# Weather Forecast Dataset

## STA 4700

By: Eugene Monforte

### Abstract:

Our dataset consists of 5 numeric predictor variables, Temperature, Humidity, Wind Speed, Cloud Cover and Pressure to predict our categorical response variable Rain. We fit a regression model and after using backward and forward stepwise regression, we found that the following equation was the best fit:

$$\log\left(\frac{p(Rain)}{1-p(Rain)}\right) = -12.0819 - .2683 \cdot Temperature + .1356 \cdot Humidity + .0822 \cdot Cloud\ Cover$$

After, we used Wald Test to determine that Temperature, Humidity and Cloud Cover all had statistically significant relationships in the likelihood of rain. We were also able to get the odds ratio that a 1°C increase in Temperature reduces the odds of Rain by $\approx 20-26\%$, a 1% increase in Humidity increases the odds of Rain by $\approx 13-17\%$, and a 1% increase in Cloud Cover increases the odds of Rain by $\approx 7.5-9.8\%$.

## Introduction (Description of Data):

Source: Kaggle (https://tinyurl.com/any73fwb)

I am using a dataset set from Kaggle about Weather forecast. The dataset has 2500 observations with 6 variables. Kaggle didn't provide much information about the variables or how they were collected. Here is a table explaining the variables based on the scope of our data and their range of values.

| Variable Name | Description | Numerical or Categorical? |
|---|---|---|
| Temperature | Degrees in Celsius | Numerical |
| Humidity | Percentage of Humidity | Numerical |
| Wind_Speed | The speed of the wind at an unknown observation time | Numerical |
| Cloud_Cover | Percentage of Humidity | Numerical |
| Pressure | Atmospheric pressure at ground level | Numerical |
| Rain | Response variable indicating if it rained or not | Categorical |

We will perform preliminary data analysis to get an idea of what the data looks like. This will consist of a summary of the data, boxplots of each variable against rain, and a stacked boxplot of rain.

Next, we will perform Logistic Regression to try and get our "best" fitting model. We will then use Wald Test to determine the significant variables and generate odds ratios with confidence intervals to be able to generate interpretations from this model.

Therefore, the statistical methods used are Logistic Regression, Wald Test, Odds Ratio and Confidence Intervals for Odds Ratio.

We will be using Backward Stepwise Regression and Forward Stepwise Regression to determine the best fitting model. Since the two were not covered in class, a brief introduction will be provided. Backward Stepwise Regression is where we start with the full model and eliminate one variable at a time based on Akaike Information Criterion (AIC) until the model does not improve anymore. Similarly, Forward Stepwise Regression is where we start with the null model and proceed by adding one variable at a time with the variable that improves AIC the most until the model does not improve at all. AIC is a metric to compare models based on fit and complexity of the model.

## Preliminary Data Analysis:

We begin with preliminary data analysis to get a better understanding of our data. We create a summary of our data to get the following values:

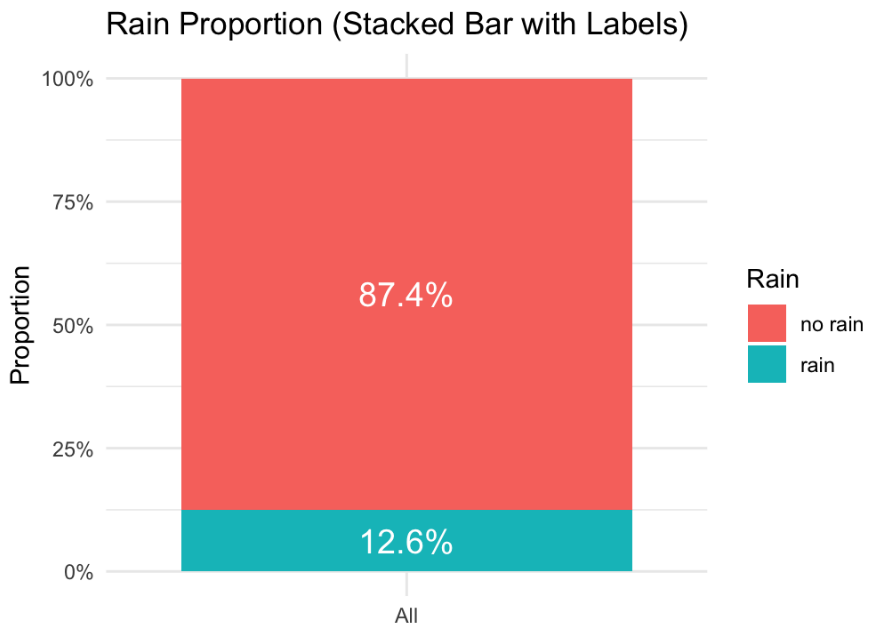
| **Variable** | **Min.** | **$1^{st}$ Quartile** | **Median** | **Mean** | **$3^{rd}$ Quartile** | **Max** |
|---|---|---|---|---|---|---|
| Temperature | 10.0 | 16.36 | 22.54 | 22.58 | 28.98 | 35.00 |
| Humidity | 30.01 | 47.34 | 63.92 | 64.35 | 81.56 | 100.00 |
| Wind_Speed | .01 | 4.76 | 9.91 | 9.91 | 14.95 | 19.99 |
| Cloud_Cover | .02 | 23.90 | 49.49 | 49.66 | 75.32 | 99.99 |
| Pressure | 980.0 | 996.9 | 1013.4 | 1014.3 | 1031.7 | 1050.0 |

The table provides an insight into the range and distribution of each of our variables. Looking at the mean and median of each variable, it appears to be roughly in the midpoint of the minimum and maximum of our data. This suggests that our data appears to be symmetric and not heavily skewed.

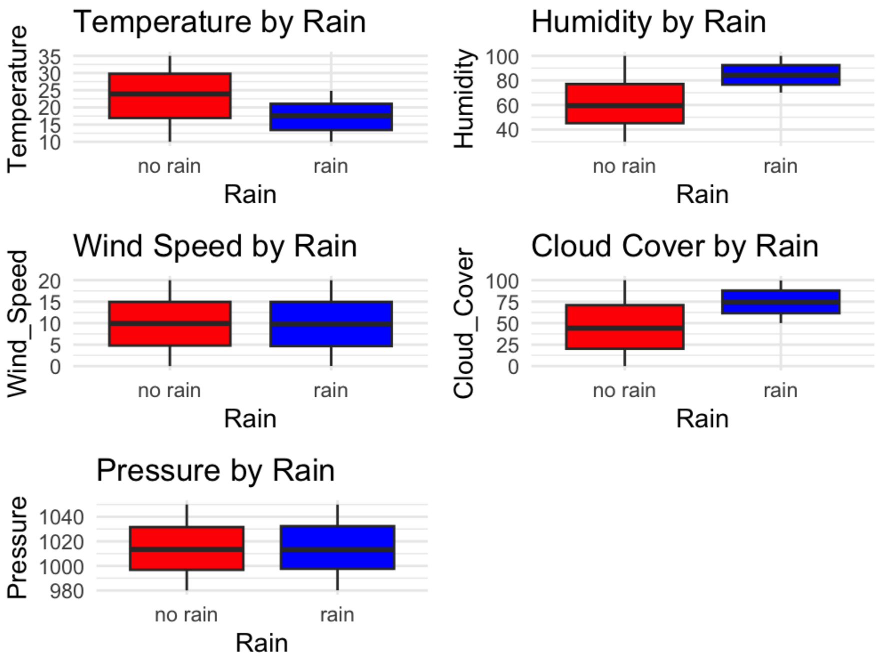
We are also able to get the following frequency table from the summary statistics.

| Rain | No Rain |
|---|---|
| 314 | 2186 |

There appears to be a lot more No Rain than Rain in our dataset. Visualizing this we get:



Only 12.6% of our Rain variable did the weather rain, while 87.4% of the days did it not rain. Now creating boxplots for each variable by rain, we get the following 5 boxplots:



The boxplot for Temperature by Rain shows that No Rain is a result of a higher Temperature. The boxplot for Humidity by Rain shows that higher Humidity results in Rain occurring. The boxplot for Cloud Cover by Rain shows that the higher percentage of Cloud Cover results in Rain. Wind Speed and Pressure do not appear to show a relationship on whether or not it rains since the median of the boxplots appear to be even. The relationships shown here may give us an idea of what variables will appear to be important during our data analysis.

**Data Analysis using Logistic Regression:**

We begin our data analysis by fitting a logistic regression model using all 5 of our variables. We get the following equation,

$$\log\left(\frac{p(Rain)}{1-p(Rain)}\right) = -12.24 - .2683 \cdot Temperature + .1357 \cdot Humidity$$

$$-.0059 \cdot Wind\ Speed + .0822 \cdot Cloud\ Cover + .0002\ \mathrm{Pr}\,e\,suure$$

Using the summary statistics, we perform the Wald Test to assess whether each coefficient is significantly different than 0 using 95% confidence level.

$$H_0\colon\ \beta_i = 0\ where\ i\ =\ 1,2,3,4,5$$

$$H_A\colon\ \beta_i \neq 0\ where\ i\ =\ 1,2,3,4,5$$

| $i$ | p-value |
|---|---|
| 1 | <2e-16 |
| 2 | <2e-16 |
| 3 | .7199 |
| 4 | <2e-16 |
| 5 | .9648 |

We see that $i = 1,2,4$ have a coefficient statistically different than 0 at the 95% confidence level. This shows that Temperature, Humidity and Cloud Cover are significantly associated with Rain. Contrastly, $i = 3,5$ have a coefficient not statistically different than 0 at the 95% confidence level. Thus, Wind Speed and Pressure are not significantly associated with Rain. The results here align with our preliminary analysis.

Using the step() function in R, which automatically does forward and backward stepwise model selection based on AIC. The function determines the following model is our "best" model in terms of AIC:

$$\log\left(\frac{p(Rain)}{1-p(Rain)}\right) = -12.0819 - .2683 \cdot Temperature + .1356 \cdot Humidity$$

$$+ .0822 \cdot Cloud\ Cover$$

Now we perform Wald's Test again for each regression coefficient.

$$H_0\colon\ \beta_i = 0\ where\ i\ =\ 1,2,3$$

$$H_A\colon\ \beta_i \neq 0\ where\ i\ =\ 1,2,3$$

| $i$ | p-value |
|---|---|
| 1 | <2e-16 |
| 2 | <2e-16 |
| 3 | <2e-16 |

So, at the 95% confidence level, we determine that Temperature, Humidity and Cloud Cover are all statistically significant to Rain.

Based on our model, we can create a table of our Odds Ratio Confidence Intervals for each variable.

| **Variable** | **Odds Ratio w Confidence Interval** | **Interpretation** |
|---|---|---|
| Temperature | .765 (.736 , .792) | A 1°C increase in Temperature reduces the odds of Rain by $\approx 20-26\%$ |
| Humidity | 1.145 (1.127 , 1.165) | A 1% increase in Humidity increases the odds of Rain by $\approx 13-17\%$ |
| Cloud Cover | 1.086 (1.075 , 1.098) | A 1% increase in Cloud Cover increases the odds of Rain by $\approx 7.5-9.8\%$ |

## Conclusion

In conclusion, we used Logistic Regression to fit a model with all our variables. Using Wald Test, we determined that not all our variables were statistically significant. After using the step() function in R, backward and forward stepwise model selection was used to determine our best model. Our best model was the following:

$$\log\left(\frac{p(Rain)}{1-p(Rain)}\right) = -12.0819 - .2683 \cdot Temperature + .1356 \cdot Humidity$$

$$+ .0822 \cdot Cloud\ Cover$$

From the model, we were able to determine the Odds Ratio of each of our variables. We determined: A 1°C increase in Temperature reduces the odds of Rain by $\approx 20-26\%$ , a 1% increase in Humidity increases the odds of Rain by $\approx 13-17\%$ , and a 1% increase in Cloud Cover increases the odds of Rain by $\approx 7.5-9.8\%$ .