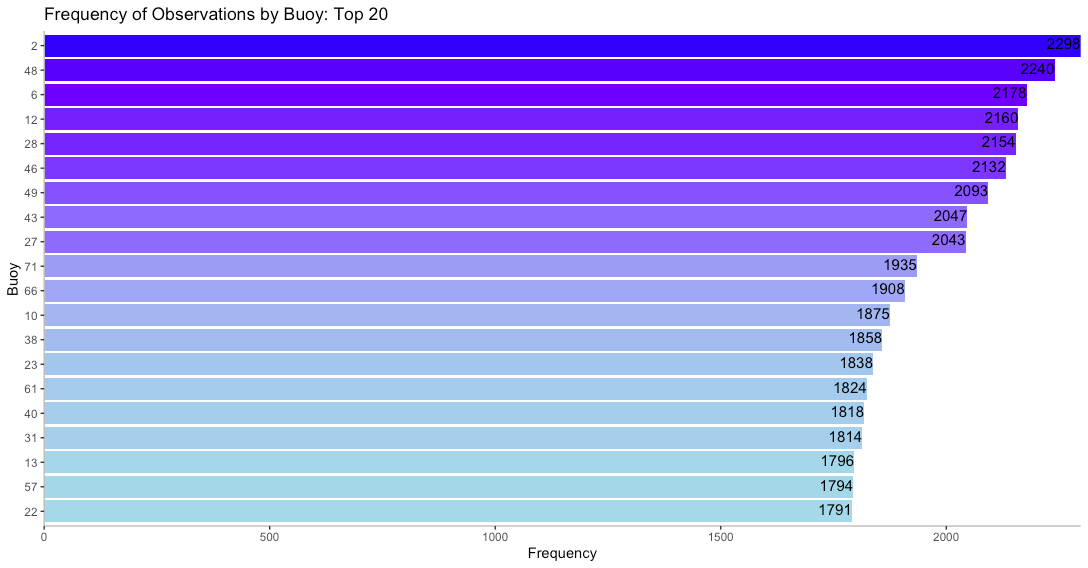
A2: Regression -- Invincibles

# Part 1: Descriptive statistics

The “El Nino” data we analyzed in Assignment 1 was measured by 72 buoys distributed in significantly different time periods. The different properties of each buoy would introduce omitted variable bias if we would regress all of them. As a result of this, we chose only one buoy to conduct the regression analysis. We use buoy 2 for our analysis since it contains the most observations (2298 entries) as shown in the barplot below:

The cleaned data looks like the table below:

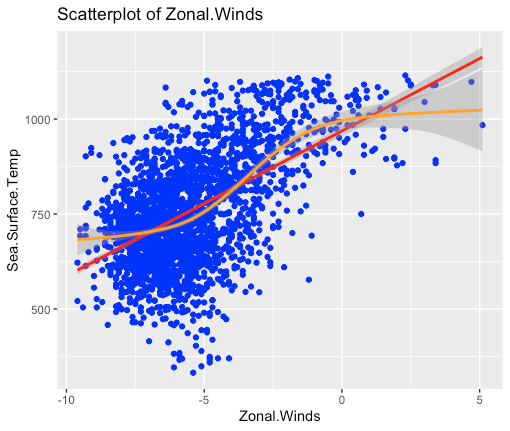
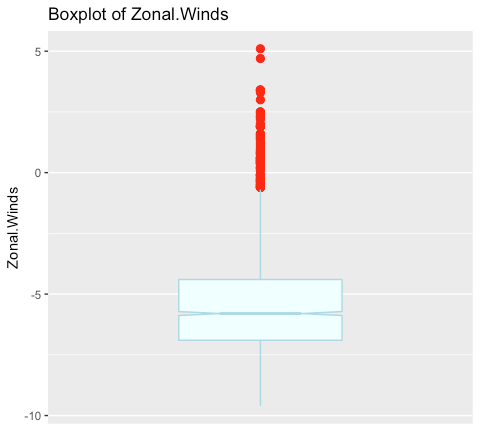
|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Observation** | **New.Date** | **Latitude** | **Longitude** | **Zonal.Winds** | **Meridional.Winds** | **Humidity** | **Air.Temp** | **Sea.Surface.Temp** |
| **1** | 12401 | 19900501 | 0.02 | -139.93 | -5.9 | 2.1 | 87.1 | 27.21 | 851 |
| **2** | 12402 | 19900502 | 0.01 | -139.92 | -4.4 | -1.6 | 87.9 | 27.12 | 831 |
| **3** | 12403 | 19900503 | 0.02 | -139.92 | -5.8 | 1.1 | 86.5 | 27.56 | 830 |
| **4** | 12404 | 19900504 | 0.01 | -139.92 | -4.1 | 0.8 | 86.3 | 27.32 | 832 |
| **5** | 12405 | 19900505 | 0.01 | -139.92 | -3.3 | -0.2 | 84.1 | 27.23 | 844 |

We are specifically interested in relevant and important variables, including Zonal.Winds, Meridional Winds, Humidity, Air Temp and Sea.Surface.Temp. The target variable is Sea.Surface.Temp, because this El Nino data set studies the El Nino effect in tropical pacific, and Sea.Surface.Temp indicates the change of the El Nino effect. As a result, we want to investigate the association between Sea.Surface.Temp and the rest of the variables, so that we can know whether the rest of the variables showcase any significant sign of the El Nino effect. If those variables are not statistically significantly associated with Sea.Surface.Temp, in future, we can remove those measures from the buoys.

The predictor variables are Zonal.Winds, Meridional Winds, Humidity and Air Temp.

Here are the plots and descriptive statistics below. In the box-and-whisker plots, the outliers are plotted as red dots. In the scatterplots, both linear(red) and curvilinear(orange) regression are applied to the plots.

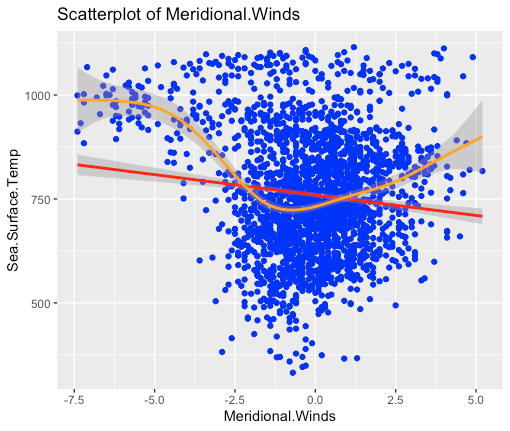
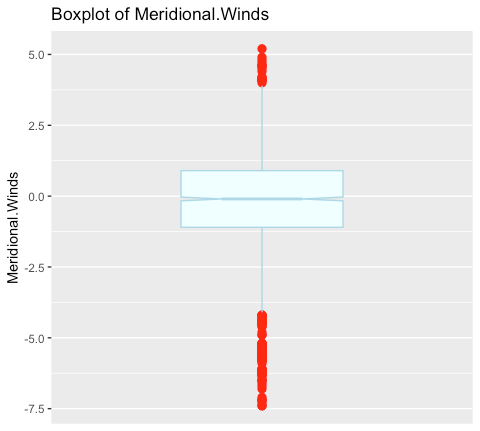
### Descriptive statistics of Zonal.Winds:



min max median mean var std.dev mode

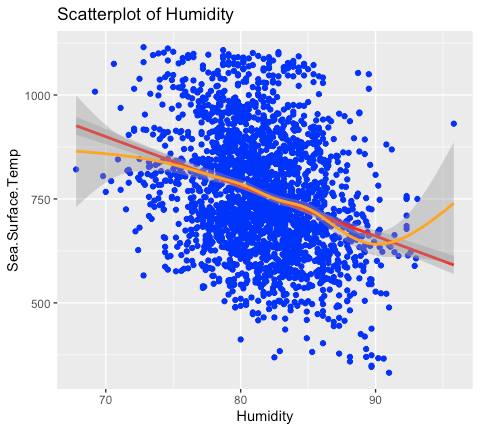
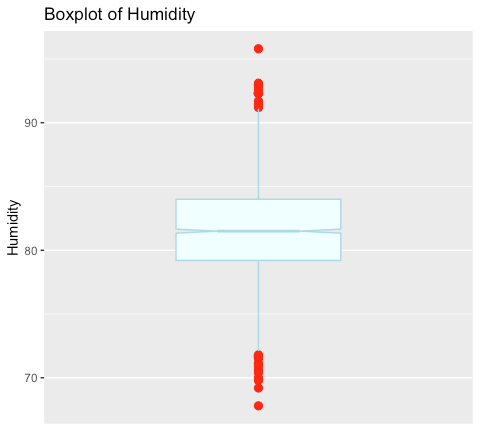
-9.60 5.10 -5.80 -5.44 4.33 2.08 -6.8

### Descriptive statistics of Meridional.Winds:

min max median mean var std.dev mode

-7.40 5.20 -0.10 -0.17 3.28 1.81 0.4

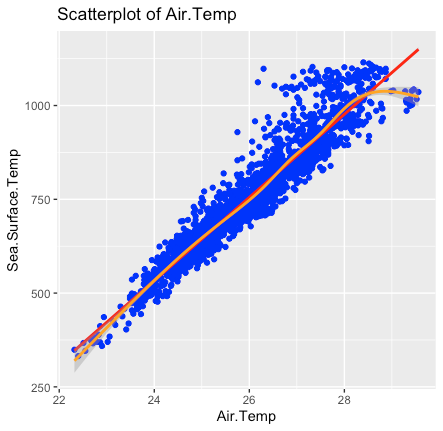
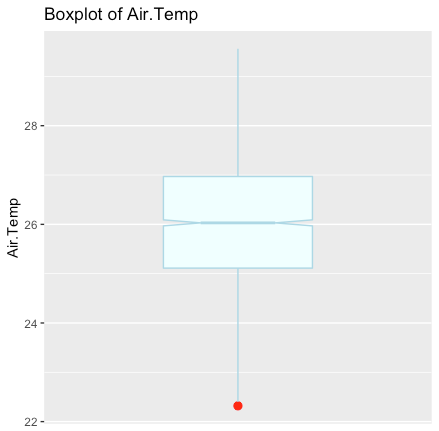
### Descriptive statistics of Humidity:



min max median mean var std.dev mode

67.80 95.80 81.50 81.61 14.48 3.81 79.8

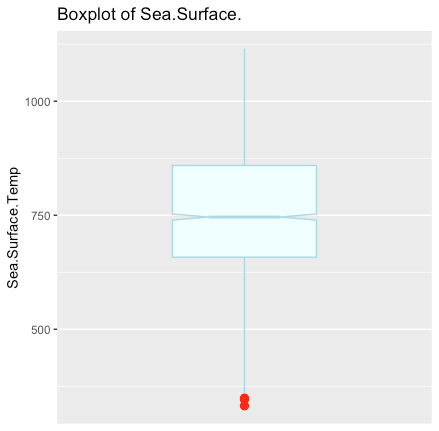
### Descriptive statistics of Air.Temp:



min max median mean var std.dev mode

22.32 29.56 26.03 26.06 1.58 1.26 25.96

### Descriptive statistics of Sea.Surface.Temp:



Because Sea.Surface.Temp is the target variable, a scatterplot cannot be applied to this variable.

min max median mean var std.dev mode

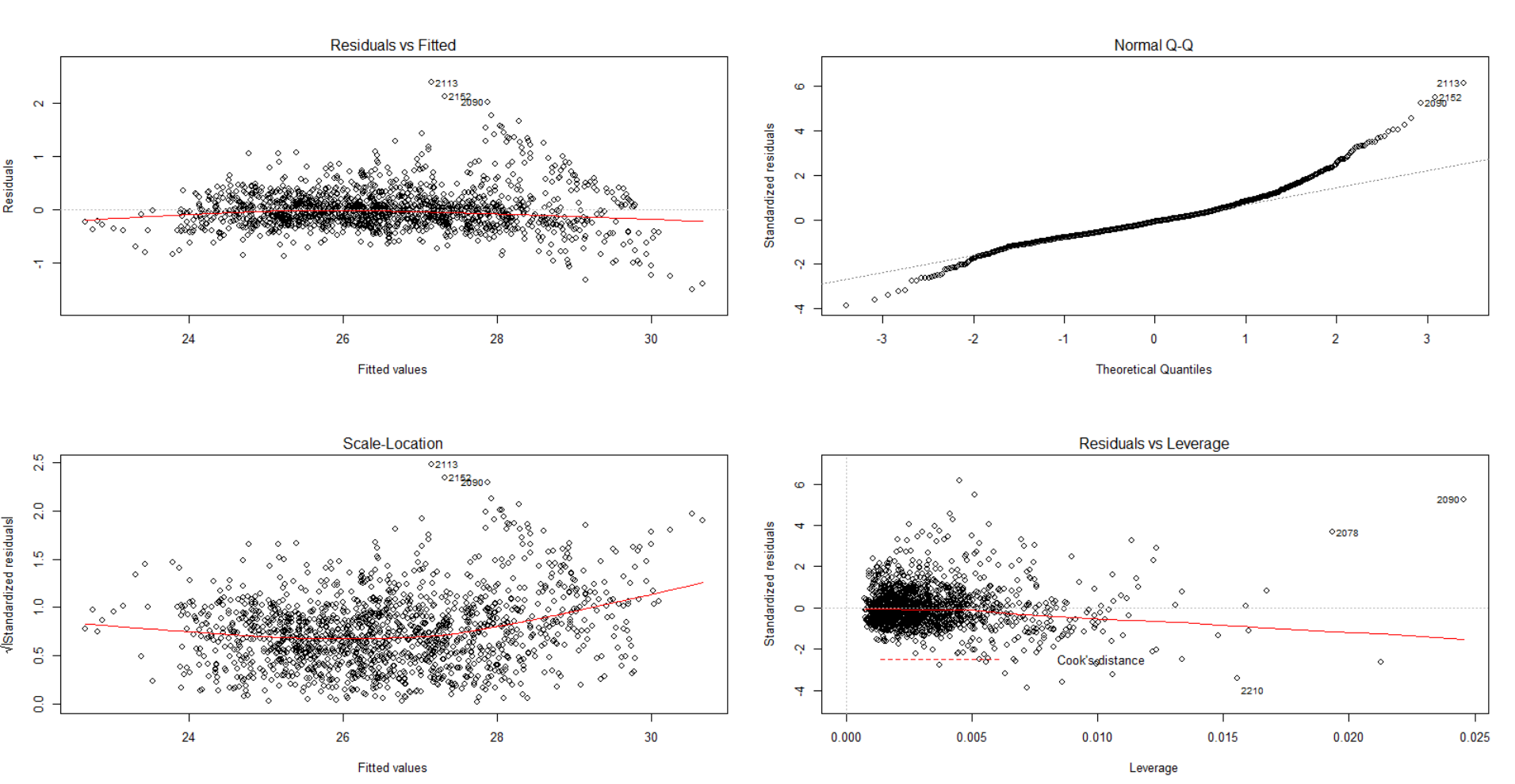
332.00 1115.00 746.00 761.09 21557.46 146.82 713

# Part 2: Multiple Linear Regression

**Simple Linear Regression:**

We randomly parted the training and validation sets. The training set originally included 1500 data points and the validation set included 798 data points. All four variables' p-values are close to 0. They can be regressed as predictors because they are statistically significant. Thereby, we regressed “Sea.Surface.Temp” on four independent variables: “Zonal.Winds”, “Meridional.Winds”, “Humidity”, and “Air.Temp.” By plotting the residuals, we discovered numerous residuals and eventually removed 87 data from the training set.

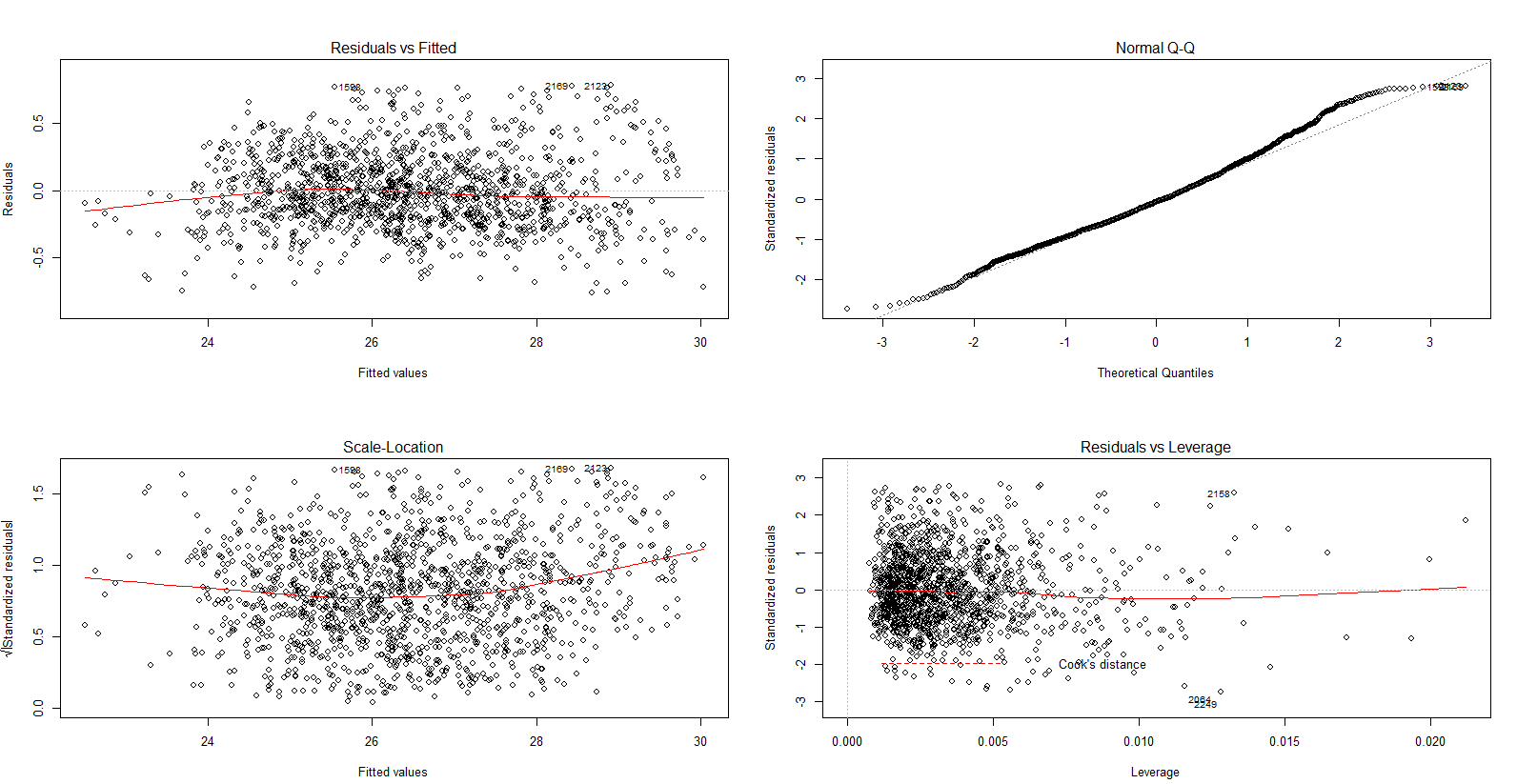
### Residual Plotting (Original)



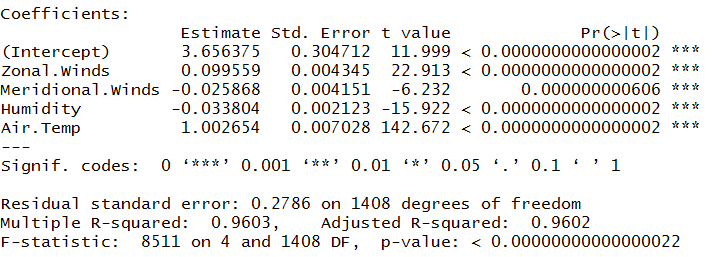
### VIF (Original): all smaller than 5



### Residual Plotting (After Removing Outliers)



### Summary of Multiple Regression (After Removing Outliers)



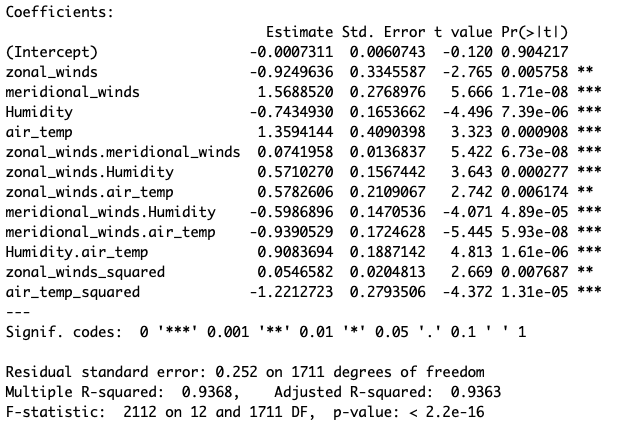
Best Model Statistics (After Removing Outliers)

Adjusted R-squared: 0.9602

RMSE: 0.4092914

**Linear Regression With Interaction And Squared Terms:**

The above model is the simplest model that we explored. We proceeded to explore a more complex model with the addition of squared and interaction terms for each of the variables. We then used the step AIC algorithm with “backward” selection in order to find the most relevant variables. With “backward”, we start with all the variables and then the variable that gives the minimum AIC when dropped, is dropped for the next iteration until there is no significant drop in AIC is noticed. Note that the data was standardized before carrying out this regression. Below are the results from this model:



Model Statistics

Adjusted R-squared: 0.9363

RMSE: 0.2673344

The above model achieves lower adjusted R squared as compared to the simpler model. All the variables are also statistically significant at 10 percent statistical significance as a result of applying the AIC criterion. However, this outcome is not a guarantee that this is the most parsimonious model. We proceeded to check for multicollinearity among the variables by calculating the Variance Inflation Factors(VIFs).

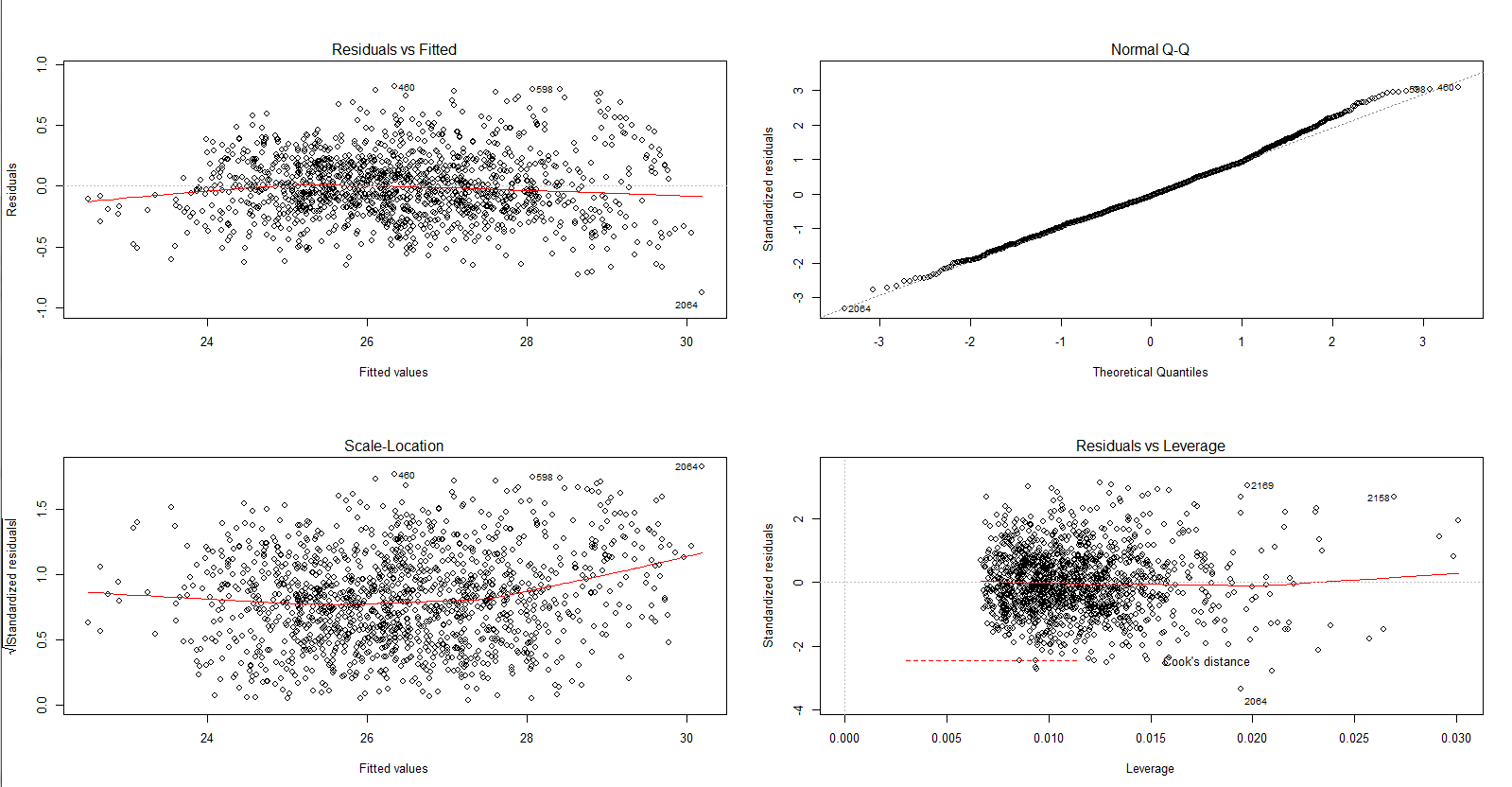
Below is table that shows the VIFs for the variables in the model above:

|  |  |
| --- | --- |
| Variable | VIF |
| meridional\_winds\_squared | 1.651 |
| zonal\_winds.meridional\_winds | 4.869 |
| zonal\_winds\_squared | 7.680 |
| zonal\_winds.Humidity | 7.799 |
| Humidity | 538.930 |
| meridional\_winds.Humidity | 623.315 |
| Humidity.air\_temp | 797.862 |
| meridional\_winds.air\_temp | 827.671 |
| air\_temp\_squared | 1700.301 |
| meridional\_winds | 2118.720 |
| air\_temp | 3295.693 |

The table above suggests that we should only include four predictors in the model whose VIFs are below 10. These are *meridional\_winds\_squared* , *zonal\_winds:meridional\_winds, ,zonal\_winds\_squared* and *zonal\_winds.Humidity*. High values of the VIF by themselves do not discount the results of regression analyses nor call for the elimination of one or more independent variables from the analysis. Rather they suggest the need to use alternative regression approaches or require combining independent variables into a single index.[[1]](#footnote-0) With this weather data, we know that weather statistics of the previous days will affect today's weather. In the next section of this report, we explore seasonality in the data and apply autoregression techniques.

**Adding Dummies to the Model**

Although the data was not collected evenly every year, there might be seasonality in the series. We added monthly dummy variables in our dataset and ran the regression with dummies. We removed the same 73 observations from the training set. The removed data accounts for only 5% of the total training data. It’s reasonable because we have not neglected too much data and we have still captured most of the information.

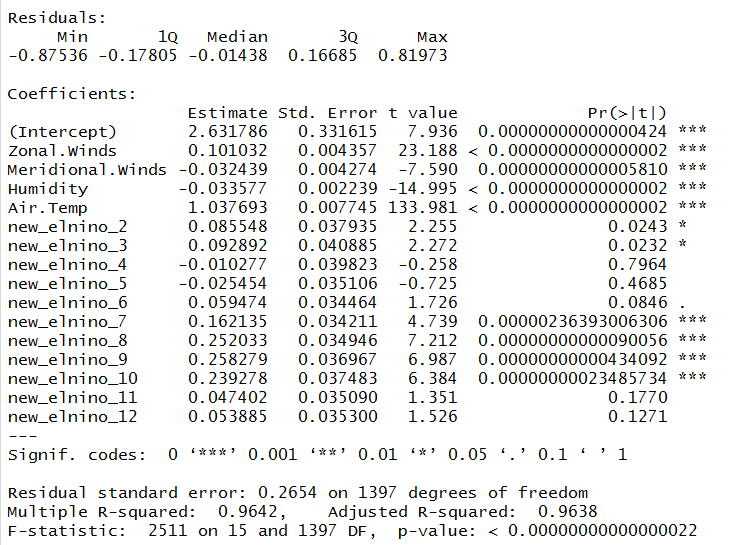


Dummies Model Statistics

Adjusted R-squared: 0.9638

RMSE: 0.39707

The main four variables’ coefficients differ from the best model without dummy variables, but mostly they are aligned with the previous best model’s coefficients. We removed the “January dummy” to avoid the dummy variable trap. The coefficients of dummies in April and May have negative values. These months may be relatively cooler in a year. Also, it seems that the sea surface temperature was still quite low just before the summer. This may be caused by the water temperature lag compared to the continent temperature. In these months, the value of dependent variable, sea surface temperature, should be reduced. August, September and October were the months with the highest temperature. The dummy coefficients in these months have the highest positive values. Note that the buoy traveled around the equatorial area. The temperature seasonal variance should be significantly lower than other places on the earth. Below are the results for this model:



**Dummies, Interaction and Squared Terms Regression:**

In the end, we decided to wrap all variables in the same model, and the backward selection suggests the following variables for regression. Their respective VIFs are shown as well.

## 

However, as the last line (RMSE=0.2476) in the above table suggests, this model actually predicts the test data well, despite the fact that it fits the training data well. The R squared on training data is 0.94, similar to models run above. The following table shows the result of the regression.

## 

# Model interpretations and reflections

Results from the multiple regression suggests that we need a simpler model for this data.

We have run four regressions above, including simple linear regression, regression with squared and interaction terms, simple linear regression with monthly dummy variable and regression with all variables. All four regressions have similar R-squared values above 0.9, however, regression with squared and interaction terms and regression with all variables have significantly lower RMSE values, i.e. 0.26 and 0.24 respectively. It makes sense, since the temperature varies from time to time in a year and there are some interactive effects among different variables, e.g. zonal winds and humidity as the above table suggests.

# 

1. <https://link.springer.com/article/10.1007/s11135-006-9018-6> [↑](#footnote-ref-0)