# Predictive Analytics

# Part One—Identify a Focal Point and a Dependent Variable

Target of analysis: Predicting the apparent temperature given the humidity

In our daily life we do our works whatever weather is maybe good or bad, we need to do our job every day. Someday we see very pleasant weather, someday very hot, someday too cold or someday is totally rainy or cloudy.

Obviously we don’t have any control over the nature’s behavior but it would be very interesting to analyze this wonderful behavior of nature. In this project I’m going to analyze how these temperature is changing every day. What are the factors those are controlling the temperature so in turn the everyday weather. I’m not going to forecast the weather but I will try to find some relationship between temperature as a dependent variable and other factors those are controlling the weather considering humidity as a major factor which will be one of the independent variable.

**Dependent variable:** Temperature

|  |  |  |
| --- | --- | --- |
| **Independent Variable** | | |
| **Summary of independent variable** | **Categorical or quantitative?** | **Argument for / description of the associates with the dependent variable** |
| Humidity | Quantitative | With increase in humidity there is a temperature decrease. |
| Visibility | Quantitative | With increase in visibility there is small increase in temperature (Very weak relationship) |
| Wind Speed | Quantitative | There is no clear relationship with Wind speed. (Very weak relationship) |
| Pressure | Quantitative | Scatter plot with pressure is a white noise so we can say that there is no relationship with pressure and temperature. |

## **Part Two—Map Decisions to Outcomes**

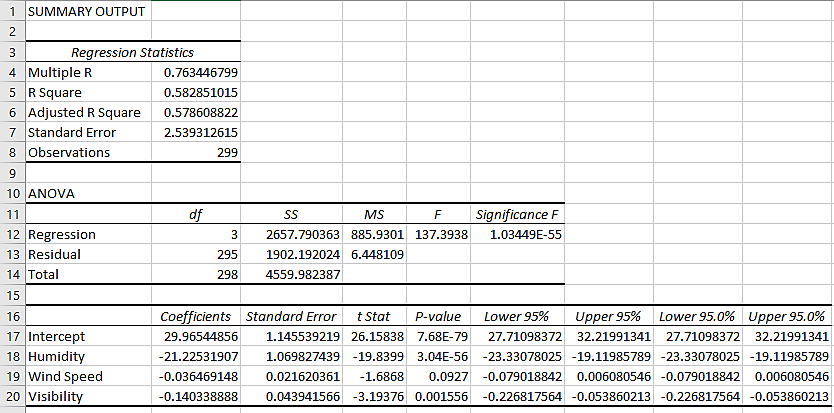
Data Sample Preview:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Formatted Date | Precip Type | Humidity | Wind Speed (km/h) | Visibility (km) | Pressure (millibars) | Temperature (C) |
| 2006-04-01 00:00:00.000 +0200 | Partly Cloudy | 0.89 | 14.1197 | 15.8263 | 1015.13 | 9.472222222 |
| 2006-04-01 01:00:00.000 +0200 | Partly Cloudy | 0.86 | 14.2646 | 15.8263 | 1015.63 | 9.355555556 |
| 2006-04-01 02:00:00.000 +0200 | Mostly Cloudy | 0.89 | 3.9284 | 14.9569 | 1015.94 | 9.377777778 |
| 2006-04-01 03:00:00.000 +0200 | Partly Cloudy | 0.83 | 14.1036 | 15.8263 | 1016.41 | 8.288888889 |
| 2006-04-01 04:00:00.000 +0200 | Mostly Cloudy | 0.83 | 11.0446 | 15.8263 | 1016.51 | 8.755555556 |
| 2006-04-01 05:00:00.000 +0200 | Partly Cloudy | 0.85 | 13.9587 | 14.9569 | 1016.66 | 9.222222222 |
| 2006-04-01 06:00:00.000 +0200 | Partly Cloudy | 0.95 | 12.3648 | 9.982 | 1016.72 | 7.733333333 |
| 2006-04-01 07:00:00.000 +0200 | Partly Cloudy | 0.89 | 14.1519 | 9.982 | 1016.84 | 8.772222222 |

|  |  |  |
| --- | --- | --- |
| **Candidate Independent variables** | | |
| **Independent variable** | **Regression equation** | **Screenshot of scatterplot** |
| Humidity | y = -19.92x + 26.987 |  |
| Visibility | y = 0.0988x + 11.104 |  |
| Wind Speed | y = 0.0746x + 11.254 |  |

From the above scatter plot, we can see the strong negative correlation between humidity and temperature means that when humidity increases temperature decreases. But if we look into the scatter plot with wind speed and visibility, there is very little change in temperature with wind speed and visibility. So we can say very weak relationship there.

## Multiple Regression Analysis:

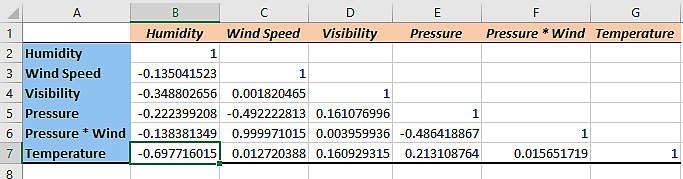


# Part ThreeGenerate a Revised Regression Equation

### Residual Plots:

## Addressing nonlinear relationships:

From the above residual plots we can’t see any non-linearity in dependent variable with respect to the considered independent variables.



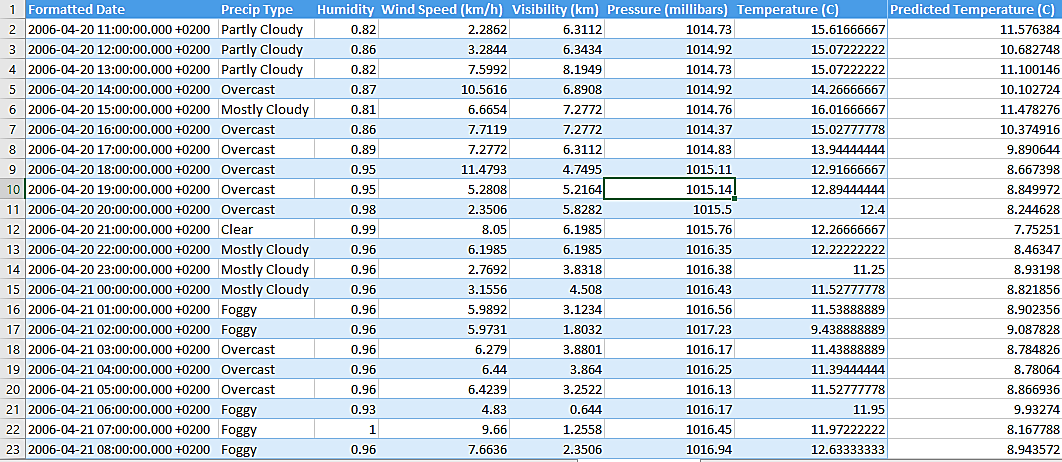
In the above correlation table, we can see that the correlation between humidity and temperature is significant. But the other variables have no or very less significant impact on the temperature. Also as the correlation between considered independent variables is very less so multi-collinearity error has very less impact on the regression analysis.

## Revised Regression Equation:

Temperature (Y) = 29.96 - 21.23 \* Humidity – 0.04 \* Wind Speed – 0.14 \* Visibility

# Part Four—Validate Your Model

## Predictive analysis using the model:



Finally we have applied our model on the holdout sample data. As we saw that there is no relation with pressure we have not considered the variable in regression equation. From the above preview we can see that our prediction value is close the actual value. As we can see humidity is the most impacting variable also we can see the p value for humidity is very small or close to zero so there is a relationship between humidity and temperature.