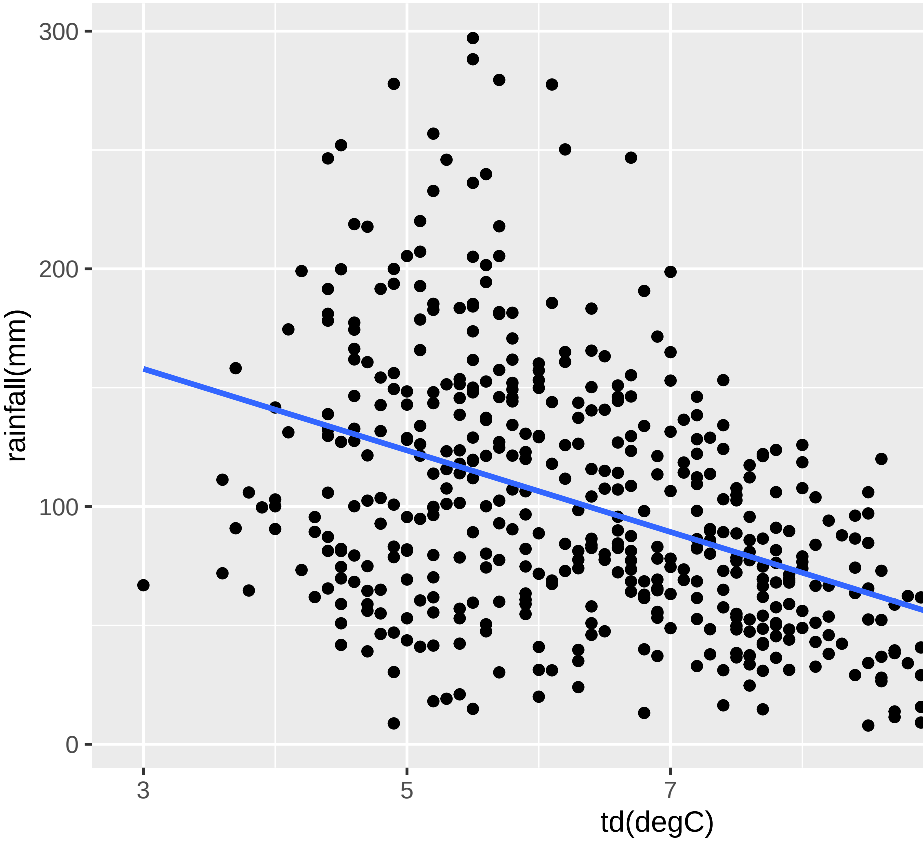
To conclude, the assumptions might not be valid in this model and the model might not be a good fit to the data.

CI for Regression Parameters

CIs have been shown in Table 4.

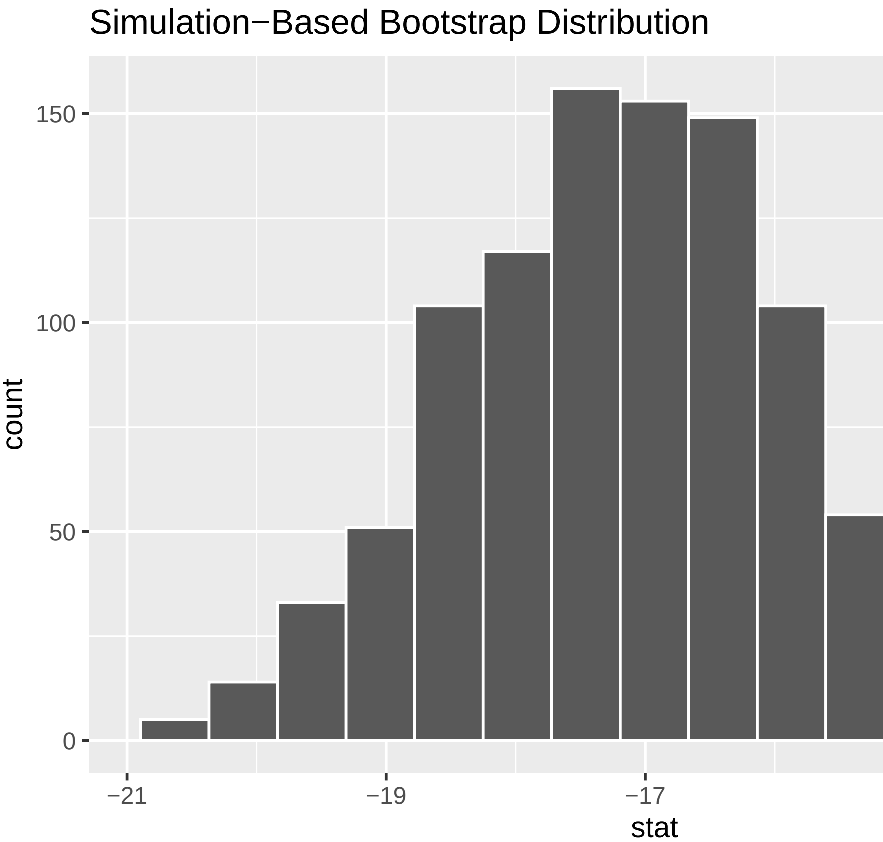
Futher Expansion

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 can



view the bootstrap distribution as a histogram.
CI for the slope:

\begin{table}[!h]

\caption{95% confidence interval}

lower_ci	upper_ci
-19.66	-14.6

\end{table}

The 95% CI for the slope parameter 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.