STATS 101A - Final Project Report: Predicting Wind Speed from Key Weather Variables

Introduction

The goal of this research project is to develop a predictive model for wind speed using a set of commonly measured weather variables. In this project, the response variable is wind speed (measured in km/h), and the predictors are temperature, humidity, visibility, and pressure. The dataset utilized in this study was obtained from Kaggle.

- Wind speed (km/h): The response variable represents the speed of the wind.
- Temperature (°C): Air temperature in Celsius
- Humidity (as a proportion): The relative humidity
- Visibility (km): The horizontal visibility
- Pressure (millibars): The atmospheric pressure

The data pulls information from 2006 to 2016 in Szeged, Hungary: Szeged Weather Data

Understanding how temperature, humidity, visibility, and pressure influence wind speed is critical not only for weather forecasting but also for a wide range of applications that affect our daily lives and various industries, such as agriculture, aviation, and renewable energy. The study aims to quantify the relationship between these predictors and wind speed, providing insights into the factors that most strongly drive variations in wind dynamics. Multiple linear regression was employed to model the relationship between the predictor and response variables.

Data description

The minimum, first quartile, median, mean, third quartile, and maximum for each variable of study are summarized below.

	Wind	Temperature	Humidity	Visibility	Pressure
Min	0.00000	-12.16111	0.2000000	0.00000	989.100
Q1	6.02140	5.00000	0.6100000	8.42030	1011.890
Median	9.82100	12.08889	0.7900000	10.25570	1016.260
Q3	14.24850	18.87222	0.9000000	14.95690	1020.500
Max	38.73660	36.06111	1.0000000	16.10000	1042.970
Mean	10.83443	12.14104	0.7382234	10.47119	1016.419

"Wind Speed" is the response variable representing the speed of the wind in kilometers per hour(km/h). The mean wind speed is 10.83443 km/h, with a minimum of 0.00000 km/h and a maximum of 38.73660 km/h.

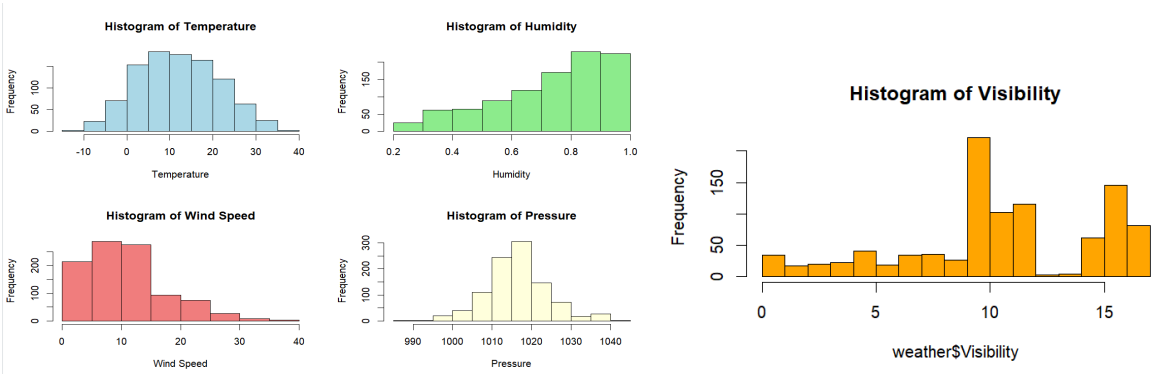
"Temperature" represents the air temperature in Celsius (°C). The mean temperature is 12.14104°C, with a maximum of 36.06111°C and a minimum of -12.16111°C. The distribution appears fairly symmetric, suggesting that most temperature values fall between 5°C and 19°C. Extreme values indicate occurrences of cold weather (-12.16 °C) and high temperature (36.06 °C), reflecting a wide range of climate conditions.

“Humidity” represents relative humidity, a measure of moisture in the air relative to its maximum capacity. The mean humidity is 0.7382 (73.82%) with a minimum of 0.2000 (20%) and a maximum of 1.0000 (100%).

"Visibility" represents horizontal visibility in kilometers (km), which indicates the distance at which an object can be clearly seen. The mean visibility is 10.47119 km, with a minimum of 0.0000 km and a maximum of 16.1000 km.

"Pressure" represents atmospheric pressure measured in millibars (mb). The mean atmospheric pressure is 1016.419 mb, with a minimum of 989.100 mb and a maximum of 1042.970 mb. The distribution appears fairly symmetric, indicating that most pressure values fall within the typical atmospheric range. The minimum pressure of 989.100 mb suggests the presence of low-pressure systems, which are commonly associated with storms, heavy rainfall, and unstable weather conditions. The maximum pressure of 1042.970 mb indicates occurrences of high-pressure systems, which are generally linked to clear skies and stable weather. The range of pressure values suggests that the dataset includes both low and high-pressure conditions, which may influence temperature, humidity, and wind patterns.

Histograms for each variable (Temperature, Humidity, Wind speed, Pressure, Visibility)



Results and Interpretation

We employed a multiple linear regression model to assess the Wind Speed of 985 randomly sampled days between the years 2006 to 2016. We first look at the full model of the data.

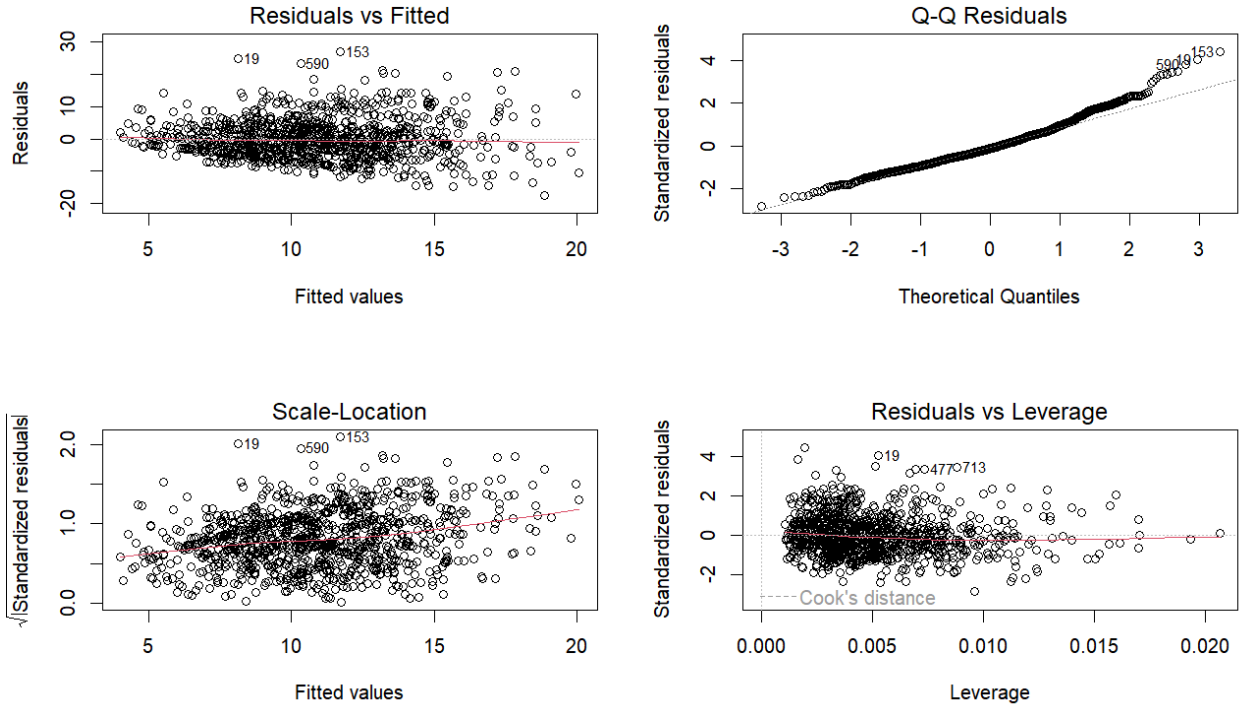
From this summary, we can see that the variables – temperature, humidity, and pressure – all have significant p-values for their coefficients, while the visibility doesn’t. However, the overall F-test is

```
Call:
lm(formula = Wind ~ Temperature + Humidity + Visibility + Pressure,
    data = weather)

Residuals:
     Min       1Q   Median       3Q      Max
-17.4375  -4.0908  -0.7865   3.3118  27.0258

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 284.06530   27.72360  10.246   <2e-16 ***
Temperature  -0.29305    0.02888 -10.146   <2e-16 ***
Humidity    -17.54419    1.36276 -12.874   <2e-16 ***
Visibility   -0.02311    0.05028  -0.460    0.646
Pressure     -0.25234    0.02687  -9.392   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.141 on 980 degrees of freedom
Multiple R-squared:  0.1811,    Adjusted R-squared:  0.1777
F-statistic: 54.17 on 4 and 980 DF,  p-value: < 2.2e-16
```

significant with a p-value less than 0.05, suggesting that at least one predictor variable significantly contributes to explaining wind speed.

Additionally, below are the 4 diagnostic plots of the model for the randomly sampled 1000 days within 2006 to 2016. Based on Residuals vs. Fitted model, there isn't any apparent pattern as most of the residuals are randomly dispersed around zero. However, there are still some residuals towards the right that are more scattered, suggesting minor heteroscedasticity. Looking at the Normal Q-Q Plot, most of the points follow the diagonal line, but the right tail deviates away from the line, indicating some skewness. This shows the plot mainly follows a normal distribution with slight skewness on the right. As for the Scale-Location Plot, it somewhat follows the same pattern in the Residuals vs Fitted plot, where the right side of the model has minor heteroscedasticity because of some variance of residuals at higher fitted values. Lastly, for the Residuals vs Leverage Plot, there doesn't seem to be any influential points that could influence the model.

Diagnostic Plots



Also, when analyzing the Variance Inflation Factor, all the variables have a VIF < 10 so we can reasonably conclude that multicollinearity is not a major concern for our model.

VIF Table

Temperature	Humidity	Visibility	Pressure
1.899484	1.822250	1.196549	1.116475

Based on the 4 diagnostic plots, there was minor heteroscedasticity on the right tail, but the majority of the data is normally distributed. Because of this, there isn't too much evidence

needed to convince us of any transformations; hence, we didn't apply any transformations to our model.

Furthermore, we ran a partial F- test comparing two models: 1) without the visibility predictor and 2) with the visibility predictor. From the code attached below, the p-value (0.6459) is greater than 0.05, suggesting that the visibility predictor is insignificant. The ANOVA table supports the full model that we showed earlier in this section, proving that we are in favor of the reduced model without the visibility predictor.

```
Analysis of Variance Table

Model 1: Wind ~ Temperature + Humidity + Pressure
Model 2: Wind ~ Temperature + Humidity + Visibility + Pressure
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    981 36968
2    980 36960  1    7.9659 0.2112 0.6459
```

Transformations

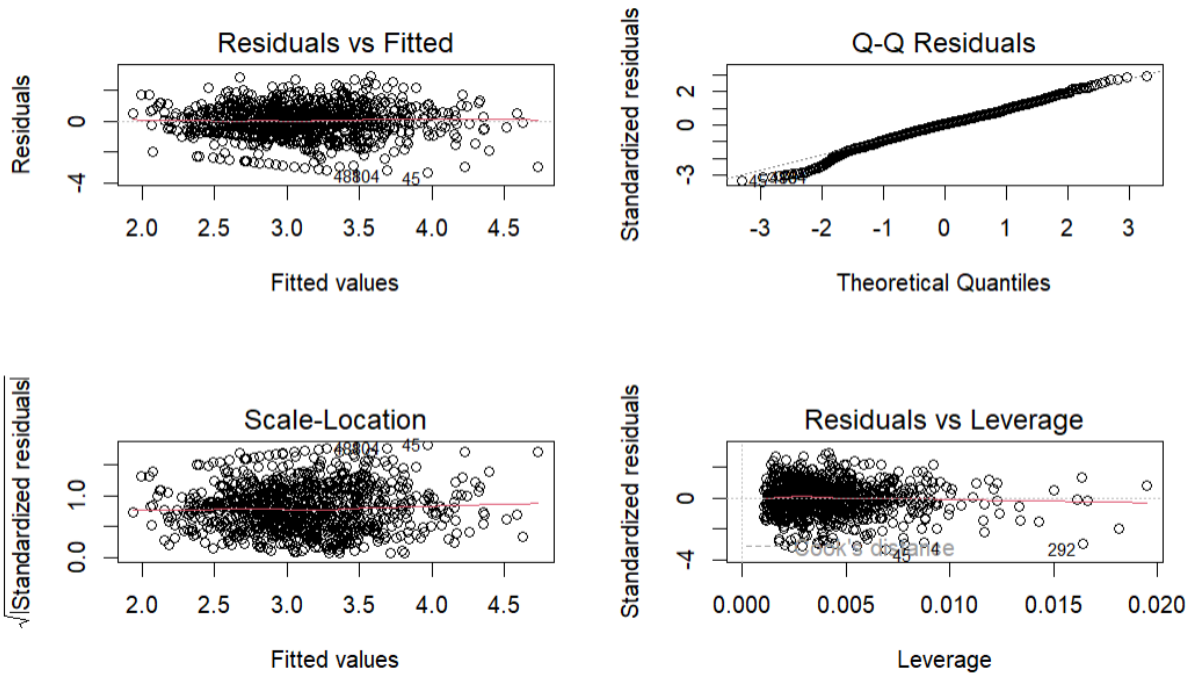
We looked to see if transformations were necessary on this reduced model so we performed a box-cox transformation on Y and all the predictors in the reduced model. These were results:

```
Warning: Convergence failure: return code = 52bcPower Transformations to
Multinormality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.5402        0.50       0.4908       0.5896
Y2    0.9474        1.00       0.7712       1.1235
Y3    1.9610        2.00       1.7397       2.1822
Y4   -2.9105       -2.91      -3.0824      -2.7386

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (0 0 0 0)	1406.922	4	< 2.22e-16

	LRT <dbl>	df <int>	pval <chr>
LR test, lambda = (1 1 1 1)	319.3841	4	< 2.22e-16

These results show that all transformation parameters are not "log", but that transformations are needed. Because of this we tested out a transformed model with the suggested rounded powers and looked at the summary of the new model and its residual plots. The results of the summary of the new model were very similar to the untransformed reduced model, but the adjusted R^2 value slightly decreased.



The only visible difference in the residual plots is that the upper tail on the Normal Q-Q plot, but a slightly heavier lower tail was created. For the reasons of the decreased adj. R^2 value and the appearance of a lower tail on the Normal Q-Q plot, we were not convinced to use the transformations that the box-cox suggested as it would only complicate the interpretation of our model.

Discussion

Through this project, our goal was to gain insights into the meteorological factors influencing wind speed and assess their statistical significance. By analyzing weather data collected from Szeged, Hungary, we developed a multiple linear regression model that accounts for 18% of the variance in wind speed. The variables included in the model were Temperature, Humidity, Pressure, and Visibility, with Humidity and Pressure showing statistical significance, while Visibility was found to be insignificant. The resulting equation, rounded to the thousandths place, is as follows: Wind Speed = 284.065 − 0.293(Temperature) − 17.544(Humidity) − 0.252(Pressure) − 0.023(Visibility).
Since Visibility was not statistically significant (p-value = 0.646), we refined the model by removing this variable. The updated model, which only includes the significant predictors, is:

Wind Speed = 282.486 − 0.295(Temperature) − 17.426(Humidity) − 0.251(Pressure). We can interpret this equation as for every one unit increase in temperature, humidity, and pressure, wind speed decreases by 0.295 for temperature, 17.426 for humidity, and 0.251 for pressure.

Summarizing the steps taken to develop this model, we first assessed the assumptions of multiple linear regression by examining diagnostic plots. Our analysis confirmed that the model met key assumptions. The Residuals vs. Fitted plot showed no strong patterns, and the line of best fit was nearly horizontal, supporting the linearity assumption. The Normal Q-Q plot

indicated that the residuals followed an approximately normal distribution, satisfying the normality assumption.

Additionally, the Scale-Location plot displayed a nearly horizontal trend, suggesting that the homoscedasticity assumption was met, though minor heteroscedasticity was observed at higher fitted values. Next, we evaluated multicollinearity among the variables and found no significant issues, as all VIF scores were below 5.

Subsequently, we assessed the significance of each variable using simple linear regression models. The analysis showed that Humidity and Pressure had statistically significant relationships with wind speed, while Temperature (p-value = 0.454) and Visibility (p-value = 0.646) were not statistically significant.

To further validate our model, we examined residual diagnostics. The Breusch-Pagan test for heteroscedasticity resulted in a p-value of 3.696e-13, indicating the presence of non-constant variance in the residuals. This suggests that additional transformations or alternative modeling approaches might improve accuracy.

Lastly, we assessed variance inflation factors (VIFs) for each predictor, with all values being below 2, confirming the absence of severe multicollinearity. This structured approach allowed us to develop a statistically valid model for understanding the relationship between meteorological factors and wind speed. However, limitations such as heteroscedasticity and the relatively low explanatory power ($R^2 = 0.18$) suggest that while our predictor variables still remain significant, additional ones would further improve prediction accuracy. This is common in linear models for predicting weather and wind speeds as there are so many variables that can contribute to or interfere with the data.

Real World Applicability

The model helps us to predict wind speeds based on temperature, humidity, visibility, pressure, and cloud cover. It can help with real world applications with climate analysis, weather forecasting, and decision-making. Determining wind speeds can prevent hazardous situations such as wildfires, tornados, and hurricanes. This specific dataset, which was all taken from the city of Szeged in Hungary in a span of ten years, will help predict wind speeds for the city under different conditions. The city will be able to take precautions when necessary when certain predictors are starting to show signs of high wind speeds.

Limitations

Because we didn't collect this data, there are several limitations that prohibit us from getting a clearer weather model for Szeged, Hungary between the years 2006 to 2016. For instance, we restricted the given variables to only temperature, humidity, visibility, and pressure, so this would limit us to obtaining a more comprehensive understanding of wind speed. With no specific dates from Szeged included within our model, this prevents us from capturing seasonal trends, which then limits us to predicting long-term climate trends or periodic fluctuations in wind speed over the ten-year period. Additionally, the exclusion of the "Wind Bearing" variable limits us from determining the wind behavior and, in turn, wind speed. Because we only decided to use four predictor variables in our model, it might not be enough to explain the full complexity of the data, indicating why our R^2 value is low regardless of having a significant p-value. There were definitely more predictor variables available to use; however, we didn't want to complicate our model so we focused on the four chosen and the R^2 value reflects wind dynamics not fully accounted for in Szeged.