Simple Linear Regression

Simple linear regression is a linear regression model with a single explanatory variable. That is, it concerns two-dimensional sample points with one independent variable and one dependent variable (conventionally, the x and y coordinates in a Cartesian coordinate system) and finds a linear function (a non-vertical straight line) that, as accurately as possible, predicts the dependent variable values as a function of the independent variable. The adjective *simple* refers to the fact that the outcome variable is related to a single predictor.

$$Y = A + BX + e$$

Where, A = Population Intercept

B =Population Regression coefficient

e =error

We can use simple linear regression to know:

1. How strong the relationship is between two variable
2. The value of dependent variable at a certain value of the independent variable.

Problem Statement

Meteoblue weather presents hourly weather data with datasets for various meteoroid indicators, water resource planning, rainfall, and others from across various parts of India. It also contains databases for several other parameters such as temperature, pressure, relative humidity, precipitation amount, wind speed, solar radiation, among others.

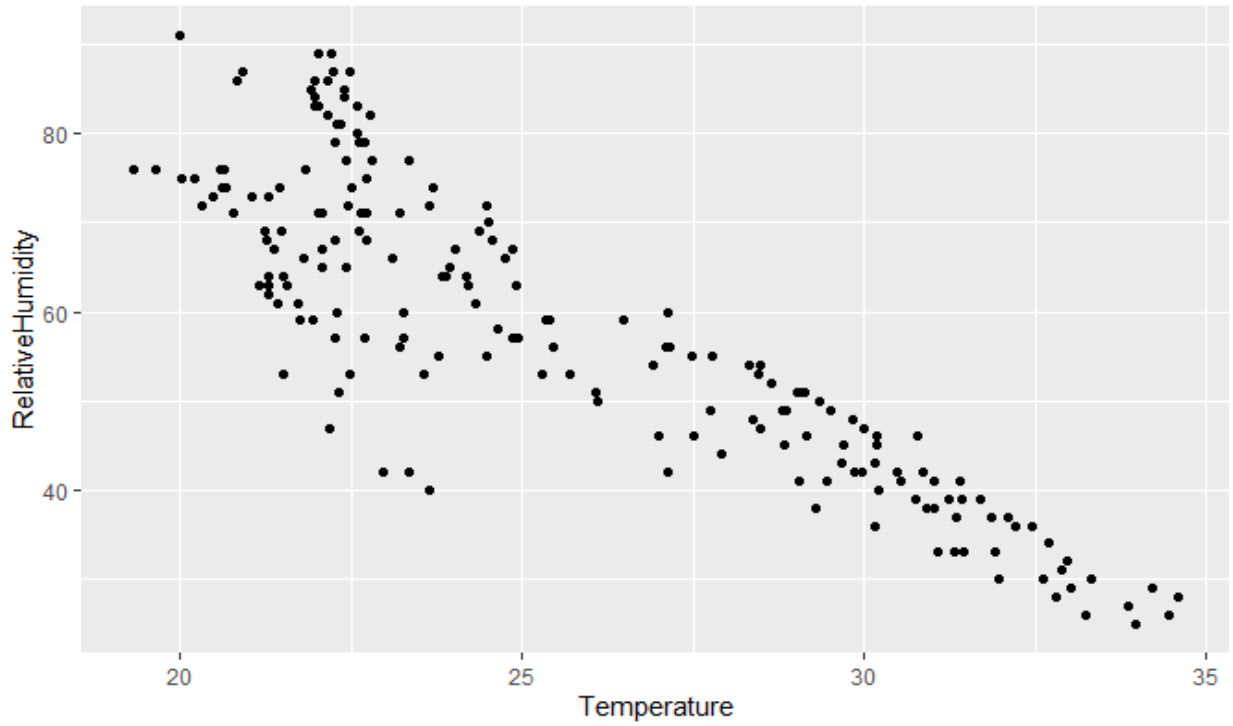
We are considering the report from March 13, 2021 to March 20, 2021 where X is temperature (in ℃)(independent variable) and Y is Relative Humidity (in %)(dependent variable) at the level of 2m. Objective is to examine the relationship between relative humidity(in %) and temperature (in ℃). The Regression model for our problem statement is Relative Humidity= A+(B* Temperature)+e.

There are 192 observations in the data and it was decided to take a 5% level of significance for the Hypothesis testing and Interval estimations.

Also,

Predict the relative humidity given that, temperature is 24.25 ℃.

Graphical Representation



Estimation of Parameters(Intercept and slope)

Method : Least Square Estimation

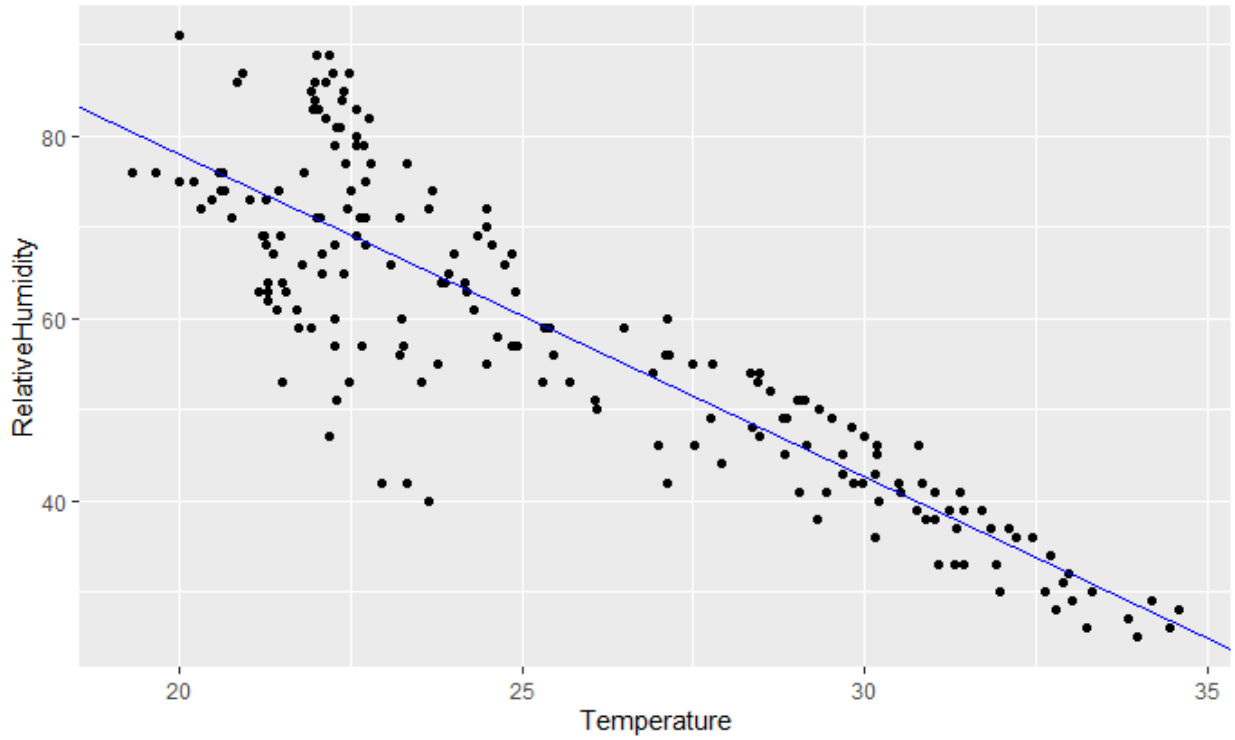
The parameters of the model a(intercept) and b(slope) are obtained by finding values of a and b that minimizes the sum of squared differences between Y and Y_{est}.i.e $\sum(Y - Y_{est})^2$

$$a = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

$$b = \frac{n\sum(XY) - (\sum X)(\sum Y)}{n(\sum X^2)}$$

	Coefficients
Intercept	148.7165915
Temperature(X)	-3.53714178

Graphical Representation



Our regression equation is,

$$\widehat{Relative\ Humidity} = 148.716 - 3.537 * Temperature$$

Interpretation of the slope and intercept

$a$is estimated average value of Y,when the value of $X$is zero
b is estimated change in average value of Y as a result of one-unit change in X

Our model here is $\widehat{Y} = a + bX$

For the Relative humidity within the ranges of temperature observed, 148.716 is the portion of relative humidity not explained by the variable temperature

For one degree increase in the temperature, the relative humidity will be decreased by 3.537%

Measures of Variation

Total variation is made up of two parts

$$Total\ variation\ (SST) = Explained\ variation\ (SSR) + Unexplained\ variation\ (SSE)$$

$$\sum (Y - \overline{Y})^2 = \sum (Y_{est} - \overline{Y})^2 + \sum (Y - Y_{est})^2$$

In our model,

	SS
SSR	41085.63896
SSE	12583.84021
SST	53669.47917

Coefficient of Determination(r^2)

The r^2 is a portion of the total variation in the dependent variable which is explained by the variation in the independent variable. It is also called as r-squared.

$$r^2 = \frac{Explained\ variation}{Total\ variation}$$

In our model,

Regression Statistics	*Value*
Multiple R	0.8749461659
R Square	0.7655307932
Adjusted R Square	0.7642967447
Standard Error	8.138226954
Observations	192

76.55% of the variation in relative humidity is explained by the variation in temperature
As the value of slope is negative, it means *r*(Correlation Coefficient)= -0.8749, so there exists a strong negative relationship between Relative humidity and Temperature

F-Test for significance(ANOVA)

Null and alternate hypothesis

$H_0 : B = 0$(Slope is not significant)
$H_1 : B \neq 0$ (Slope is significant)

Test Statistic

$$F_{cal} = \frac{MSR}{MSE}$$

where,

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n-k-1}$$

where, F follows an $F - Distribution$ with k numerator and $n - k - 1$ denominator degree of freedom
n =no of observations
k = no of independent variables

Result

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	41085.63896	41085.63896	620.3409509	9.49E-62
Residual	190	12583.84021	66.23073795		
Total	191	53669.47917			

Decision Criteria

Rej H_0 if $F_{cal} > F_{crit}$ or $p - value < \alpha$
F_{crit}= F(0.05,1,192-2)=3.89
As $F_{cal} > F_{crit}$, We Rej H_0 at 0.05 level of significance

Conclusion
There is enough evidence that the temperature affects relative humidity

t-test for intercept and its Confidence Interval

t −Test for $y - intercept$ and its Confidence interval

Null and alternative hypothesis:

$H_0 : A = 0$(Intercept is not significant)
$H_1 : A \neq 0$(Intercept is significant)

$$t_{cal} = \frac{a-A}{s.e(a)}$$

where,

$$s.e(a) = \frac{S}{\sum_i^n (X_i - \bar{X})} * \sqrt{\frac{\sum X}{n}}$$

$S = \sqrt{MSE}$

Decision Criteria
Rej H_0 if $t_{cal} > \left| t_{\alpha/2,n-2} \right|$ at α level of significance.

Confidence Interval for y-intercept - $a \pm t_{\alpha/2,n-2} * se(a)$

Results

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	148.7165915	3.687250317	40.33265407	4.31E-95	141.4433865	155.9897965

As, 40.332>1.9725 we reject H_0at 5% level of significance
i.e; There is enough evidence to show that y-intercept is significant in our model
And, with 95% Confidence we can say that the true value of intercept will lie between 141.4433 to 155.9897

t-test for Slope and its Confidence interval

Null and alternative hypothesis
$H_0 : B = 0$(Slope is not significant)
$H_1 : B \neq 0$ (Slope is significant)

$$t_{cal} = \frac{b-B}{s.e(b)}$$

where,

$$s.e(b) = \frac{S}{\sum_i^n (X_i - \bar{X})}$$

Decision Criteria

Rej H_0 if $t_{cal} > \left| t_{\alpha/2,n-2} \right|$ at α level of significance

Confidence Interval for Slope- $b \pm t_{\alpha/2,n-2} * se(b)$

Result

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Temperature(X)	-3.53714178	0.1420159889	-24.90664471	9.49E-62	-3.817272314	-3.257011246

Conclusion
As, -24.9066 < -1.9725 we reject H_0 at 5% level of significance
I.e There is enough evidence that temperature affects humidity
And, with 95% Confidence we can say that the true value of slope will lie between -3.817 to -3.257

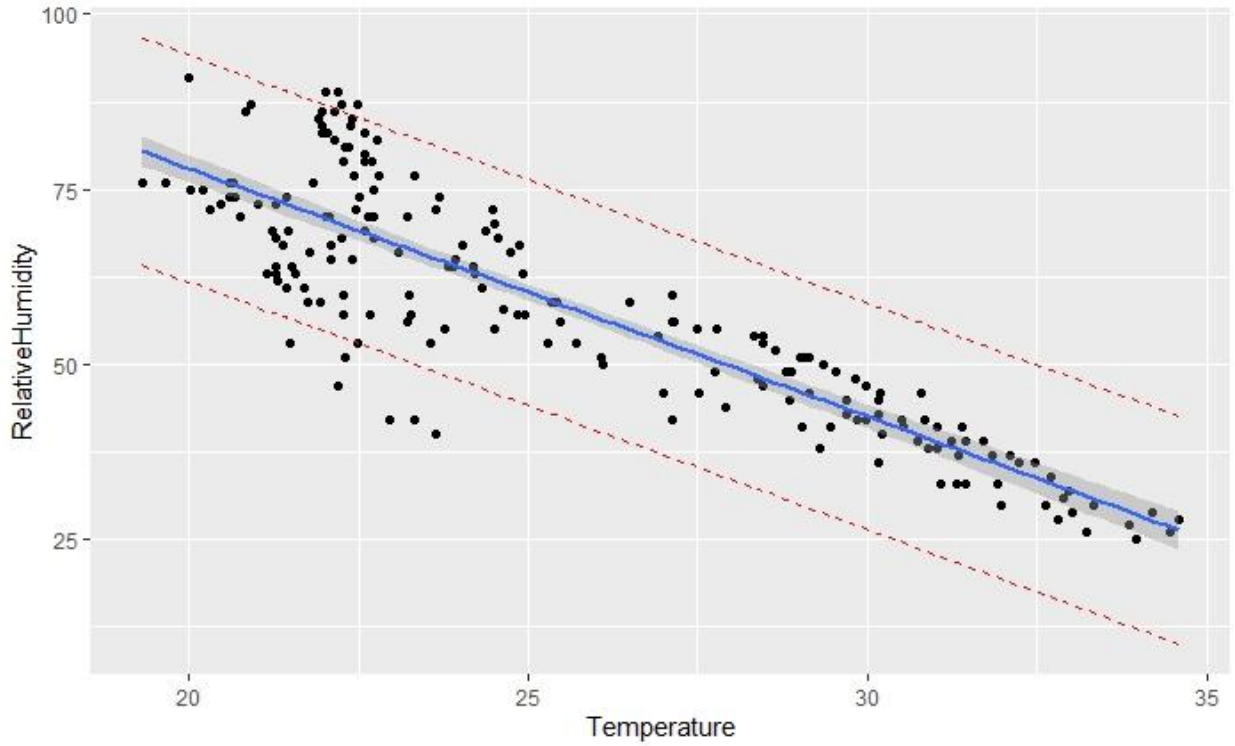
Confidence Interval and Prediction interval for Y_{est}

Confidence Interval

$$C.I = Y_{est} \pm t_{\frac{\propto}{2},n-2}\, S\sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum(X_i - \overline{X})^2}}$$

Prediction Interval

$$P.I = Y_{est} \pm t_{\frac{\propto}{2},n-2}\, S\sqrt{\frac{1}{n} + 1 + \frac{(X_i - \overline{X})^2}{\sum(X_i - \overline{X})^2}}$$

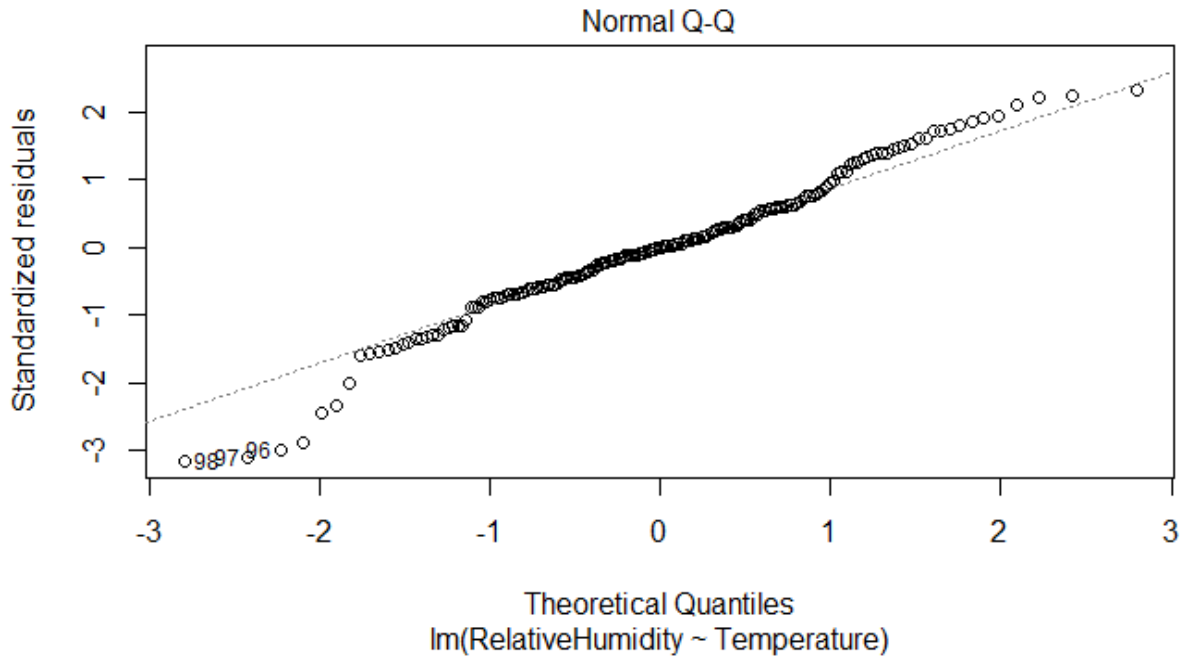


The 95% Confidence intervals and the Prediction intervals for the Y-estimates are given in the sheet

Adequacy of Regression Model

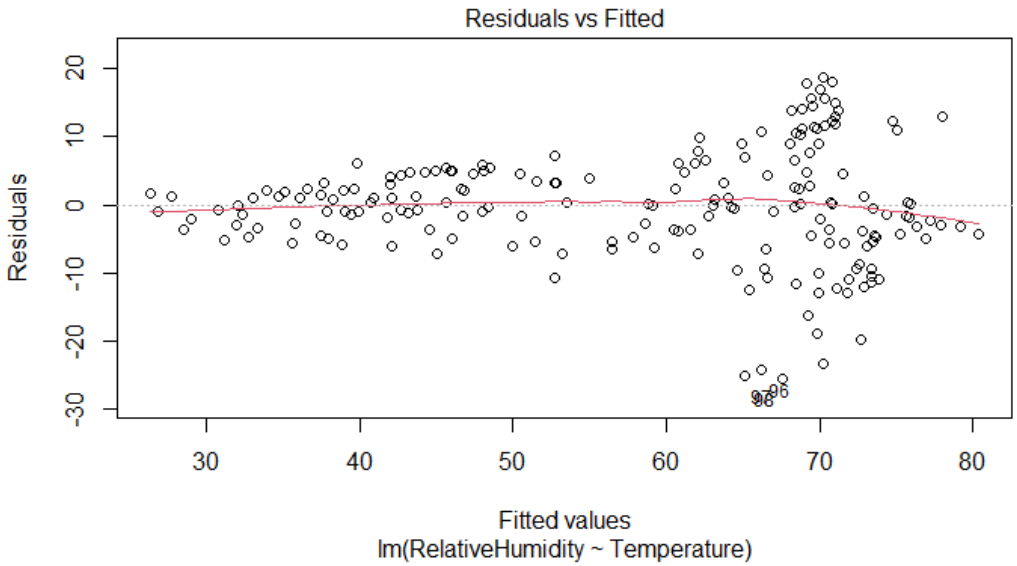
1) **Normal Q-Q (Normality of Errors)**

The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line. In our example, most of the points fall approximately along this reference line, so we can assume normality. That means that our data is normally distributed.



2) **Residuals vs Fitted. (Linearity)**

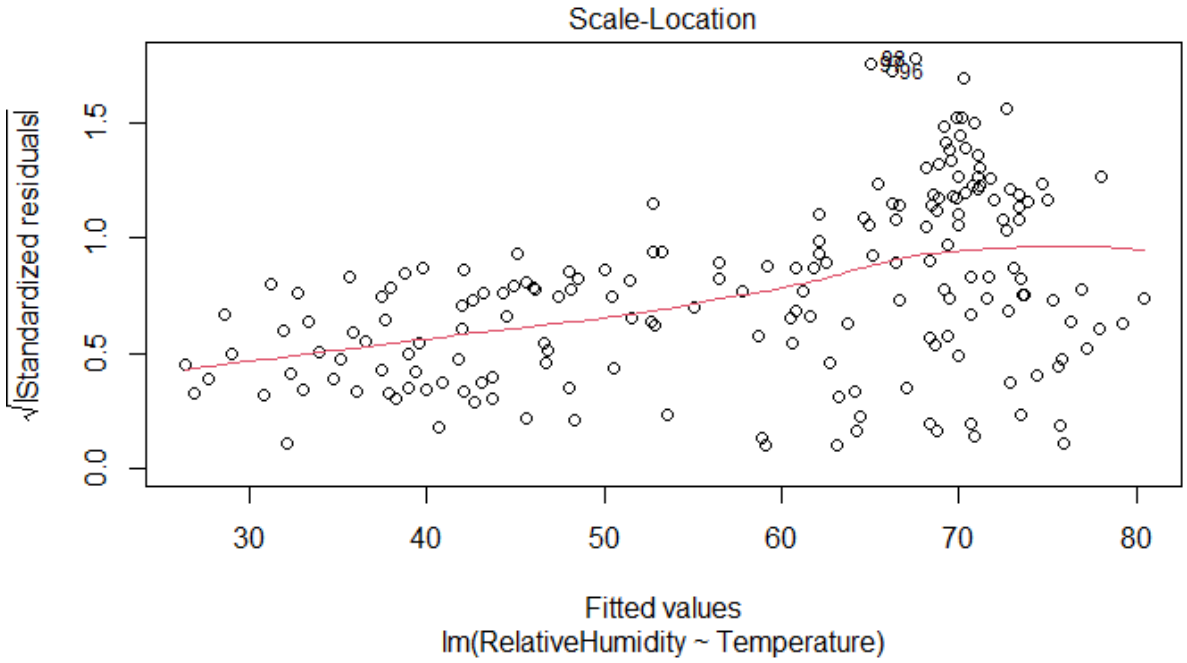
Used to check the linear relationship assumptions. A horizontal line, without distinct patterns is an indication for a linear relationship, what is good.



The residual plot showed a pattern. The presence of a pattern may indicate a problem with some aspect of the linear model.

3) **Scale-Location. (No heteroscedasticity)**

Used to check the homogeneity of variance of the residuals (homoscedasticity). Horizontal line with equally spread points is a good indication of homoscedasticity.



This plot shows if residuals are spread equally along the ranges of predictors. It's good if you see a horizontal line with equally spread points.

In this case, it is clear that the variability (variances) of the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residuals errors (or *heteroscedasticity*).

To predict the relative humidity given temperature(24.25 ℃)

$$\widehat{Relative\ Humidity} = 148.716 - 3.537 * 24.25$$
$$\widehat{Relative\ Humidity} = 62.94$$

The estimated relative humidity when the temperature is 24.25℃ is 62.94.
Also, with 95% confidence(C.I) we can say that the mean of relative humidity with 24.25 ℃ would lie between 61.78 and 64.09
And, with 95% confidence(P.I) we can say that the true value of relative humidity with 24.25 ℃ would lie between 46.84 and 79.03

Conclusion

There exist a significant negative relationship between relative humidity and temperature
For a one degree increase in temperature within the observed range values, the relative humidity is supposed to be decreased by 3.537 percent