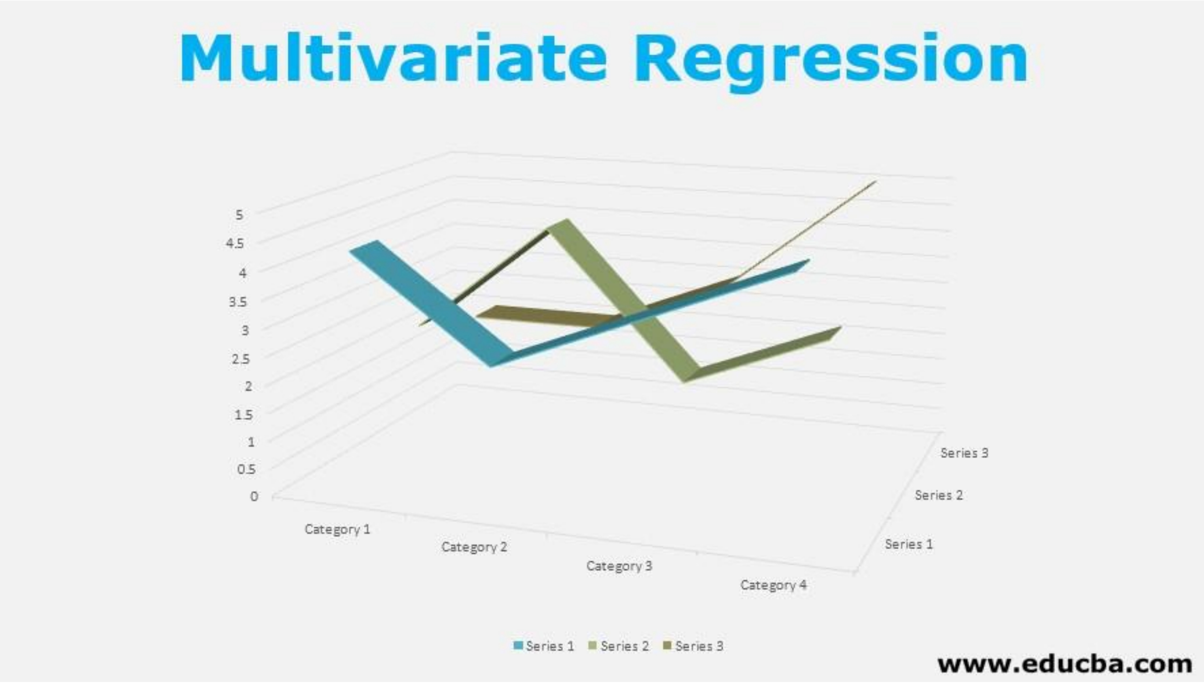
# Applied Multivariate Analysis: Temperature Prediction Using Multiple Regression

A project harnessing advanced multivariate techniques for precise temperature forecasting.




By: Uzma Naeem

# REPORT

## 1. DATA DESCRIPTION:

The data was gathered for the time period of January 2nd,2023 till 31st January 2023. The API gave the following continuous (quantitative) data variables recorded on an hourly basis for the above-mentioned time period:

1. **Y** Variable: Temperature_2m above ground (°C)
2. $1^{st}$ X variable: Relativehumidity_2m above ground (%)
3. $2^{nd}$ X variable: Dewpoint_2m above ground (°C)
4. $3^{rd}$ X variable: Apparent_temperature (°C)
5. $4^{th}$ X variable: Pressure_msl (hPa)
6. $5^{th}$ X variable: Cloud cover (%)
7. $6^{th}$ X variable: Windspeed_100m (km/h),

To create ONE discrete **(qualitative) variable** in the data, the following procedure was followed:

**a)** Using the date time variable, hours of the day between 6 PM and 6 AM were labeled as 'night' time. Other hours were labeled as 'day'.

**b)** Data were grouped and averaged based on day, month, and night/daytime thereby giving us 2 recordings of average temperature (one for the night and one for the day) for each day between 2nd January 2023 and 31st January 2023.

## 2. SOURCE:

The data was taken from: https://open-meteo.com/en/docs/historical-weather-api

Location details that were used with the API on the website were: latitude, longitude, elevation, utc_offset_seconds, timezone, timezone_abbreviation
33.700005,73.0,585.0,18000, Asia/Karachi, PKT

## 3. INITIAL CHOICE OF VARIABLES:

The initial selection of these specific X variables to forecast the Y variable (temperature_2m above ground) can be supported by their established correlations with temperature and their importance in capturing various components that influence temperature changes. Here's an explanation for why each **X** variable was chosen:

### a) 1st X Variable - Relative humidity:

It's well-known that humidity influences temperature. The air can feel warmer or cooler depending on the amount of humidity present. Because of this, it makes sense to think about relative humidity as a potential temperature prediction.

### b) 2nd X Variable - Dewpoint:

The temperature at which condensation starts to happen is known as the dew point. It is associated with humidity and can give information about how much moisture is present in the air. The temperature may be affected by greater dew points since they are linked to higher moisture content. Dew point, therefore, has a role in temperature prediction.

c) **3rd X Variable - Apparent Temperature:**

A person's perception of the temperature is affected by several variables, including temperature, humidity, wind speed, and radiation. These variables are all considered while calculating apparent temperature. Instead of just looking at the temperature as it is measured, we also take into account how people perceive the temperature by including apparent temperature as a predictor.

d) **4th X Variable - Pressure:**

A change in atmospheric pressure may signal a change in weather and may have an impact on temperature. In contrast, low-pressure systems can bring clouds and lower temperatures. For instance, high-pressure systems are frequently associated with clear skies and warmer temperatures. As a result, adding pressure as a predictor helps account for how the atmosphere affects temperature.

e) **5th X Variable - Cloud cover:**

The percentage of the sky that is covered with clouds is referred to as cloud cover. By either reflecting or absorbing sunlight, clouds can change the amount of incoming solar radiation that reaches the surface, which can have an impact on temperature. Lower temperatures can be attained by increasing cloud cover, whereas higher temperatures can be achieved by decreasing cloud cover. So, cloud cover matters when predicting temperature.

f) **6th X Variable - Windspeed:**

By enabling heat transfer between the atmosphere and the surface, wind speed can affect temperature. Lower wind speeds can enable the surface to retain more heat, producing a warmer feeling, while higher wind speeds can improve convective heat transfer, resulting in a cooler feeling. Hence, wind speed is a valuable predictor of temperature.

## Category Variable - Time of Day:

The temperature is affected by the time of day due to differences in solar radiation. As a result of the sun's direct heat, days are often warmer than nights. On the other side, temperatures are often lower at night because heat is lost to space. The model can capture these diurnal temperature patterns by including the time of day as a category variable.

## Y Variable: Temperature

The Y variable, temperature_2m above ground (°C), refers to the measured temperature at a height of 2 meters above the Earth's surface. It represents the air temperature in a certain area, which is a crucial sign of the local thermal conditions. The measurement of temperature is normally done in degrees Celsius (°C), and it affects a variety of processes in the ecosystem as well as characteristics of the weather and climate. Insights into temperature variations and assistance in comprehending the underlying causes of temperature changes can both be gained from the prediction of this variable.

## 4. EQUATION WITH EXPLANATION:

Following are the two equations for day and night (Day=0, Night=1).

**a) DAY:**

$$\hat{y} = -184.226 + 0.197X(p) - 0.116X(rh) + 0.546X(dp)$$

**b) NIGHT:**

$$\hat{y} = -178.419 + 0.197X(p) - 0.116X(rh) + 0.546X(dp)$$

where X(p) = X of Pressure

X(rh) = X of Relative humidity

X(dp) = X of Dewpoint

In the given equations, ŷ represents the predicted value of the temperature variable (y) during the day and night, respectively. The equations are multiple regression models that relate the temperature (y) to three predictor variables: Pressure (X(p)), Relative Humidity (X(rh)), and Dewpoint (X(dp)).

The coefficients in front of each predictor variable (0.197, -0.116, 0.546) represent the estimated effects of that predictor on the temperature.

In equation

**a) for the daytime,**

The predicted temperature (ŷ) is calculated by taking the sum of the intercept (-184.226) and the product of each predictor variable with its corresponding coefficient. The equation suggests that an increase in Pressure (X(p)) by one unit leads to an increase in the predicted temperature by 0.197 units, a decrease in Relative Humidity (X(rh)) by one unit leads to a decrease in the predicted temperature by 0.116 units, and an increase in Dewpoint (X(dp)) by one unit leads to an increase in the predicted temperature by 0.546 units.

Similarly, equation

**b) for the nighttime,**

It follows the same structure but with different coefficient values and intercepts (-178.419). The coefficients indicate the expected change in temperature associated with a one-unit change in each respective predictor variable during the nighttime.

These equations allow us to estimate the temperature based on the given predictor variables. By plugging in the values of X(p), X(rh), and X(dp) into the equations, the predicted temperature values can be obtained for the daytime and nighttime, respectively.

(Detailed Calculation - Appendix-4C)

## 5. VARIABLES ENDED UP IN THE FINAL EQUATION:

Based on the final regression analysis, these variables are present in the final equation, which shows that they have a considerable impact on explaining the variation in the Y variable (temperature_2m above ground):

1. Time of Day (day=0, night=1)
2. Pressure MSL (hPa)
3. Relative Humidity 2m (%)
4. Dewpoint 2m (°C)
5. Time of Day * Relative Humidity
6. Time of Day * Dew Point

Considering these variables and their interactions can help improve our understanding and prediction of temperature variations. It focuses on the significance of diurnal cycles, atmospheric pressure, moisture content, and their dynamic interactions with temperature. Further research into these relationships might offer insightful information about local climate dynamics and assist in making decisions about weather forecasts.

## 6. RELATIONSHIP OF FINAL X VARIABLES WITH Y VARIABLE:

These results suggest that the X variables; dew point, humidity, atmospheric pressure, and time of day all have significant effects on the Y Variable, temperature at 2 meters above the ground. It is implied by the inclusion of interaction terms that the relationships between the time of day and relative humidity and dew point are not constant but instead change depending on the time period.

### a) Time of Day:

This variable denotes the time of day, which can be classified as daytime (0) or nighttime (1). It captures the pattern of temperature variations throughout the day. Due to the direct heating influence of sunlight, daytime temperatures are often greater, but nighttime temperatures are typically lower because of heat loss to space.

### b) Pressure MSL:

A predictor for pressure at mean sea level (MSL) has been added because variations in atmospheric pressure might affect temperature. Warmer temperatures are frequently brought on by high-pressure systems, whereas cooler temperatures are frequently brought on by low-pressure systems. The result reveals how changes in atmospheric conditions affect temperature fluctuations by including pressure MSL.

### c) Relative Humidity:

The quantity of moisture in the air that is actually present as opposed to the maximum amount that it might contain at a particular temperature is known as relative humidity. The air may feel warmer or cooler depending on its relative humidity; higher relative humidity can make the air seem warmer. The outcome of the data work shows how differences in moisture content affect temperature by including relative humidity at a height of 2 meters.

### d) Dewpoint:

The model's inclusion of the dew point at 2 meters above ground level enables a thorough analysis of moisture's influence on temperature variations. The dew point, or the temperature at which air gets saturated and condensation occurs, is a useful measure of the amount of moisture in the air. Higher moisture content is associated with higher dew points, which can affect temperature. The model's ability to accurately anticipate temperature changes is improved by using dew point as a predictor. This effectively represents the complex link between moisture and temperature.

### e) Time of Day * Relative Humidity:

The combined impact of these factors on temperature is accounted for by the interaction term between the time of day and relative humidity. It accurately depicts how temperature and humidity change throughout the day. For instance, due to variations in solar radiation and moisture evaporation rates, the effect of relative humidity on temperature may change from day to night.

### f) Time of Day * Dew Point:

The interaction between time of day and dew point considers the combined impact of these variables on temperature, much like the previous interaction term. It illustrates how, considering elements like solar radiation and moisture content, the connection between temperature and dew point changes during the day.

## 7. DATA RE-EXPRESSIONS/ TRANSFORMATION:

The analysis of half slopes indicates that the three statistically significant X variables do not necessitate transformation. This conclusion is based on the observed differences in the slope coefficients' signs for the first and second variables. The third variable deviates from the 0.5-2 range. Therefore, considering the characteristics of the slope coefficients and their magnitudes, it is concluded that transforming these variables is not necessary for the present analysis. **(Appendix-4B)**

# APPENDIX

## 1. Tables of Sums of Squares:

### 1st Regression

| | | | | | |
|---|---|---|---|---|---|
| **1st Regression**<br>(with all 6 X Variables & 1 indicative Variable) | | | | | |
| SUMMARY OUTPUT | | | | | |
| *Regression Statistics* | | | | | |
| Multiple R | 0.94127251 | | | | |
| R Square | 0.88599394 | | | | |
| Adjusted R Square | 0.87064697 | | | | |
| Standard Error | 1.1211638 | | | | |
| Observations | 60 | | | | |
| **ANOVA** | | | | | |
| | ***df*** | ***SS*** | *MS* | *F* | *Significance F* |
| **Regression** | 7 | 507.9773 | 72.56818 | 57.73087 | 2.7552E-22 |
| **Residual** | 52 | 65.36443 | 1.257008 | | |
| **Total** | 59 | 573.3417 | | | |

### 2nd Regression

| | | | | | |
|---|---|---|---|---|---|
| **2nd Regression**<br>with 5 X Variables & 1 indicative Variable<br>(excluding Apparent Temperature) | | | | | |
| SUMMARY OUTPUT | | | | | |
| *Regression Statistics* | | | | | |
| Multiple R | 0.94121131 | | | | |
| R Square | 0.88587873 | | | | |
| Adjusted R Square | 0.87295935 | | | | |
| Standard Error | 1.11109738 | | | | |
| Observations | 60 | | | | |
| **ANOVA** | | | | | |
| | ***df*** | ***SS*** | *MS* | *F* | *Significance F* |
| **Regression** | 6 | 507.9112 | 84.65187 | 68.56971 | 3.2522E-23 |
| **Residual** | 53 | 65.43048 | 1.234537 | | |
| **Total** | 59 | 573.3417 | | | |

### 3rd Regression

| | | | | | |
|---|---|---|---|---|---|
| **3rd Regression**<br>with 4 X Variables & 1 indicative Variable<br>(excluding Apparent Temperature & Windspeed) | | | | | |
| SUMMARY OUTPUT | | | | | |
| *Regression Statistics* | | | | | |
| Multiple R | 0.93843106 | | | | |
| R Square | 0.88065286 | | | | |
| Adjusted R Square | 0.8696022 | | | | |
| Standard Error | 1.12568244 | | | | |
| Observations | 60 | | | | |
| **ANOVA** | | | | | |
| | ***df*** | ***SS*** | *MS* | *F* | *Significance F* |

| | | | | | |
|---|---|---|---|---|---|
| **Regression** | 5 | 504.915 | 100.983 | 79.69233 | 1.1146E-23 |
| **Residual** | 54 | 68.42669 | 1.267161 | | |
| **Total** | 59 | 573.3417 | | | |

## 4th Regression

| **4th Regression**<br>with 3 X Variables & 1 indicative Variable<br>(excluding Apparent Temperature, Windspeed & Cloudcover) | | | | | |
|---|---|---|---|---|---|
| SUMMARY OUTPUT | | | | | |
| *Regression Statistics* | | | | | |
| Multiple R | 0.93638728 | | | | |
| R Square | 0.87682113 | | | | |
| Adjusted R Square | 0.86786267 | | | | |
| Standard Error | 1.13316597 | | | | |
| Observations | 60 | | | | |
| **ANOVA** | | | | | |
| | ***df*** | ***SS*** | *MS* | *F* | *Significance F* |
| **Regression** | 4 | 502.7181 | 125.6795 | 97.87629 | 2.4527E-24 |
| **Residual** | 55 | 70.62358 | 1.284065 | | |
| **Total** | 59 | 573.3417 | | | |

## 5th Regression

| **5th Regression**<br>with **3** X Variables, **1** indicative Variable & **3** indicative variables **times** X variable.<br>(excluding Apparent Temperature, Windspeed & Cloudcover) | | | | | |
|---|---|---|---|---|---|
| SUMMARY OUTPUT | | | | | |
| *Regression Statistics* | | | | | |
| Multiple R | 0.94365957 | | | | |
| R Square | 0.89049339 | | | | |
| Adjusted R Square | 0.87575211 | | | | |
| Standard Error | 1.09881675 | | | | |
| Observations | 60 | | | | |
| **ANOVA** | | | | | |
| | ***df*** | ***SS*** | *MS* | *F* | *Significance F* |
| **Regression** | 7 | 510.557 | 72.93671 | 60.40817 | 9.7892E-23 |
| **Residual** | 52 | 62.78471 | 1.207398 | | |
| **Total** | 59 | 573.3417 | | | |

## 6th Regression

| **6th Regression**<br>with **3** X Variables, **1** indicative Variable & **2** indicative variables **times** X variable.<br>(excluding Apparent Temperature, Windspeed & Cloudcover) | | | | | |
|---|---|---|---|---|---|
| SUMMARY OUTPUT | | | | | |
| *Regression Statistics* | | | | | |
| Multiple R | 0.94253887 | | | | |
| R Square | 0.88837953 | | | | |
| Adjusted R Square | 0.87574325 | | | | |
| Standard Error | 1.09885595 | | | | |
| Observations | 60 | | | | |

| ANOVA | | | | | |
|---|---|---|---|---|---|
| | *df* | *SS* | *MS* | *F* | *Significance F* |
| **Regression** | 6 | 509.345 | 84.89084 | 70.30388 | 1.8177E-23 |
| **Residual** | 53 | 63.99667 | 1.207484 | | |
| **Total** | 59 | 573.3417 | | | |

## 2. Tests of Hypotheses (Ho, Ha), Calculated F statistics, F table values:

1. *Testing Ho β (apparent temperature) = 0 given β (time of day) β (pressure) β (windspeed) β (relative humidity) β (dewpoint) β (cloud cover)*

   ***VS***

   *Testing Ho β (apparent temperature) <> 0 given β (time of day) β (pressure) β (windspeed) β (relative humidity) β (dewpoint) β (cloud cover).*

**F*** = [ SSReg(X(at),X(tod),X(p),X(ws),X(rh),X(dp),X(cc)) – SSReg(X(tod), X(p),X(ws),X(rh),X(dp),X(cc)] / df (X(at),X(tod), X(p),X(ws),X(rh),X(dp),X(cc) – dfX(tod), X(p),X(ws),X(rh),X(dp),X(cc)]

---

[SSCT-SSReg[(X(at),X(tod), X(p),X(ws),X(rh),X(dp),X(cc)]/(n-1 - #Xvariables)

**F*** = 507.977-507.911/ (7-6)

573.341-507.977/ (60-1-7)

**F*** = 0.052545994

**F(table)** = 4.0266314

Since **F*< F(table)**, it indicates that the observed F-statistic is not statistically significant at the chosen significance level. In other words, there is insufficient evidence to reject the null hypothesis. This means that the apparent temperature **(X(at))** variable is not statistically significant in explaining the variation in the dependent variable (y) when considering the other independent variables (X(tod), X(p), X(ws), X(rh), X(dp), X(cc)).

2. *Testing Ho β (windspeed) = 0 given β (time of day) β (pressure) β (relative humidity) β (dewpoint) β (cloud cover)*

   ***VS***

   *Testing Ho β (windspeed) <> 0 given β (time of day) β (pressure) β (relative humidity) β (dewpoint) β (cloud cover)*

**F*** = [ SSReg X(ws),X(tod),X(p),X(rh),X(dp),X(cc)) – SSReg(X(tod), X(p),X(rh),X(dp),X(cc)] / df X(tod), X(p),X(ws),X(rh),X(dp),X(cc) – df X(tod), X(p),X(rh),X(dp),X(cc)]

---

[SSCT-SSReg[(X(ws),X(tod),X(p),X(rh),X(dp),X(cc)]/(n-1 - #Xvariables)

**F*** = 507.911- 504.915/ (6-5)

573.341-507.911/ (60-1-6)

**F*** = 2.426990567

**F(table)** = 4.023016998

Since **F*< F(table)**, it indicates that the observed F-statistic is not statistically significant at the chosen significance level. In other words, there is insufficient evidence to reject the null hypothesis. This means that the wind speed **(X(ws))** variable is not statistically significant in explaining the variation in the dependent variable (y) when considering the other independent variables (X(tod), X(p), X(rh), X(dp), X(cc)).

3. *Testing Ho $\beta$ (cloud cover) = 0 given $\beta$ (time of day) $\beta$ (pressure) $\beta$ (relative humidity) $\beta$ (dewpoint)*

   ***VS***

   *Testing Ho $\beta$ (cloud cover) <> 0 given $\beta$ (time of day) $\beta$ (pressure) $\beta$ (relative humidity) $\beta$ (dewpoint)*

**F*** = [ SSReg X(cc),X(tod),X(p),X(rh),X(dp)) – SSReg(X(tod), X(p),X(rh),X(dp)] / df X(cc), X(tod), X(p),X(rh),X(dp) – df X(tod), X(p),X(rh),X(dp)]

---

[SSCT-SSReg[(X(cc),X(tod),X(p),X(rh),X(dp)]/(n-1 - #Xvariables)

**F*** = (504.915-502.718) / (5-4)

(573.341-504.915)/ (60-1-5)

**F*** = 1.733710126

**F(table)** = 4.01954096

Since **F*< F(table)**, it indicates that the observed F-statistic is not statistically significant at the chosen significance level. In other words, there is insufficient evidence to reject the null hypothesis. This means that the cloud cover **(X(cc))** variable is not statistically significant in explaining the variation in the dependent variable (y) when considering the other independent variables (X(tod), X(p), X(rh), X(dp)).

4. *Testing Ho $\beta$ (times of day*pressure) = 0 given $\beta$ (time of day) $\beta$ (pressure) $\beta$ (relative humidity) $\beta$ (dewpoint) $\beta$ (times of day*pressure) $\beta$ (times of day*relative humidity) $\beta$ (times of day*dewpoint)*

   ***VS***

   *Testing Ho $\beta$ (times of day*pressure) <> 0 given $\beta$ (time of day) $\beta$ (pressure) $\beta$ (relative humidity) $\beta$ (dewpoint) $\beta$ (times of day*pressure) $\beta$ (times of day*relative humidity) $\beta$ (times of day*dewpoint)*

**F*** = [ SSReg X(tod*p),X(tod),X(p),X(rh),X(dp), X(tod*rh) X(tod*dp)) – SSReg(X(tod), X(p),X(rh),X(dp), X(tod*rh), X(tod*dp)] / df X(tod*p), X(tod), X(p),X(rh),X(dp), X(tod*rh)

X(tod*dp) – df X(tod), X(p),X(rh),X(dp), X(tod*rh), X(tod*dp)]

[SSCT-SSReg[(X(tod*p),X(tod),X(p),X(rh),X(dp), X(tod*rh) X(tod*dp))]/(n-1 - #Xvariables)

**F*** = (510.557-509.345) / (7-6)

(573.341-510.557)/ (60-1-7)

**F*** = 1.003781594

**F(table)** = 4.0266314

Since **F*< F(table)**, it indicates that the observed F-statistic is not statistically significant at the chosen significance level. In other words, there is insufficient evidence to reject the null hypothesis. This means that the times of day*pressure **X(tod*p)** variable is not statistically significant in explaining the variation in the dependent variable (y) when considering the other independent variables (X(tod), X(p), X(rh), X(dp), X(tod*rh), X(tod*dp).

## 3. Assumptions Check:

If all four assumptions of regression are satisfied, it generally means that the linear regression model is valid and can be used to make predictions with a reasonable degree of accuracy. The four assumptions of regression are:

### a. Homoscedasticity:

Homoscedasticity means that the variability or spread of the residuals (errors) is consistent or constant across all levels of the independent variable. This assumption is important because violation of homoscedasticity can lead to biased and inefficient estimates of the regression coefficients. It can also affect the validity of hypothesis tests and confidence intervals.

According to this rule, first, calculate the standard deviation of the errors (residuals) for the first half of the data and divide it by the standard deviation of the errors for the second half of the data. If the resulting ratio falls within the range of **0.5 to 2**, it suggests that the variance of the errors is relatively constant across the two halves of the data, indicating homoscedasticity.

In the data, the value obtained from the division of the two standard deviations is **0.897447528** which falls within the range. Thus, confirms homoscedasticity.

### b. Autocorrelation:

The Durbin-Watson test is a statistical test that helps detect autocorrelation in the residuals of a regression model. It specifically tests for the presence of first-order autocorrelation, which is the correlation between adjacent residuals. The Durbin-Watson statistic (DW) ranges from **0 to 4**, with a value of **2** indicating no autocorrelation. The test statistic is calculated as follows:

$$D = \Sigma \{e(t)-e(t-1)\} * \{e(t)-e(t-1)\} / \Sigma e(t) * e(t)$$

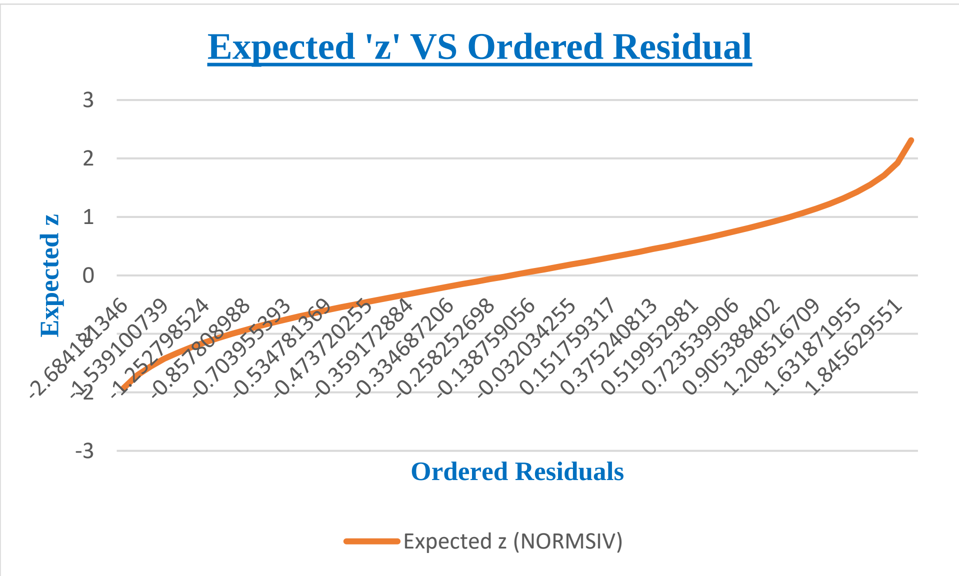
To interpret the Durbin-Watson statistic:

DW close to 2 (around 2) indicates no significant autocorrelation.
DW significantly less than 2 (towards 0) suggests positive autocorrelation.
DW significantly greater than 2 (towards 4) suggests negative autocorrelation.
Moreover, if the observed value of d is **below** dl, then the null hypothesis is rejected and there is autocorrelation and vice versa. But if the observed value of d is **between** dL and dU then it means, there is no sufficient evidence to prove that there is auto correlation.

In the data, the value obtained from the calculation of this formula is **1.630821761.** Here the critical values obtained from the table, given that $\alpha=0.05$, $k=6$, and $N= 60$, are **dL = 1.372** and **dU = 1.808**. Since the value obtained from the calculation is between dL and dU, therefore, there is no sufficient evidence to prove that there is autocorrelation.

### c. Normality:

This is called a normal probability plot. Should look like a straight line if errors are normally distributed. If the middle looks like a straight line but ends, one or both, do not follow the straight line, then the tails are either too light or too heavy for a normal distribution. The tails help determine whether your claim level of significance is what is really the case. However, some deviation from normality is often expected in real-world data, and small departures from normality may not significantly impact the validity of the regression model.




**Plot ordered errors <u>versus</u> expected Z (NORMSINVERSE)**

## <u>4. Procedures/Steps:</u>

### A. Regression

1. Initially, the data was taken from this website: https://open-meteo.com/en/docs/historical-weather-api
2. Then the data was sorted and according to the requirements of the project, 60 observations were picked.

3. First, I started to work on my data with the **BACKWARD** method by running a Regression for all six variables of X along with the indicative variable.
4. The result of that regression reveals three significant "**P**-values" (i.e., less than 0.05); **Time of Day** (indicative variable) - the most significant, **Pressure** - the next significant, and **Relative Humidity** - the third significant.
5. Apparent Temperature, Windspeed, Dewpoint, and Cloud cover reflected the insignificant variables as their P-values were greater than 0.05.
6. To check the significance of these four variables again, they are not dropped at once but instead one by one, so that it is evident that perhaps some variables are in real significant but showing insignificance due to the presence of any of the others.
7. For this reason, the Apparent Temperature dropped at first as its P-value was the highest among the four.
8. Then another regression was run which reveals that the **Dewpoint** is showing significance as its P-value is now less than 0.05. However, wind speed and cloud cover still show insignificance, windspeed being the most.
9. In the next step, Windspeed was dropped before running the third regression. The result reveals that Cloud cover is still insignificant.
10. So, the next regression only considered the four significant variables; **Time of Day, Pressure, Relative Humidity, and Dewpoint** to check their significance again.
11. The result of regression confirmed the significance of all four variables.
12. Since one of the significant variables involves an indicative variable, for this reason, three new columns were added indicating the indicative variable times the x variable. These include: i) **Time of Day*Pressure,** ii) **Time of Day*Relative Humidity, and** iii) **Time of Day*Dewpoint.**
13. Afterward, another regression was taken which consists of four significant X variables and three additional variables that show indicative variables times X variables as disclosed above.
14. The result reveals that the **Time of Day*Pressure** is not significant.
15. variables and 2 X variables times indicative variable.
16. The result reveals that the indicative variable is insignificant. This marks the end of regression.

## B. Transformation Check:

1. To check the data if there is a need for transformation, the first step is to take the **half slope**.
2. For this reason, at first, the data is divided into **three** equal parts, since the number of observations is 60. Therefore, three sets of 20 equal observations are made.
3. For each set of observations, the **median** is taken for Pressure, Relative Humidity, and Dewpoint along with the median of the **Y** Variable. **Note:** As the data is a **time series**, therefore, median must be taken without sorting the data from lowest to highest observation.
4. After taking the medians, first, the difference between medians is taken (in such a way that first the **difference of median** between the **second** set of observations and the **first** set of observations and then the difference of median between the **third** set of observations and the **second** set of observations) and then the ratio of that difference of Y by X is obtained.

5. Based on the presented calculation, it is evident that the data set does not require transformation due to the divergent signs observed in the two half slopes variables, with the third variable being outside the permissible range.

1. **First Variable:**

| **Calculations of Half Slope** | | | |
|---|---|---|---|
| | **Pressure Midpoints** | **Y Midpoints** | |
| M**p**1….> | 1021.945833 | 10.96666667 | <….M**Y**1 |
| M**p**2….> | 1017.470833 | 8.841666667 | <….M**Y**2 |
| M**p**3….> | 1019.025 | 9.8875 | <….M**Y**3 |
| | | | |
| first difference of m2 and m1 of both then the ratio of y by x | 0.474860335 | | |
| first difference of m3 and m2 of both then the ratio of y by x | 0.672922252 | | |
| ratio of 1st over second | 0.705668944 | | |

2. **Second Variable:**

| **Calculations of Half Slope** | | | |
|---|---|---|---|
| | **Relative Humidity Midpoints** | **Y Midpoints** | |
| M**rh**1….> | 67.45833333 | 10.96666667 | <….M**Y**1 |
| M**rh**2….> | 70.625 | 8.841666667 | <….M**Y**2 |
| M**rh**3….> | 71.75 | 9.8875 | <….M**Y**3 |
| | | | |
| first difference of m2 and m1 of both then the ratio of y by x | -0.671052632 | | |
| first difference of m3 and m2 of both then the ratio of y by x | 0.92962963 | | |
| ratio of 1st over second | -0.721849444 | | |

3. **Third Variable:**

| **Calculations of Half Slope** | | | |
|---|---|---|---|
| | **Dewpoint Midpoints** | **Y Midpoints** | |
| M**d**1….> | 1.420833333 | 10.96666667 | <….M**Y**1 |
| M**d**2….> | 1.35 | 8.841666667 | <….M**Y**2 |
| M**d**3….> | 4.0625 | 9.8875 | <….M**Y**3 |
| | | | |
| first difference of m2 and m1 of both then the ratio of y by x | 30 | | |
| first difference of m3 and m2 of both then the ratio of y by x | 0.385560676 | | |
| ratio of 1st over second | 77.80876494 | | |

## C. Calculation of Equation:

(Handwritten done on the hardcopy)

## **5. Conclusions**:

In conclusion, this report analyzed weather data and developed regression models to predict temperature based on various predictor variables such as temperature, relative humidity, dew point, apparent temperature, pressure, cloud cover, and wind speed. The final equation for predicting temperature during the day and night included the variables of pressure, relative humidity, and dew point. The analysis revealed that these variables have significant effects on temperature. The time of day was also included as a categorical variable, capturing the diurnal temperature patterns. The coefficients of the predictor variables in the equations indicated the estimated effects on temperature. The analysis did not require any transformations of the variables. The findings confirmed the initial choice of variables based on their known relationships with temperature. The regression models can provide valuable insights into temperature variations and contribute to weather forecasting and understanding local climate dynamics.