Predicting the Mean Temperature for a location in Ontario in July 2024

Imtiaz Kidwai

2024-06-12

Introduction

This analysis focuses on predicting the mean temperature across the coming month i.e July 2024 for a location in Ontario using data from July 2023. Meteorology has always been personally fascinating to me as I have always been curious about how weather patterns are predicted, and while I may not be able to satisfy that curiosity yet, learning linear regression gave me tools to make a reasonable prediction. Predicting temperature has increasing relevance today due to the impact of climate change. In particular, summers are getting hotter and thus it is important for the average person to be aware of what temperatures may look like so that they can make necessary preparations. The phenomenon of climate change is studied by scientists all around the world and thus, weather prediction models may be useful to their research. Even without the context of climate change, weather prediction models hold much relevance in everyday life. Weather forecasts are used to plan almost every outdoor occasion, and these models are significant to research in Physics, Earth Science and Environment Science.

There is a lot of research done on predicting temperatures. One research article displays the use of linear regression to predict daily maximum temperatures (Massie et al., 1997, pp. 799). The article used Eta Geopotential Thickness Forecasts to model the temperature, which in itself is modelled by other variables (Massie et al., 1997, pp. 801). One research article also involves forecasting temperature at weather stations, but they used deep neural networks instead of linear regression (Saha Roy, 2020, pp. 38). There is a research article that used multiple linear regression to predict monthly temperature and it involves using maximum and minimum temperatures, so the article provided the most insight on what my analysis may be missing (Gupta et al. , 2022, pp.1). The model also includes maximum and minimum dew points as predictor variables, which my model does not account for (Gupta et al. , 2022, pp. 4). This shows that dew points also impact temperature. Overall, there is a lot of research on predicting weather, using techniques such as linear regression and deep neural networks. Weather forecasting is a widely popular topic of research.

Methods

I acquired the data from Environment Canada's Monthly Summaries website, where I set the location to Ontario, month to July and year to 2023. For each weather station in Ontario, the dataset had information on the mean temperature through the month (in celsius), the number of days with the mean temperature, the highest maximum temperature recorded by the weather station in the month (in celsius), the number of days with that maximum temperature, the lowest minimum temperature recorded by the weather station in the month (in celsius), the number of days with that minimum temperature, total precipitation (in mm), total snowfall (in cm), snow on the ground last day (cm), number of days with precipitation with 1.0 mm or more, days with valid bright sunshine, cooling degree days and heating degree days. For some of the aforementioned variables, data was either missing or not related to mean temperature. Additionally, any data related to snowfall was irrelevant due to lack of snowfall in July. Lastly, the number of days with mean temperature, cooling degree days and heating degree days were modelled or calculated using mean temperature, which is exactly what we wanted to predict, so I omitted those variables as well.

This left eight variables that could possibly be incorporated into the model: Precipitation (P), High-

est Maximum Temperature (Tx), Lowest Minimum Temperature (Tn), Days with Precipitation (DwP), Days with Minimum Temperature (DwTn), Days with Maximum Temperature (DwTx), Longitude (Long) and Latitude (Lat).

To determine the best possible model, I decided to start with all eight parameters, and drop the parameter with the highest P-value for the t-test. This is because a low P-value implies that the coefficient is likely not zero i.e the predictor likely affects the response variable. Thus, by eliminating the parameter with the highest P-value, we eliminate the variable that is most likely to not affect the model. To decide when to stop, we used a mix of adjusted R-squared, Akaike's Information Criterion and Bayesian Information Criterion. A better model would have higher Adjusted R-Squared and/or lower AIC and BIC. Table 1 in the Results section was used to organize this process.

To check for multicollinearity between the predictors, I computed the Variance Inflation Factor (VIF) $\frac{1}{1-R_j^2}$ for each $j \in \{1, ..., n\}$ where R_j^2 is the coefficient of determination for the fitted model where X_j is the response variable for every other predictor. I checked whether the VIF was greater than 5 for any predictor, and if it was, I would have to use Principal Components Analysis as a remedy. Table 2 in the results section displays the VIF for all the predictors

To verify the assumptions for linear regression, I used the residuals vs fitted plot and the normal QQ plot (Figure 1 and Figure 2 in the results section respectively).

In particular, we can analyze linearity, independence of the error terms and if the variances are constant using the residuals vs fitted plot. Linearity is violated if there is an obvious systematic pattern, independence of error terms is violated if there is a large cluster of points that are obviously separated form the rest, and the constant variance assumption is violated if there is a fanning pattern in the scatter plot. The normality of the error terms can be analyzed using the normal QQ plot. The assumption is violated if there is too large of a deviation from the diagonal line at the endpoints.

To validate the model, I made sure that all the analysis I performed so far was done on a training set, which contained 60% of the original data. A test set contains the other 40%. Table 3 shows the estimated coefficients for the fitted model for both the training and test set. A fitted a model for the test data as well. For the sake of simplicity, we will call it the test fit, while the model that was fitted based on the training data will be called the training fit. I analyzed the estimated coefficients and the goodness-of-fit of both the fits to see if they are similar, and I checked if the violation status of the assumptions of linear regression was the same for each assumption for both models.

Results

The data can be summarized with a small table about variable characteristics as well as histograms on each variable. Table 4 (Appendix) displays the minimum value, maximum value, mean, median and variance for each of the four predictors and response. Note that the training and test sets are combined here. An interesting observation is that the mean of the mean temperature through the month (Tm) is just the midpoint of the mean of Tn and the mean of Tx. Figure 3 in the appendix shows the histogram for all five variables. We can notice that there is no symmetry in the data for any of the variables.

The results for finding the best model can be found below.

Table 1 shows the process of deciding the best model

Table 1: Finding the Best Model

Number of Parameters	Highest P-value	Parameter with Highest P-value	Adjusted R-Squared	AIC	BIC
8	0.965	DwP	0.9355	198.1917	223.6247
7	0.649	DwTn	0.9362	196.1939	219.0835
6	0.382	P	0.9368	194.4222	214.7686
5	0.331	DwTx	0.9370	193.2511	211.0542
4	0.0674	Long	0.9370	192.2652	207.5249
3	1.94 e^(-15)	Lat	0.9353	193.8175	206.534

Starting with eight parameters, I identified the parameter with highest P-value, and recorded the adjusted R-squared, AIC and BIC of the regression model. I would then remove that parameter, and repeat the process until I was penalized for it i.e the adjusted R- squared decreased or AIC or BIC increased. Notice how once there were three parameters left, removing the parameter with the highest P-value would be punishing since the adjusted R-squared decreased and both the AIC and BIC increased. Thus, the model with the three remaining parameters (Tx, Tn, Lat) was the best model.

However, notice also that Longitude was the parameter that was removed immediately before that. Since my research question involves using location to model mean temperature, I made the decision to continue with longitude as a predictor even though it did not do a great job of predicting mean temperature. Thus, our final model uses Highest Maximum Temperature (Tn), Lowest Minimum Temperature (Tx), Latitude (Lat) and Longitude (Long) as predictor variables.

In terms of problematic observations, the residuals vs leverage plot, which can found in the appendix, shows two observations that had a standardized residual fairly higher than 2. Every other observation was almost within the interval [-2, 2]. Thus, I decided that the data with those two observations removed would fit a better model.

We also avoided a problematic amount of multicollinearity between the predictors.

Table 2 proves the lack of multicollinearity between the predictors.

Table 2: Variance Inflation Factor for all the variables

Variable	VIF
X_1	1.28123
X_2	2.941176
X_3	4.384042
X_4	2.308936

Recall that the rule of thumb is that multicollinearity becomes problematic if the VIF is greater than

5. Since that was not the case here, we do not have to worry about addressing multicollinearity in our model.

The model also satisfies all the assumptions for linear regression. Figure 1 is used to verify that linearity of the model and the independence of the error terms are not violated and that the variances are constant.

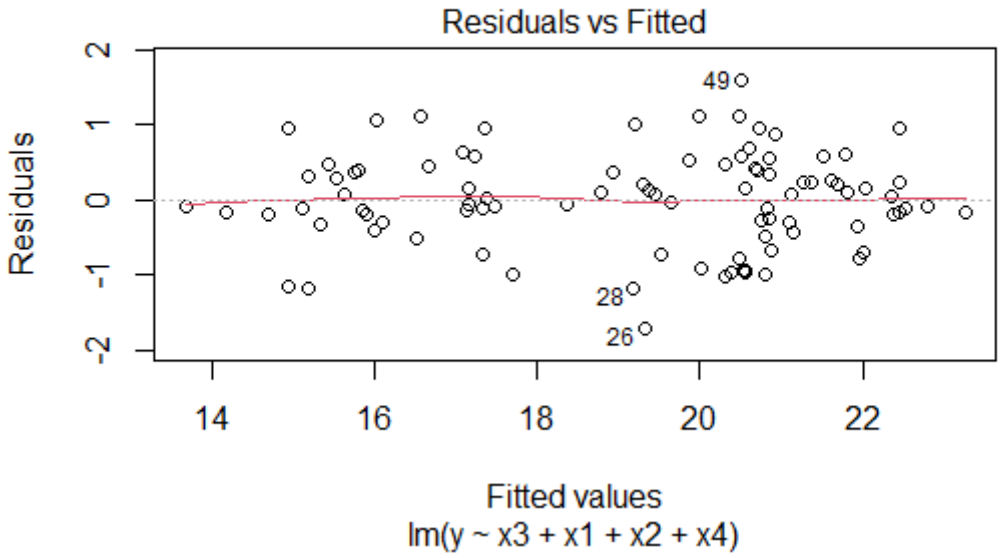

Figure 1: Residuals vs Fitted plot

We can notice a lack of a systematic pattern in the scatter plot, which suggests that the linearity of the model and the constant variance assumptions hold.

In Figure 2, we can notice a straight diagonal line for most of the points, and only some deviation at the end. This suggests that the error terms indeed follow a normal distribution.

Finally, analyzing the test fit and training fit shows that our model is valid. Table 3 shows the estimated regression coefficients for both the test model and training model as well as the differences between the coefficients. This gave us a sense of how much the models agree with each other.

Table 3: Estimated Regression Coefficients for the model fitted from two different sets of data

Variable	Estimated Regression Coefficient (Training Dataset)	Estimated Regression Coefficient (Test Dataset)	Difference in absolute value
Intercept	21.66233	21.624348	0.037982
X_1	0.34867	0.418024	0.069354

Variable	Estimated Regression Coefficient (Training Dataset)	Estimated Regression Coefficient (Test Dataset)	Difference in absolute value
X_2	0.37744	0.335042	0.042398
X_3	-0.29357	-0.409286	0.115716
X_4	0.03622	-0.009791	0.046011

The differences between the coefficients of the training model and the coefficients of the test model are all within 0.1, except for the coefficient of the latitude variable.

Furthermore, the goodness-of-fit is 0.9401 and 0.941 for the training fit and test fit respectively. We also have a difference within 0.1 here.

Looking at the plots, we can verify that all regression assumptions for the test fit are met. The VIF for all the variables are less than 5 so we also do not have to worry about multicollinearity within the predictors in the test fit.

With such minimal differences in the models and with all the assumptions met for both fits, we can claim that the model is the same for two different sets of data, which implies that the model is indeed valid. Thus, our final model is $Y = 21.66233 + 0.34867X_1 + 0.37744X_2 - 0.29273X_3 + 0.03622X_4 + \epsilon$.

Discussion

We can look back at the training data column in Table 3. The intercept tells us that the average mean temperature in July at 0 latitude and 0 longitude (which is Null Island) if the highest maximum and lowest minimum temperature in July were both 0 degrees celsius is 21.66233 degrees celsius. We can already start noticing some flaws with the model. The coefficient for Tx tells us that the average mean temperature through July increases by 0.34867 degrees celsius for every 1 C* increase in highest maximum temperature. The coefficient for Tn tells us that the average mean temperature through July increases by 0.37744 degrees celsius for every 1 C* increase in lowest minimum temperature. The coefficient for Lat tells us that the average latitude decreases by 0.29357 C* for every 1 degree increase in latitude. The coefficient for Long tells us that the average mean temperature through July decreases by 0.03622 C* for every 1 degree increase in longitude.

The coefficient for latitude being negative tells us that the lower the latitude is i.e the further South we go in Ontario, the higher the average mean temperature is, which makes intuitive sense. Similarly, we can see that an increase in Tx and Tn corresponds to a higher average mean temperature. Longitude is more difficult to analyze because the coefficient is relatively close to zero which tells us that the variable has a negligible impact on mean temperature.

We can use this to predict the mean temperature for July for any location. For example, where I live the latitude and longitude are 43.595310 and -79.640579 respectively. If I predict that the highest maximum temperature in July will be 34 C* and the lowest minimum temperature will be 11 degrees celsius. Then, we can predict that the mean temperature here is approximately 21.66233

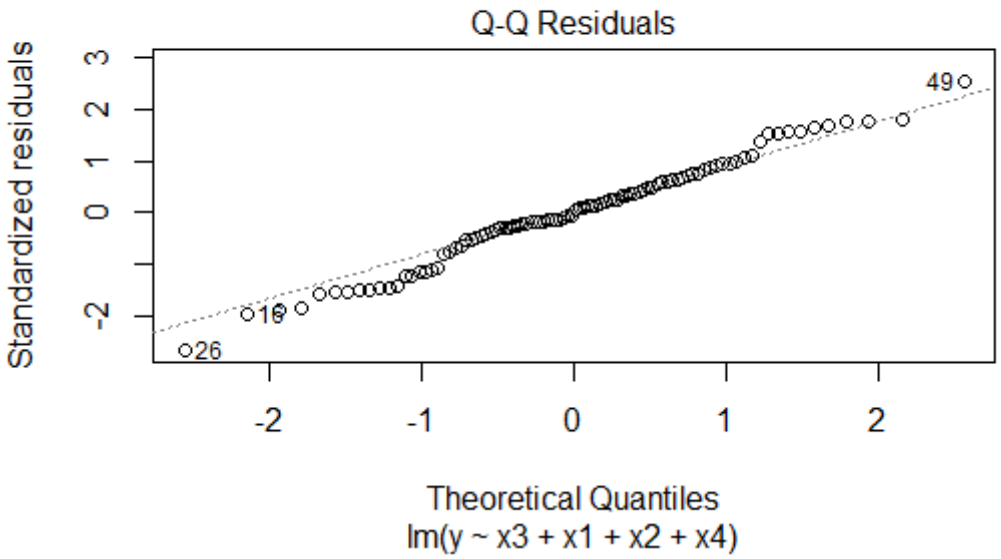

Figure 2: Normal QQ Plot of the model

+ 0.34867(34) + 0.37744(11) - 0.29273(43.595310) + 0.03622(-79.640579) = 22.02271 degrees celsius.

There are a lot of problems with this model and this is why there are meteorologists out there with much more education and experience doing weather prediction. There a lot of factors other than the predictors in this model that affect temperature. We already saw in the introduction section that dew points and wind speed are some examples. Incorporating all these variables into the model allows for a more accurate prediction of the mean temperature.

Furthermore, two of the variables are highest maximum temperature in the month and lowest minimum temperature. While we use these variables to predict mean temperature through the month, we do not actually know what the values for these variables might be. We have to go through the process of estimating the values for these variables before we estimate mean temperature itself. Depending on how difficult it would be to estimate these variables, this model may not be the most efficient way to predict mean temperature.

Lastly, the usefulness of the research question itself can be questioned here. A simple internet search provides robust and accurate information on temperature, precipitation, wind speed, etc for each hour. Being able to predict the mean temperature through a specific month in a specific province in a specific country does not serve any usefulness to the average person. It may still be useful in other research contexts, but even then there are more reliable sources for this information. Ideally, a more useful model would predict the temperature for any time period in any area in the world.

All these limitations stem from the fact that climate data is difficult and expensive to collect and manage, which is why it is not possible to have enough free public data available to answer a research question like this. An ideal data set would have many more variables for more time periods available for every region around the world. This would allow us to form a more useful research

question that could be answered with a more accurate model. Of course, it would also require more time to analyze!

Appendix

Table 4: Numerical Summary of the data

Variable	Max	Min	Median	Mean	Variance
Longitude	-74.75	-94.38	-80.69	-82.34	24.9729
Latitude	56.02	41.95	45.11	46.32	10.98146
Highest Maximum Temperature	34.60	23.00	31.35	31.06	3.54255
Lowest Minimum Temperature	16.800	0.200	9.050	8.681	12.30707
Mean Temperature	23.70	12.70	19.65	19.28	6.612226

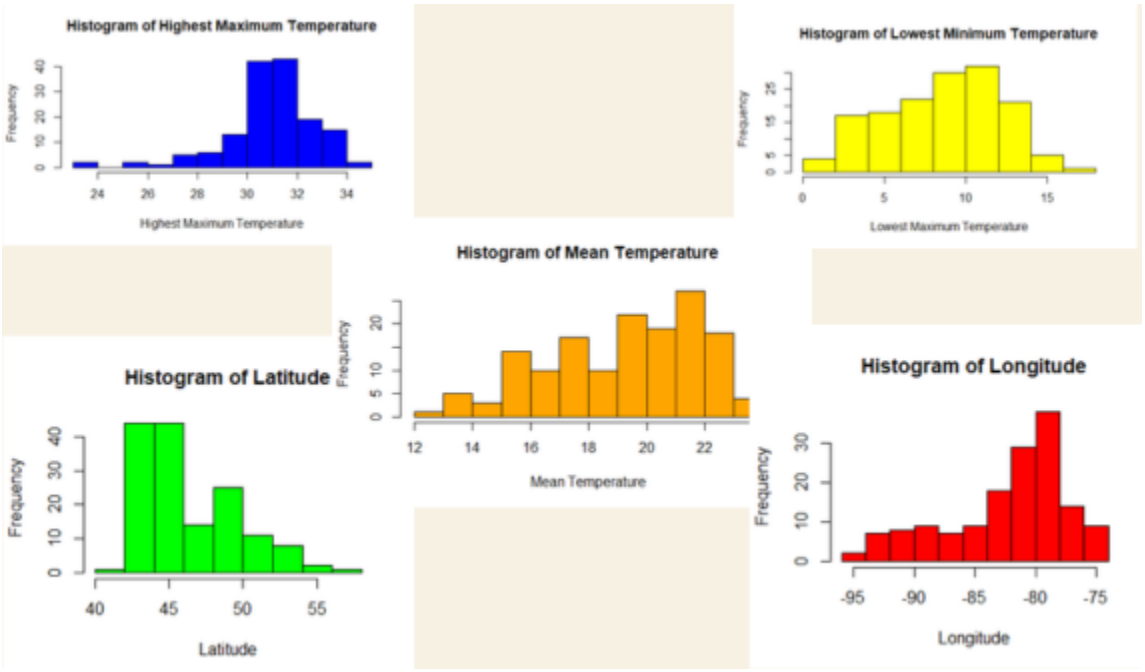

Figure 3: Visual Summary of the data

References

Gupta, I., Mittal, H., Rikhari, D., & Singh, A. K. (2022). *MLRM: A Multiple Linear Regression based Model for Average Temperature Prediction of A Day*. https://doi.org/10.48550/arxiv.2203.05835

Massie, D. R., & Rose, M. A. (1997). Predicting Daily Maximum Temperatures Using Linear Regression and Eta Geopotential Thickness Forecasts. Weather and Forecasting, 12(4), 799-807. https://doi.org/10.1175/1520-0434(1997)012%3C0799:PDMTUL2.0.CO;2

Roy, D. S. (2020). Forecasting The Air Temperature at a Weather Station Using Deep Neural Networks. Procedia Computer Science, 178, 38–46. https://doi.org/10.1016/j.procs.2020.11.005

Roy, D. S. (2020). Forecasting The Air Temperature at a Weather Station Using Deep Neural Networks. *Procedia Computer Science*, *178*, 38–46. https://doi.org/10.1016/j.procs.2020.11.005