STAT 306 Group Project

Group 27

9 April 2021

Introduction

Dew point temperature is the temperature at which air is saturated with water vapor at its current state. In this analysis, we wish to explain how dew point temperature is influenced by other meteorological factors such as temperature, relative humidity, wind speed and air pressure.

The present data set is historical weather data collected by a weather station operated by NAVCAN at the Vancouver International Airport. The data was collected in February 2021. The variable definitions are as follows, described with guidance from [1]:

- Temperature: "degree of hot or cold of the air, as measured by a thermometer" ([1])
- Dew point temperature: the temperature where air becomes saturated with water vapor currently in the air; dew forms when air is cooler than this temperature
- Relative humidity: the ratio of the current amount of water vapor to the saturated amount of water vapor in the air at the current temperature, as a percentage
- Wind speed: the speed of wind in km/h at 10m above ground measured in one-, two- or ten minute intervals
- Air pressure: "atmospheric pressure in kilopascals (kPa) at station elevation" ([1])

The standards to which these variables are measured can be found in [2].

Analysis

We start by plotting scatterplots of dew point temperature against all predictors individually.

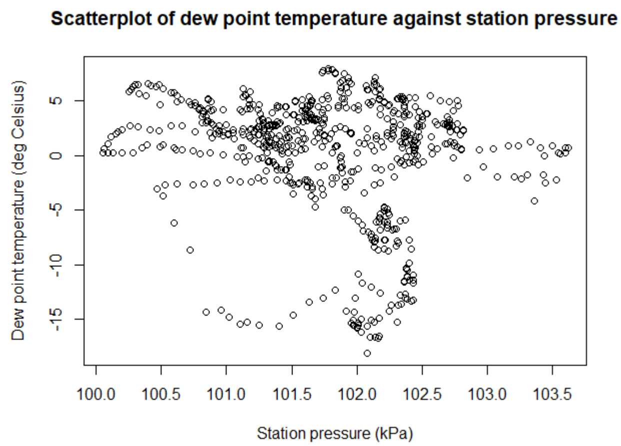

Figure 1: Scatterplot of dew point temperature against atmospheric pressure.

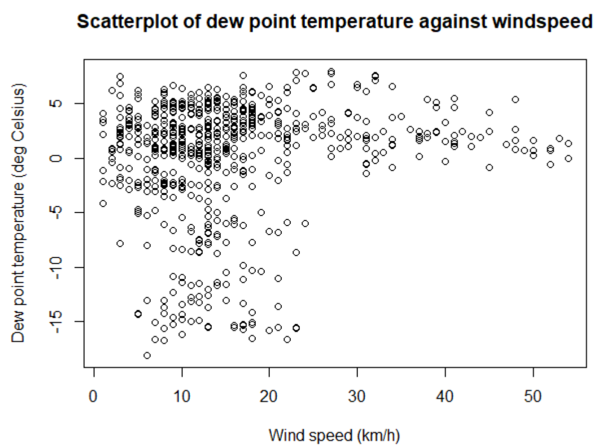

Figure 2: Scatterplot of dew point temperature against wind speed.

The above scatterplots lack any pattern, so they provisionally suggest that wind speed and atmospheric pressure have no effect on dew point temperature. Regardless, we include these predictors in the analysis for the sake of completeness.

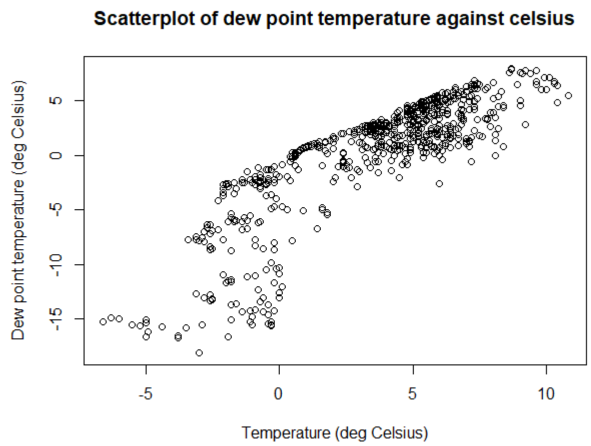

Figure 3: Scatterplot of dew point temperature against temperature.

Now, the points in the scatterplot clearly follow a pattern, suggesting a relationship between dew point temperature and temperature. This relationship seems to be different for temperatures below and above zero degrees celsius. A naive inspection suggests a quadratic relationship for temperatures below zero and a linear relationship for temperatures above zero, both of which will be investigated in the analysis.

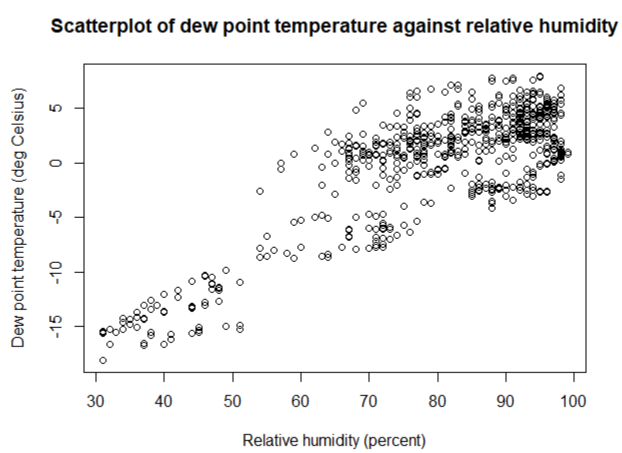

Figure 4: Scatterplot of dew point temperature against relative humidity.

The points in this scatterplot also clearly follow a pattern, so a significant relationship between dew point temperature and relative humidity is likely. The points generally seem to follow a single line, suggesting a simple linear relationship. Again, this will be investigated in the analysis that follows.

Model 1

We first fit a full linear model that includes all predictors without any interaction and normal errors. The R summary of the fit and the residual plot are below.

```
Console  Terminal
D:/Study/Year 3/STAT 306/Group project/Data analysis/
> summary(lin_full)

Call:
lm(formula = weather$Dew.Point.Temp..Â.C. ~ weather$Temp..Â.C. +
    weather$Rel.Hum.... + weather$Wind.Spd..km.h. + weather$Stn.Press..kPa.)

Residuals:
     Min       1Q   Median       3Q      Max
-2.00052 -0.19316  0.00993  0.24716  0.87877

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -40.927977   2.405406 -17.015  < 2e-16 ***
weather$Temp..Â.C.        1.000613   0.005136 194.817  < 2e-16 ***
weather$Rel.Hum....       0.201549   0.001084 185.855  < 2e-16 ***
weather$Wind.Spd..km.h.   0.011352   0.001618   7.016 5.82e-12 ***
weather$Stn.Press..kPa.   0.208280   0.023295   8.941  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3815 on 643 degrees of freedom
Multiple R-squared:  0.9953,    Adjusted R-squared:  0.9953
F-statistic: 3.392e+04 on 4 and 643 DF,  p-value: < 2.2e-16

> |
```

Figure 5: R summary of full linear model

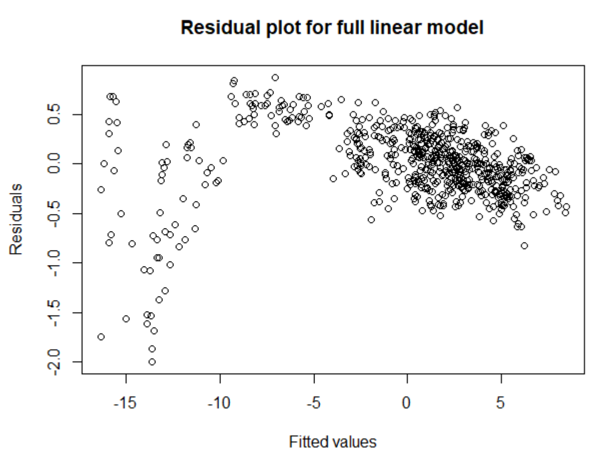

Figure 6: Residual plot of full linear model

Observing the residual plot, it is clear the points show a curved shape. This observation led us to believe that a higher order model in terms of the predictors is likely to be more appropriate. Attempts at fitting models that include interaction terms did not resolve this issue, so interaction terms between the predictors are not included in the analysis from here on.

Model 2

Next, we fit a model that included some quadratic terms. We added two new quadratic terms for temperature and relative humidity, as the shape of the previous residual plot shows that a quadratic relationship is plausible. We did not add quadratic terms for the other predictor variables because the scatterplots of dew point temperature against them lacks any patterns, and it is in our interest to minimize the size of the model.

Also, we recalled that the shape of the scatterplot of dew point temperature against temperature is different for temperatures above and below zero degrees Celsius. We decided to create a new dummy variable, called 'isAbove0' that has two levels: '1' indicating a temperature above zero and '0' indicating below. We added this variable into the model and included its interaction with the linear and quadratic terms for temperature and relative humidity. Lastly, we included interaction between the isAbove0 factor and temperature, and between relative humidity.

The results of this quadratic model are shown below:

```
Console   Terminal ×
D:/Study/Year 3/STAT 306/Group project/Data analysis/

> summary(quad_full)

Call:
lm(formula = weather$Dew.Point.Temp..Â.C. ~ weather$Temp..Â.C. *
    weather$isAbove0 + I(weather$Temp..Â.C.^2) * weather$isAbove0 +
    weather$Rel.Hum.... * weather$isAbove0 + I(weather$Rel.Hum....^2) *
    weather$isAbove0 + weather$Wind.Spd..km.h. + weather$Stn.Press..kPa.)

Residuals:
     Min       1Q   Median       3Q      Max
-0.39004 -0.05476 -0.00226  0.05682  0.28782

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 -2.716e+01  5.834e-01 -46.549  < 2e-16 ***
weather$Temp..Â.C.                           9.783e-01  1.482e-02  66.021  < 2e-16 ***
weather$isAbove01                            2.229e+00  2.180e-01  10.221  < 2e-16 ***
I(weather$Temp..Â.C.^2)                      1.342e-02  2.681e-03   5.007 7.18e-07 ***
weather$Rel.Hum....                          4.207e-01  2.893e-03 145.415  < 2e-16 ***
I(weather$Rel.Hum....^2)                    -1.596e-03  2.253e-05 -70.823  < 2e-16 ***
weather$Wind.Spd..km.h.                     -3.628e-05  3.913e-04  -0.093    0.926
weather$Stn.Press..kPa.                      7.201e-03  5.762e-03   1.250    0.212
weather$Temp..Â.C.:weather$isAbove01         1.735e-02  1.619e-02   1.072    0.284
weather$isAbove01:I(weather$Temp..Â.C.^2)  -1.476e-02  2.749e-03  -5.369 1.11e-07 ***
weather$isAbove01:weather$Rel.Hum....       -7.042e-02  5.894e-03 -11.946  < 2e-16 ***
weather$isAbove01:I(weather$Rel.Hum....^2)   5.134e-04  3.882e-05  13.224  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08744 on 636 degrees of freedom
Multiple R-squared:  0.9998,    Adjusted R-squared:  0.9998
F-statistic: 2.358e+05 on 11 and 636 DF,  p-value: < 2.2e-16

>
```

Figure 7: R summary of quadratic model with interaction

In this model, the significance of wind speed and station pressure has been lost, while the linear and quadratic terms for temperature and relative humidity are very significant. The interaction terms of 'isAbove0' with the linear and quadratic terms of relative humidity and with the quadratic term of temperature are also very significant, so it was appropriate to include this dummy variable. The residual standard error is also much lower in this model compared to the last, which indicates that this is a much better fit.

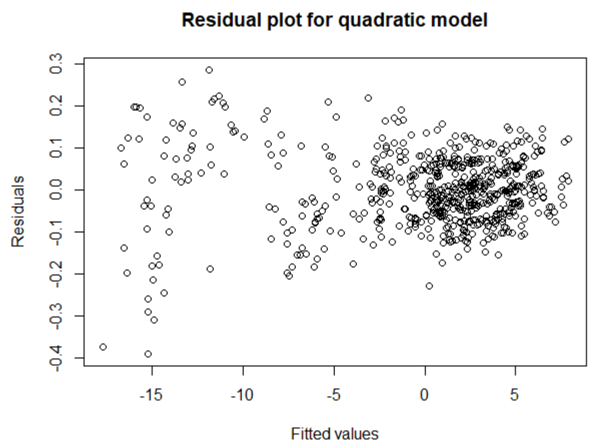

Figure 8: Residual plot of quadratic model with interaction

Upon observing the residual plot for this model, the points show less trend than the previous model, which is promising. The 'isAbove0' variable, along with the quadratic terms, seem to have reduced the pattern in the residual. However, there are a few issues. An initial examination of the residual plot suggests that there is heteroscedasticity, shown by the "funneling" behaviour of the residuals. A closer inspection through only plotting the points corresponding to temperatures below zero degrees showed that there is still a cubic pattern among these points.Therefore, although this model is a better fit than the linear model, the issues outlined here indicate that it is still a poor fit.

Model 3

In the third model, we added cubic terms for temperature and relative humidity. These predictors were chosen for the inclusion of cubic terms due to the same reasons in Model 2 (i.e. the scatterplots of dew point temperature against the other predictors lack any pattern). Interaction terms between 'isAbove0' and the cube of temperature and relative humidity are also included. The results are as follows.

Console Terminal

D:/Study/Year 3/STAT 306/Group project/Data analysis/

```
Call:
lm(formula = weather$Dew.Point.Temp..Â.C. ~ weather$Temp..Â.C. *
    weather$isAbove0 + I(weather$Temp..Â.C.^2) * weather$isAbove0 +
    I(weather$Temp..Â.C.^3) * weather$isAbove0 + weather$Rel.Hum.... *
    weather$isAbove0 + I(weather$Rel.Hum....^2) * weather$isAbove0 +
    I(weather$Rel.Hum....^3) * weather$isAbove0 + weather$Wind.Spd..km.h. +
    weather$Stn.Press..kPa.)

Residuals:
      Min        1Q    Median        3Q       Max
-0.227627 -0.050536  0.001631  0.051549  0.208683

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 -3.115e+01  5.275e-01 -59.056  < 2e-16 ***
weather$Temp..Â.C.                           1.083e+00  2.788e-02  38.852  < 2e-16 ***
weather$isAbove01                            4.595e+00  6.510e-01   7.060 4.41e-12 ***
I(weather$Temp..Â.C.^2)                      5.886e-02  1.232e-02   4.776 2.22e-06 ***
I(weather$Temp..Â.C.^3)                      5.695e-03  1.405e-03   4.052 5.70e-05 ***
weather$Rel.Hum....                          6.459e-01  1.299e-02  49.708  < 2e-16 ***
I(weather$Rel.Hum....^2)                    -5.283e-03  2.101e-04 -25.139  < 2e-16 ***
I(weather$Rel.Hum....^3)                     1.898e-05  1.077e-06  17.618  < 2e-16 ***
weather$Wind.Spd..km.h.                     -9.634e-05  3.189e-04  -0.302   0.7627
weather$Stn.Press..kPa.                      5.156e-03  4.765e-03   1.082   0.2797
weather$Temp..Â.C.:weather$isAbove01        -9.048e-02  3.087e-02  -2.931   0.0035 **
weather$isAbove01:I(weather$Temp..Â.C.^2)   -5.931e-02  1.266e-02  -4.685 3.43e-06 ***
weather$isAbove01:I(weather$Temp..Â.C.^3)   -5.761e-03  1.417e-03  -4.065 5.41e-05 ***
weather$isAbove01:weather$Rel.Hum....       -2.201e-01  2.752e-02  -7.999 6.02e-15 ***
weather$isAbove01:I(weather$Rel.Hum....^2)   3.192e-03  3.852e-04   8.287 6.99e-16 ***
weather$isAbove01:I(weather$Rel.Hum....^3)  -1.458e-05  1.778e-06  -8.197 1.37e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07118 on 632 degrees of freedom
Multiple R-squared:  0.9998,	Adjusted R-squared:  0.9998
F-statistic: 2.61e+05 on 15 and 632 DF,  p-value: < 2.2e-16
```

Figure 9: R summary of cubic model with interaction

The results of this cubic regression are promising. The linear, quadratic and cubic terms for temperature and relative humidity are significant as well as their interactions with 'isAbove0', and residual standard deviation reduced slightly. The wind speed and station pressure terms are not significant.

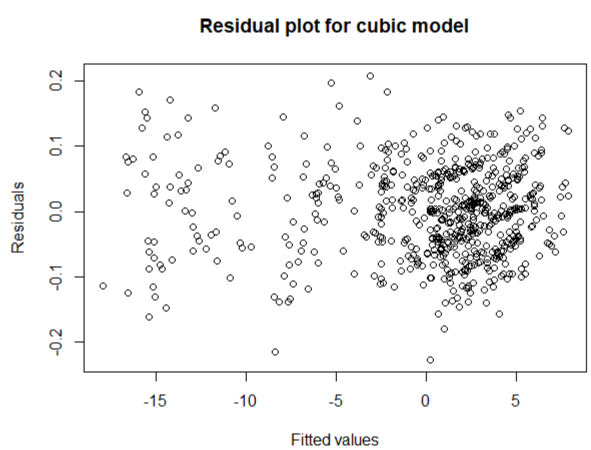

Figure 10: Residual plot of cubic model with interaction

The residual plot for this model is now patternless, so the fit of this model is satisfactory. Further plotting with standardized residuals against both fitted values and predictors also show patternless residuals with only one or two outliers. Hence we select the following estimated model for the data:

$$Y = -31.15+1.083x_1+0.05886x_1^2+0.005695x_1^3+0.6459x_2-0.005283x_2^2+(1.898\cdot 10^{-5})x_2^3$$
$$-(9.634\cdot 10^{-5})x_3+0.005156x_4+z(4.595-0.9048x_1-0.05931x_1^2-0.005761x_1^3-0.2201x_2$$
$$+0.003192x_2^2-(1.458\cdot 10^{-5})x_2^3)+\varepsilon$$

where Y is the current dew point temperature in degrees Celsius, x_1 is the current temperature in degrees Celsius, x_2 is the current relative humidity as a percentage, x_3 is the current wind speed in km/h, x_4 is the current atmospheric pressure in kPa, z is 1 when the current temperature is strictly larger than 0 and is 0 otherwise, and ε is a normally distributed error term.

The statistical interpretation of the results from R for this model is that there is enough evidence in the data to suggest that dew point temperature has a cubic relationship with temperature and relative humidity. In contrast, there is insufficient evidence in the data to support the hypothesis that wind speed and station pressure have an effect on dew point temperature. The coefficients suggest that dew point temperature generally increases as temperature and relative humidity increase, but the effect is not as strong when the temperature is above zero degrees celsius.

There are several limitations to the current data analysis. The analysis does not take into account the time dependence of the data. It is decided that this is not a significant issue because the time scale is relatively small, and that the current context does not permit time to be a reasonable independent variable. Additionally, the cubic model has a poor mean-squared error when three-fold cross-validation is applied to it, so the model has poor predictive power. This issue also raises the concern of whether this model overfits the data. Other kinds of models, such as those involving transformations of dew point temperature, interaction between predictors and/or higher order terms of wind speed and atmospheric pressure, should be fitted to this data set as attempts to find models that are simpler and/or have more predictive power. Also, the data set also has more points with temperatures above zero degrees than points with temperatures below zero degrees. On one hand, this issue might cause the analysis to be biased towards conditions involving above freezing temperatures, which may be resolved by splitting the data set into two parts, one with data points with temperatures above zero degrees and the other with data points with temperatures below zero degrees, and performing analyses on the two parts separately. On the other hand, such an approach makes a comparison between above and below freezing conditions less direct, which costs the explanatory power of the model. An exploration of this approach as well as other possible models is omitted from this study due to time constraints.

Conclusion

We have fitted a full linear model, a quadratic model and a cubic model to the data in order to investigate the effects of meteorological factors on dew point temperature. The cubic model is deemed the most appropriate out of the three. It is determined that dew point temperature has a cubic dependence on temperature and relative humidity, whose effects differ below and above zero degrees celsius; in addition, wind speed and air pressure seem to have no effect on dew point temperature according to the analysis. We have noted the limitations of this analysis, which invites further analysis of this data set.

References

[1] Environment and Climate Change Canada. *Glossary - Climate.* https://climate.weather.gc.ca/glossary_e.html.

[2] Environment and Climate Change Canada. *Manual of Surface Weather Observation Standards - Eighth Edition.* Monitoring and Data Services Directorate, Meteorological Service of Canada, 2019.