**Introduction**

Linear regression is a very powerful statistical process. In fact, most of the data problems can be solved using linear regression. In this project, I want to show you how you can use linear regression to solve a real life application: weather prediction. I will predict the weather of South Bend, where I attend University of Notre Dame, based on data of South Bend in 2017.

**Dataset**

The dataset is generated using MERRA-2. The MERRA-2 takes into the measured observations and computes the temperature based on past data. The dataset focus on Winter 2017 from December 1st 2016 to February 28th 2017. The domain is a 182x101 (longitudexlatitude) grid comprising the whole United States and Mexico. But I will only focus on the South Bend area.

The variables are:

**albedo**: the proportion of the incident light or radiation that is reflected by a surface, typically that of a planet or moon.

**CLDHGH**: an absolute number from 0 to 1, measuring the fraction of clouds at surface pressure to 400 millibars (mb) to to the top of the atmosphere, basically high clouds like cirrus clouds

**Cldmid & cldlow**: the same, but for mid cloud and low cloud.

SWGDN is the surface incoming shorwave ux, inWatts per meter square.

**radiation**: incident shortwave radiation, in Watts per square meters (W/m2)

**QLML**: specific humidity, which is an absolute number (no unit measure) from 0 to 1 denoting ratio of the mass of water vapor to the total mass of air parcel

**SPEED**: wind speed, in meters per second.

**PRECTOT**: the total precipitation, in Kg per meter square per second.

**QLML**: the specific humidity, unitless and a ratio between 0 and 1.

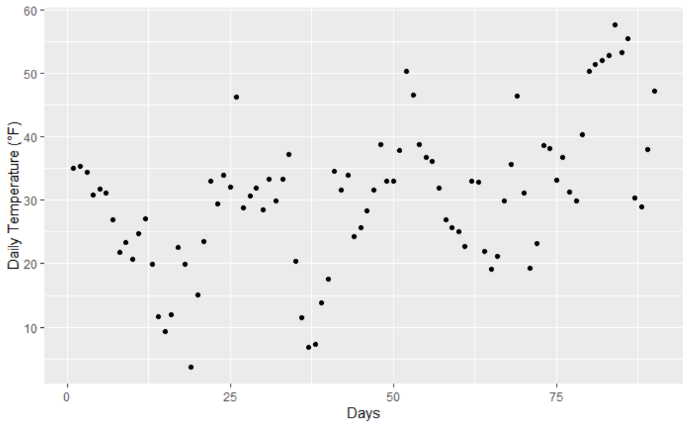
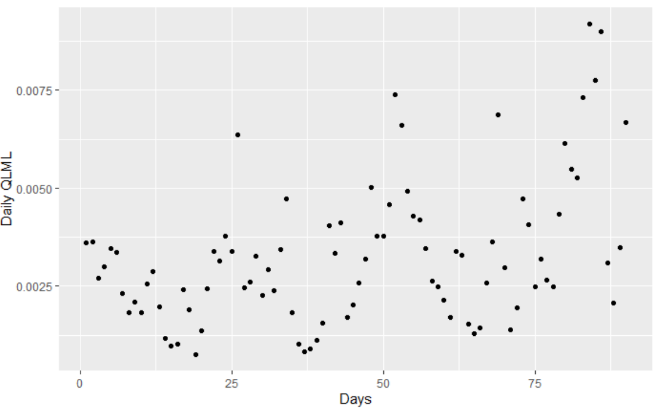
**PRECSNO**: the snowfall rate, in Kg per meter square per second.

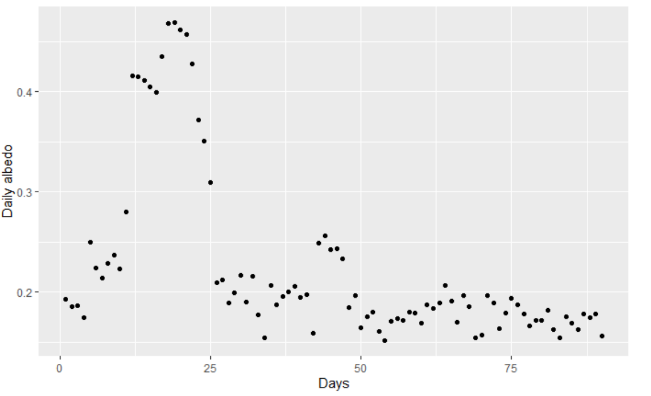
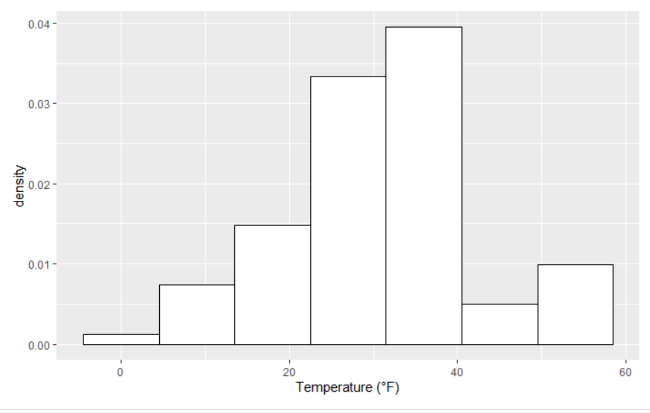
One of the potential issue of data quality is that the errors in collected observations. So the errors in collected observations is carried to the errors in the model. Some mix of observations and bias in observations and model can introduce spurious variability and trends into reanalysis output.

**Exploratory Data Analysis**

First I decide to look at the correlation to see if there is any potentially interesting predictors. From the correlation table, I find that QLML and albedo are highly correlated to temperature with correlation of 0.93 and -0.50 respectively, which indicate that they might be good indicators for temperature. However, we should be careful because QLML and albedo are correlated with correlation of -0.39. Also, snowfall rate and total precipitation are highly correlated with correlation of 0.64, which makes sense because on snow day the snow particles contain water and increase. SWGDN is negatively correlated with all the cloud predictors, which makes sense because the more cloud the less radiation can get through. Albedo has some positive correlation with snow rate because the more snow the albedo is more reflective.

Then I look at the graph of each predictor vs day (Figure 1). Looking at temperature graph vs day, we can see that the graph of temperature is very dependent on time. **The graph of humidity looks like there is a strong correlation with temperature as I mentioned, and the relationship is roughly linear, so linear model is appropriated**. Looking at the albedo, there is an odd spike on t=13 to t=25, I am not sure what they are, they might be outliers. Other graphs I don’t report because they look fine. Also on Figure 1 final graph: density of temperature, we see that the distribution of temperature is skewed to the right and has a fat tail.

**Figure 1**: Daily temperature vs days, Daily humidity vs days, Daily abedo vs days, and density of temperature

The snowfall rate and total precipitation are divided are either low or really high, which makes sense because they indicate days with snow. I decided to make all the values of snowfall rate and total precipitation that are below 0.0025 to be 0 and values that high 0.0025 to be 1. I think this is a reasonable assumption because the new value indicates that that day there is rain or snow.

**Model selection**

**Choosing predictors:**

First lets choose the model without interaction:

Running linear regression through a full model, we see that there are 3 predictors that are significant according to the p-value for testing if the coefficient is different from 0: albedo, QLML, and CLDLOW. This matches our previous understanding that albedo and humidity are important to predict. CLDLOW might be important because the cloud covers the sun and reduces the temperature.

Next I will perform an extensive search of the coefficients on all of the criterions. The criterions I will be using are: R2, R2 adjusted. Mallow Cp. Press, AIC and BIC.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| coefficients | R2 | R2 adjusted | Mallow Cp | press | AIC | BIC |
| 1 | 0.874 | 0.872 | 58.63 | 1529 | 510 | 517 |
| 2 | 0.900 | 0.898 | 30.69 | 1263 | 491 | 501 |
| 3 | 0.918 | 0.916 | 11.4 | 1069 | 475 | 487 |
| 4 | 0.923 | 0.919 | 8.40 | 1047 | 472 | 486 |
| 5 | **0.930** | **0.925** | **3.06** | **966** | **466** | **483** |
| 6 | 0.930 | 0.924 | 4.19 | 973 | 467 | 486 |
| 7 | 0.930 | 0.924 | 6.05 | 992 | 469 | 491 |
| 8 | 0.930 | 0.923 | 8.01 | 1009 | 471 | 496 |
| 9 | 0.930 | 0.912 | 10.00 | 1032 | 473 | 500 |

**Table 1**: best value of criterion for each number of predictors(bold value is the best value).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Predictors | R2(for 5 predictors) | R2 adjusted | Mallow Cp | press | AIC | BIC |
| **albedo** | T | T | T | T | T | T |
| **CLDHGH** | F | F | F | F | F | F |
| **CLDMID** | F | F | F | F | F | F |
| **CLDLOW** | T | T | T | T | T | T |
| **SWGDN** | T | T | T | T | T | T |
| **PRECTOT** | F | F | F | F | F | F |
| **QLML** | T | T | T | T | T | T |
| **SPEED** | T | T | T | T | T | T |
| **PRECSNO** | T | T | T | T | T | T |

**Table 2:** Predictors chosen based on each criterion.

Looking at table 1, for R2, the reason I choose 5 predictors because from when adding predictor 6th to predictor 9th the R2 changes very insignificantly. In table 2, We see that all the criterion agrees that the model with albedo, CLDLOW, SWGDN, QLML, and PRECSNO is the best. The predictor 6th that can be added is SPEED, and the 5th predictor that is least significant among the 5 predictor is PRECSNO.

For this reason, I decided just to make sure I will check the cross validation MSE of 3 different models: model with 4 predictors(without PRECSNO), model with 5 predictors, and model with 6 predictors(with SPEED)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | 4 predictors | 5 predictors | 6 predictors |
| In sample MSE | 29.1 | 26.6 | 26.4 |
| Out of sample MSE | 38.2 | 35 | 35.5 |

**Table 3**: cross validation of models with 4,5 and 6 predictors

Looking at table 3, because the 5 predictors have the best out-of-sample MSE. We choose the 5 predictors to be out 5 predictors interested: albedo, CLDLOW, SWGDN, QLML, and PRECSNO. The model makes physical sense because we know albedo is negatively correlated with temperature, since if the earth reflects more sun radiation the temperature is more cool. Humidity is positive correlated with temperature because it affects how water takes energy and evaporates. Cloud coverage is negatively correlated with temperature because the more cloud the less sun radiation can get through. surface incoming shortwave flux is positively correlated because the more radiation the temperature increases. The precipitation might not be chosen because it tells the same story with the snow rate. High and mid clouds might not be chosen because it tells the same story with low cloud coverage. I think the effect on wind is just too small and to make a difference on the prediction, as we see the wind speed is negatively correlated with temperature with correlation of -0.02.

The coefficients of the model with 5 predictors without interactions are:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Intercept | Albedo | CLDLOW | SWGDN | QLML | PRECSNO |
| 17.50 | -22.58 | -6.05 | 0.028 | 5364 | 3.3 |

**Table 4**: Coefficients of model without interaction

Looking at the coefficients in table 4, This model indicates that when all the predictors are 0, the temperature is 17.5F. Assuming all the coefficients are independent, the coefficient of albedo means that when humidity is 0 the temperature will decrease 2.3F per 0.1 unit of albedo. Temperature will decrease 0.6F per 0.1 unit of fraction of cloud at surface pressure up to 700mb, increase 0.028F per 1 W/m2 of surface incoming flux, and increase 5.4F per 0.001 unit of humidity.

a potential trouble is PRECSNO because increasing snow rate will increase temperature does not make physical sense. Later on, I also find that when I add interactions to the model the PRECSNO will lose significance. Also a part of snow rate data correlated with albedo data. So I dropped it.

**Choosing interations:**

Running linear regression with all the interactions, I find that there are 3 significant interactions: albedo:QLML, CLDLOW:SWGDN:QLML, and CLDLOW:QLML.

When I check how much the model improves by adding all the different interactions I mentioned above, I learn that adding albedo:QLML is the most significance in terms of criterion and out-of-sample MSE. And adding CLDLOW:SWGDN:QLML is the second most significance in terms of criterion and out of sample MSE. I find that adding interaction albedo:QLML increases the performance of the model significantly. I also find that dropping PRECSNO improves the performance of the model. I decided to not use the interaction CLDLOW:SWGDN:QLML because I cannot interpret the physical meaning of the interaction, and even though the performance of the model increases, the increase is not that significance.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 5 predictors with albedo:QLML | 4 predictors (without PRECSNO) with albedo:QLML | 4 predictors with: albedo:QLML  And CLDLOW:SWGDN:QLML |
| R2 | 0.948 | 0.947 | 0.953 |
| R2adj | 0.944 | 0.944 | 0.949 |
| Press | 723 | 724 | 670 |
| AIC | 441 | 440 | 432 |
| BIC | 461 | 457 | 452 |
| Out of sample MSE | 26.9 | 26.3 | 24.2 |

**Table 5**: criterions value and out of sample MSE of different models with interactions

So the final model that I choose has 4 predictors: Albedo, CLDLOW, SWGDN, and QLML with interaction albedo:QLML.

**Analysis of the model:**

The coefficients are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Intercept | Albedo | CLDLOW | SWGDN | QLML | Albedo:QLML |
| Coefficients | 2.44 | -50.73 | -6 | 0.0222 | 2100 | 18400 |
| P value | 0 | 1.0e-11 | 7.1e-05 | 0.012 | 0.00038 | 2.0e-08 |

**Table 6**: coefficients value and p value of hypothesis testing with null Bi=0

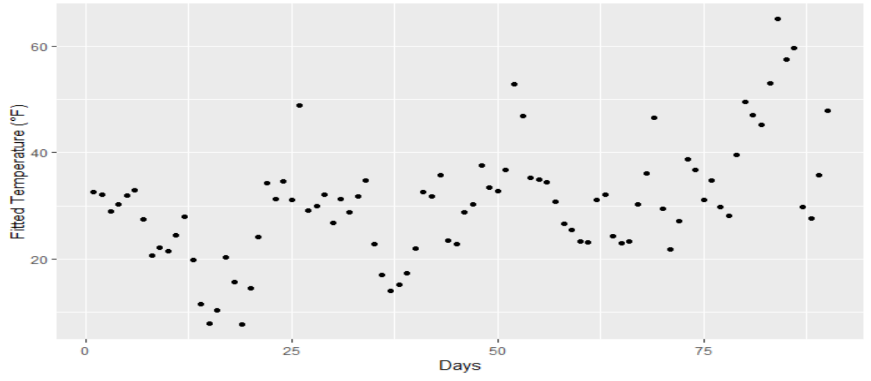
|  |  |  |
| --- | --- | --- |
| MSE(F^2) | F-statistics | R2 |
| 6.15 | 118 | 0.947 |

**Table 7**: MSE value and F statistics of the global F test

Looking at table 6, the intercept means that according to our model, when all the predictors are 0, the temperature at South Bend is 2.44F. Assuming all the coefficients are independent aside from Albedo and humidity. The coefficient of albedo means that when humidity is 0 the temperature will decrease 0.5F per 0.1 unit of albedo. Temperature will decrease 0.6F per 0.1 unit of fraction of cloud at surface pressure up to 700mb, increase 0.0222F per 1 W/m2 of surface incoming flux. And when albedo is 0, the temperature will increase 21F per 0.01 unit of humidity.

The p values for testing with null hypothesis of Bi = 0 indicate that all the coefficients are significant. From table 7, we can see that the R squared is 0.947, which means that our model performs 95% better than a model with no slope. The mean squared error is 9.86F^2, and that the standard deviation is 3.14F, which means with 95% confidence the temperature observed is varied within 6.15F. The p value of global F test is 0 at alpha level of 0.05, So we reject the null and conclude that our model provides a better fit than pure noise.

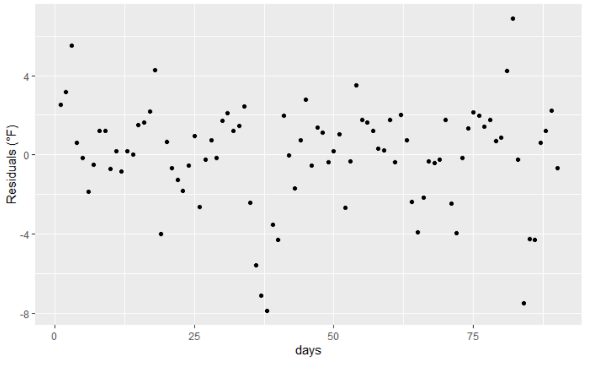
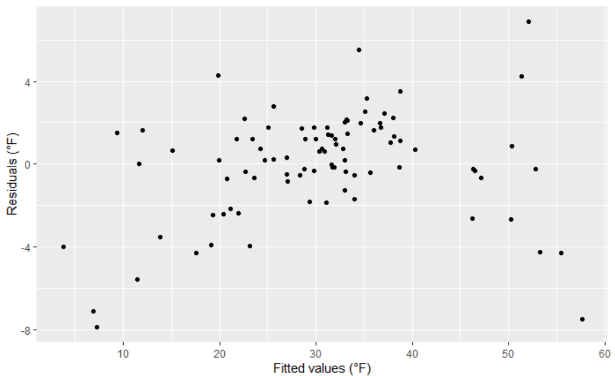
Looking at Figure 2, we can see that our predict values, we can see that the plot of fitted value really closely resembles the original plot, indicate that the linear model works well.

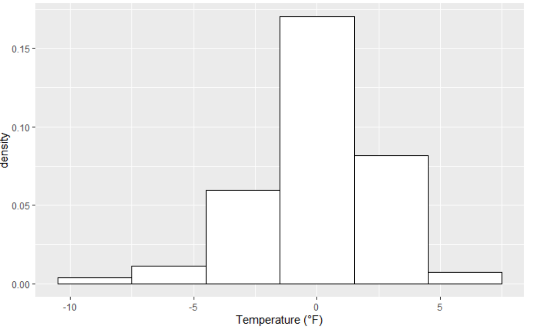


**Figure 2**: Fitted value vs days

**Residuals and assumptions:**

I find that the residual versus humidity indicates the error has non constant variance. The residual versus albedo indicates error has non-constant variance. However, this makes sense why we include an interacting term so there is no problem. The residual versus CLDLOW and SWGDN indicate no problem. So I put the 4 previous graphs in the extra material. The residual versus fitted value indicates that the error has non constant variance as well, but the effect is not big (Figure 3). Looking at the residual versus time, we see that there is no trend, so we can say that the errors are likely to be independent. From figure 3 graph 3, The error is also roughly normal, but it is skewed to the right.





**Figure 3:** Residuals vs fitted value, days. And the Density of residuals.

So the model that we have satisfies that the relationship between predicted variable and predictors are linear. However, the errors have non constant variance. There might be some existence of outlying observations, as I mentioned in the EDA. Errors are independent. And errors are not normal.

**Conclusion**

**So the final model is:**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Intercept | Albedo | CLDLOW | SWGDN | QLML | Albedo:QLML |
| Coefficients | 2.44 | -50.73 | -6 | 0.0222 | 2100 | 18400 |

So according to our model, the temperature of south bend in the winter decreases 0.5F per 0.01 unit of albedo when humidity is 0, decrease 0.6F per 0.1 unit of fraction of cloud at surface pressure up to 700mb, increase 0.0222F per 1 W/m2 of surface incoming flux, and increase 2.1F per 0.001 unit of humidity when albedo is 0. Using this model, one can predict the temperature of South Bend in the winter in the coming year.