Analysis of Rain Behaviour in Australia

Statistical Learning's Project

Candidates: Luca Dal Zotto - Francesco Ferretto

Department of Mathematics "Tullio Levi-Civita"
Master Degree in Data Science

10th July 2020

Università degli Studi di Padova

Contents

Introduction to the task

Obtaining Data

- The `weatherAUS` dataset contains over 140,000 daily weather observations from 49 Australian weather stations collected in about 10 years. All the information has been collected by the Australian Bureau Of Meteorology.

Figure: *Copyright Commonwealth of Australia 2010, Bureau of Meteorology.*

- The specific dataset used in this project has been downloaded from Kaggle at the link: https://www.kaggle.com/jsphyg/weather-dataset-rattle-package.
- It is also available in a slightly different version via the R package `rattle.data` and at the link: https://rattle.togaware.com/weatherAUS.csv.

The dataset in organized in *blocks*, where each block contains the recordings of a given weather station. These blocks are ordered by the location of the station, and each block is in chronological order.
Each row has some climatic data regarding a specific day and location.
To be precise, two variables regards the rainfall of the following day: they will be the response variable (one for regression, one for classification).
All other variables are relative to that day.

- Dimensions: 142193 observations and 24 variables;
- No duplicate observation;
- 56420 observations with at least one NA value.

Four variables (Evaporation, Sunshine, Cloud9am, Cloud3pm) has a high percentage of NAs (between 35 and 50%). The remaining variables have less than 10% of NAs.

Introduction to the task

Explanatory Variables, $\mathbb{X}$

- **Date**: The date of observation (a date object).
- **Location**: The common name of the location of the weather station
- **MinTemp**: The minimum temperature in degrees centigrade
- **MaxTemp**: The maximum temperature in degrees centigrade
- **Temp9am**: Temperature (degrees C) at 9 a.m.
- **Temp3pm**: Temperature (degrees C) at 3 p.m.
- **Evaporation**: Class A pan evaporation (in millimeters) during 24 h
- **Sunshine**: The number of hours of bright sunshine in the day
- **WindGustDir**: The direction of the strongest wind gust in the 24 h to midnight
- **WindGustSpeed**: The speed (in kilometers per hour) of the strongest wind gust in the 24 h to midnight
- **WindDir9am**: The direction of the wind gust at 9 a.m.
- **Rainfall**: The amount of rainfall recorded for the day in millimeters.
- **WindDir3pm**: The direction of the wind gust at 3 p.m
- **WindSpeed9am**: Wind speed (in kilometers per hour) averaged over 10 min before 9 a.m.
- **WindSpeed3pm**: Wind speed (in kilometers per hour) averaged over 10 min before 3 p.m.
- **Humidity9am**: Relative humidity (in percent) at 9 am
- **Humidity3pm**: Relative humidity (in percent) at 3 pm
- **Pressure9am**: Atmospheric pressure (hpa) reduced to mean sea level at 9 a.m.
- **Pressure3pm**: Atmospheric pressure (hpa) reduced to mean sea level at 3 p.m.
- **Cloud9am**: Fraction of sky obscured by cloud at 9 a.m. This is measured in "oktas," which are a unit of eighths.
- **Cloud3pm**: Fraction of sky obscured by cloud at 3 p.m.

Introduction to the task

Response(s), $\mathbb{Y}$

Rainfall variables

- **Rainfall**: amount of rainfall recorded for the day in mm;
- **RainToday**: "Yes" if precipitation in the 24 hours to 9am exceeds 1mm, otherwise "No";
- **RISK_MM**: same as Rainfall, shifted by one day (in the future);
- **RainTomorrow**: same as RainToday, shifted by one day.

- Check correctness of these variables
- Restore some missing values.

Missing data are also referred to as *non-response* or *unobserved* data, and occur in most types of studies. Missing values can occur due to:

- Failure of measurement
- Data loss
- Out-of-range data and data loading issues
- Units fail to answer all questions
- Loss of follow-up or other plausible reasons.

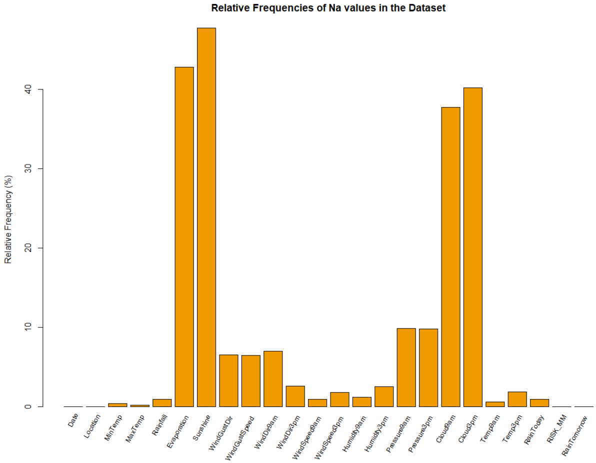
"We should be suspicious of any dataset (large or small) which appears perfect."

David J. Hand

Introduction to the task

Missing Values' Pattern

UNIVERSITÀ DEGLI STUDI DI PADOVA

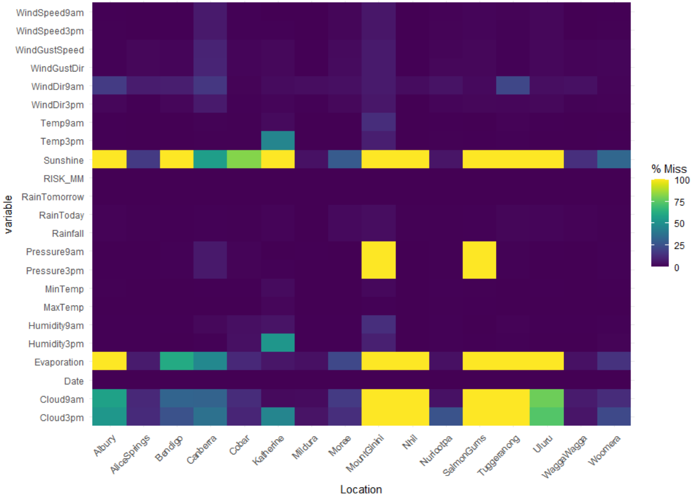


The subset of variables ***S*** $=\{$*Evaporation*, *Sunshine*, *Cloud9am*, *Cloud3pm*$\}$ have approximately 40% of the NAs.

But, first of all, how are these values missing? And *why*?

Introduction to the task

Missing Values' Pattern



This leads to another fact, that the total absence of observations seems to be related to the absence of the sensors needed to acquire the measurements rather than a missingness due to the influence of external factor(s), or our predictors. So, we might not have a pattern of missings related to other variables in the overall analysis.

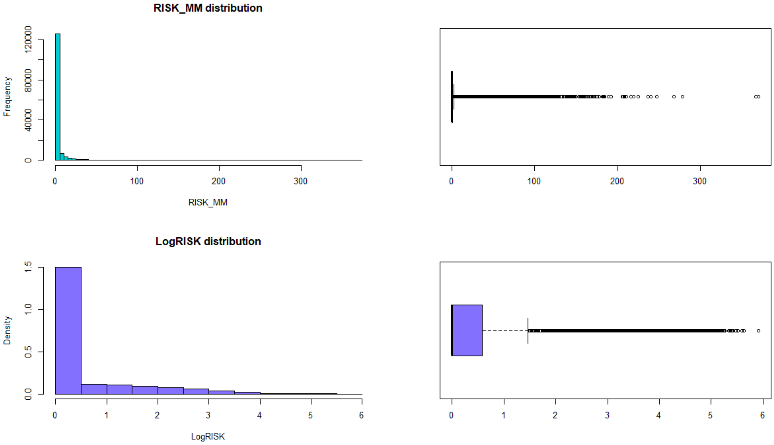
Our main strategy is to *remove the Locations with total absence of measurements wrt the variables with the highest portion of NAs*. Another possible strategy should consist in removing the predictors, but it may have bad consequences since we're deleting degrees of freedom instead of a portion of observations. Imputation in our case would be quite *ineffective*, especially with mean-median imputation of the missing data. So, our aim is to choose the best strategy without altering the structure of the dataset, in particular, relationships among explanatory variables and the response.

Data exploration

Rainfall variables

Università degli Studi di Padova

In this section we are going to analyze the distribution of the variables and observe their main features, starting from the response RISK_MM.

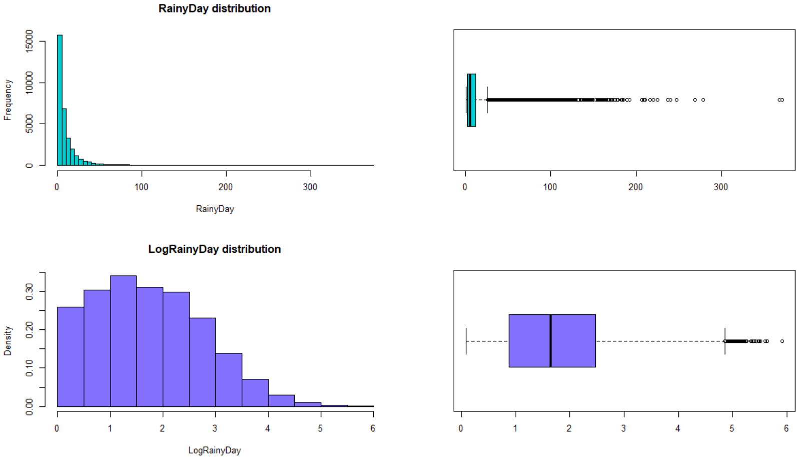


Extremely right skewed $\rightarrow$ $LogRISK = log(RISK\_MM + 1)$

Data exploration

Rainfall variables

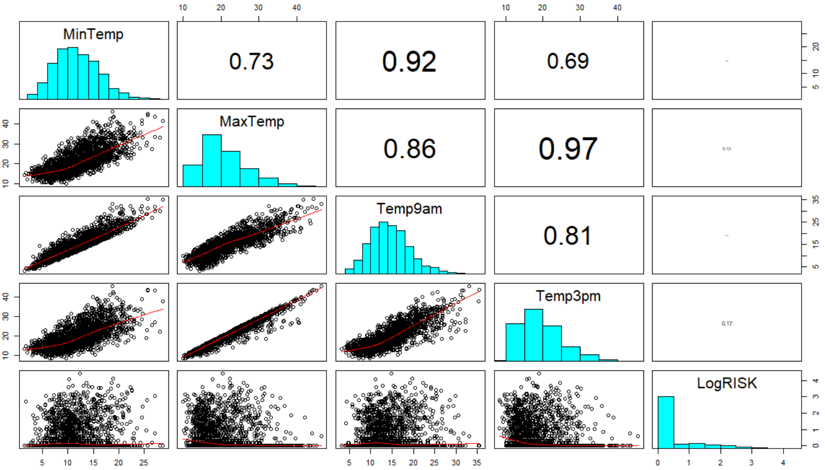
Class Frequency: "*No*" = 110314, "*Yes*" = 31877 → focus on rainy day



LogRainyDay is approx. a truncated normal

Data exploration

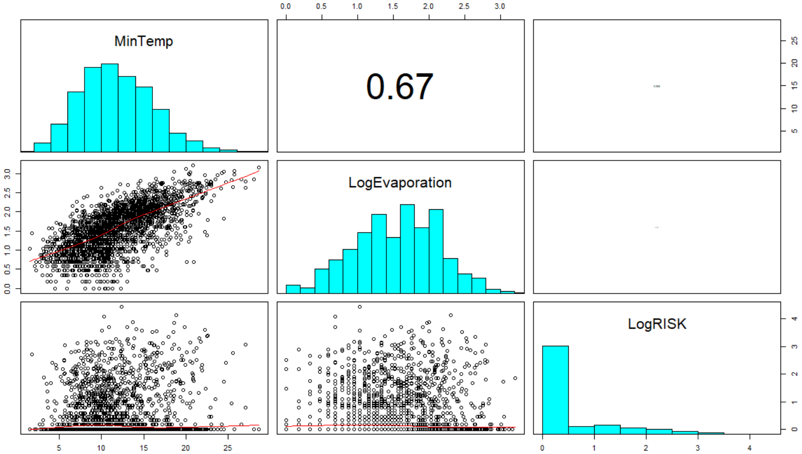
Temperature variables



Approx. bell-shaped. Highly correlated → keep MinTemp

Data exploration

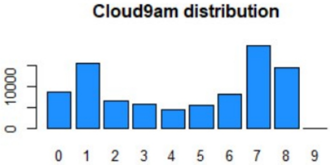
Evaporation

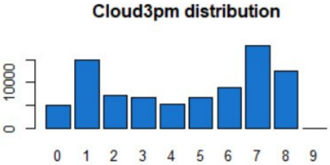


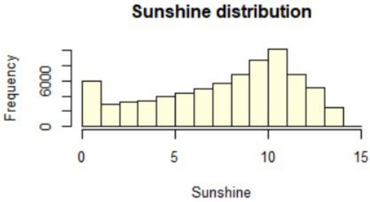
Right skewed and with lots of NAs. Correlated with Temperature variables

Data exploration

Cloud and Sunshine





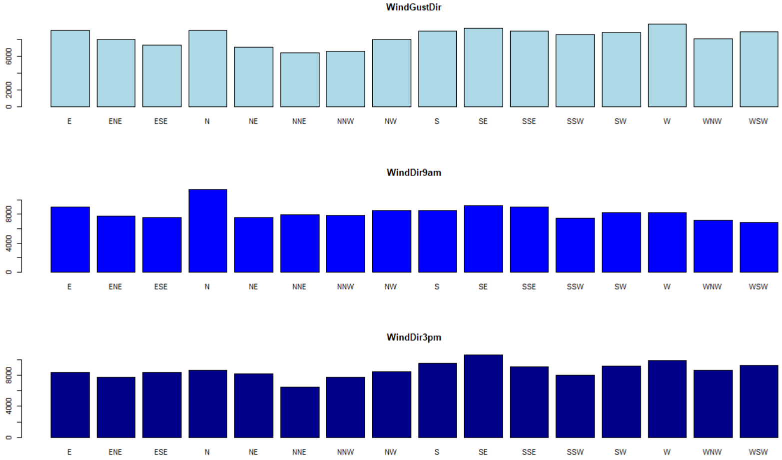


	Cloud9am	Cloud3pm	Sunshine	LogRISK
Cloud9am	1.00	0.41	-0.61	0.22
Cloud3pm	0.41	1.00	-0.69	0.36
Sunshine	-0.61	-0.69	1.00	-0.40
LogRISK	0.22	0.36	-0.40	1.00

Cloud9am and Cloud3pm are discrete. High correlation with Sunshine.

Data exploration

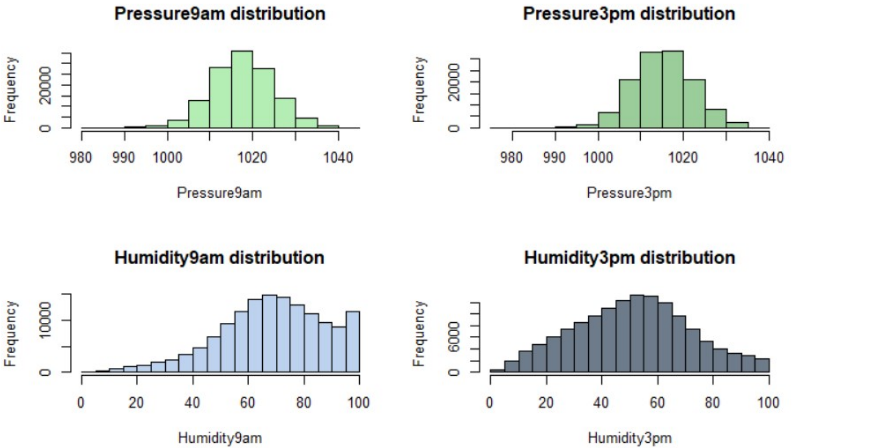
Wind variables



Categorical variables → from compass directions to Cartesian coordinates

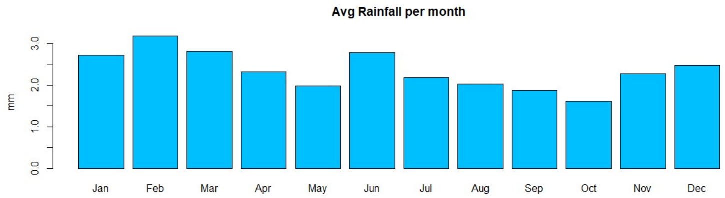
Data exploration

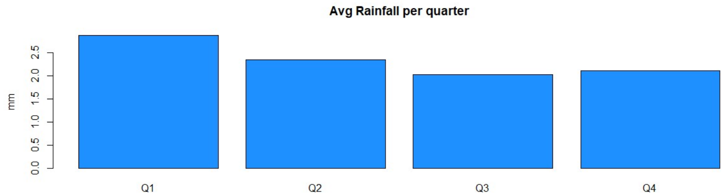
Pressure and Humidity



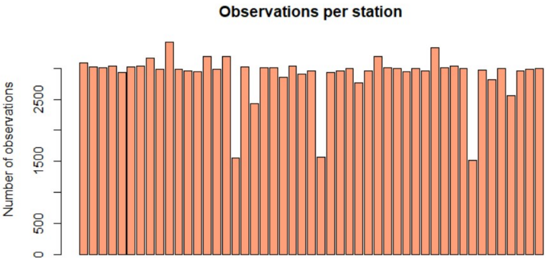
Range of pressure $[977; 1041]$. Humidity not much symmetric: pick in 100%. High correlation (0.53) between Humidity9am/3pm but might be both relevant. Very high correlation (0.96) between Pressure9am/3pm.

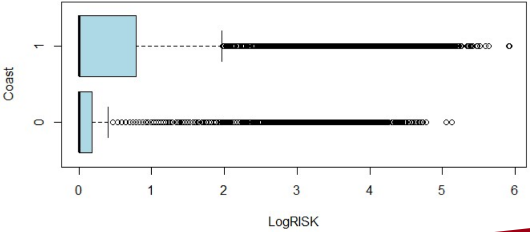
Date of observation, in the format Year-Month-Day → extract the features month and quarter





Università degli Studi di Padova





- 49 different weather station;
- Most of them provided about 3000 days of meteorological recordings, but 3 of them (Katherine, Nhil and Uluru) provided approx one half;
- the weather recordings for these 3 stations start from March 2013, while the other stations provided data from February 2008;
- use this variable to create by hand the binary variable Coast.

Problems

1. **Time Dependency**: the dataset basically is a sequence of time series, so the observations could not be *i.i.d.*;
2. **NAs**: some variables have a high percentage of missing values: certain locations have no recording for some features;
3. Tricky distribution shape of the response variable.

Possible counter-measures

1. Select one observation every 3 days;
2. Remove the locations with at least one variable of all NAs or impute NAs, e.g. with unsupervised methods (*k-NN*) using nearby stations; we might also consider a model based on only one weather station;
3. Consider a model regarding only rainy days.

UNIVERSITÀ DEGLI STUDI DI PADOVA

Handcrafted Analysis

Here we face an handcrafted analysis starting from the full model $\mathcal{M}_{23}$, potentially, we expect to encounter **multicollinearity** among the predictors, from the preliminary phase of EDA. Here we've confirmed our belief, using a statistic called *VIF* - **Inflation Variance Factor**. From the literature, in particular with a relevant number of observations, an high VIF can lie in a range between 5 and 10, according to *Hastie et Al.* ($*^1$)

	GVIF	Df	GVIF^(1/(2*Df))
MaxTemp	47.057766	1	6.859866
MinTemp	10.594707	1	3.254951
Temp3pm	55.884768	1	7.475612
Temp9am	24.198561	1	4.919203
Pressure3pm	25.280719	1	5.027994
Pressure9am	25.627725	1	5.062383
Humidity3pm	7.044856	1	2.654215

Characteristics of the full model $\mathcal{M}_{23}$

```
Call:
lm(formula = LogRISK ~ ., data = WeatherV11, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4293 -0.3913 -0.0846  0.2031  4.2671

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    24.6322468  0.5441093  45.271  < 2e-16 ***
MaxTemp         0.0422697  0.0028805  14.674  < 2e-16 ***
MinTemp        -0.0101193  0.0014848  -6.815 9.50e-12 ***
Temp3pm        -0.0218684  0.0032007  -6.832 8.44e-12 ***
Temp9am        -0.0136446  0.0021922  -6.224 4.87e-10 ***
LogEvaporation  0.1205677  0.0086129  13.998  < 2e-16 ***
Sunshine       -0.0723552  0.0014536 -49.776  < 2e-16 ***
Wind3pmX       -0.0014072  0.0002179  -6.458 1.07e-10 ***
Wind3pmY       -0.0006827  0.0002637  -2.589  0.00964 **
Wind9amX        0.0016736  0.0002565   6.526 6.83e-11 ***
Wind9amY        0.0033893  0.0002869  11.813  < 2e-16 ***
WindGustX      -0.0001506  0.0001135  -1.326  0.18480
WindGustY      -0.0011459  0.0001303  -8.792  < 2e-16 ***
Humidity3pm     0.0191808  0.0003847  49.859  < 2e-16 ***
Humidity9am    -0.0038992  0.0003216 -12.123  < 2e-16 ***
Pressure3pm    -0.0771962  0.0021417 -36.045  < 2e-16 ***
Pressure9am     0.0527827  0.0021445  24.613  < 2e-16 ***
Cloud3pm        0.0106599  0.0016736   6.369 1.91e-10 ***
Cloud9am       -0.0200855  0.0015936 -12.604  < 2e-16 ***
quarters2      -0.0325618  0.0100942  -3.226  0.00126 **
quarters3      -0.0141130  0.0106589  -1.324  0.18549
quarters4       0.0159257  0.0087930   1.811  0.07012 .
Coast1         -0.0681967  0.0076693  -8.892  < 2e-16 ***
LogRainfall     0.2091049  0.0039697  52.675  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6955 on 56442 degrees of freedom
  (30850 observations deleted due to missingness)
Multiple R-squared:  0.4027,    Adjusted R-squared:  0.4025
F-statistic:  1655 on 23 and 56442 DF,  p-value: < 2.2e-16
```

- **BIC** of the model: 119485
- **AIC** of the model: 119261.5
- R^2 of the model: 0.4024949

Reduced Model

Model with a subset the explanatory variables $\mathcal{M}_{10,BSS}$ – WeatherV11

Characteristics of the reduced model $\mathcal{M}_{10,BSS}$

```
Call:
lm(formula = LogRISK ~ . - MinTemp - Temp9am - quarters - Temp3pm -
    Wind3pmX - Wind3pmY - Wind9amX - Wind9amY - WindGustX - Pressure3pm -
    WindGustY, data = WeatherV11, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3128 -0.4092 -0.0891  0.2034  4.2274

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.7650934  0.5327769  38.975  < 2e-16 ***
MaxTemp       0.0126529  0.0006517  19.415  < 2e-16 ***
LogEvaporation 0.0434223 0.0080193   5.415 6.16e-08 ***
Sunshine     -0.0763245  0.0014242 -53.590  < 2e-16 ***
Humidity3pm   0.0154717  0.0002565  60.312  < 2e-16 ***
Humidity9am  -0.0027297  0.0002568 -10.629  < 2e-16 ***
Pressure9am  -0.0202503  0.0005156 -39.278  < 2e-16 ***
Cloud3pm      0.0130019  0.0016845   7.718 1.20e-14 ***
Cloud9am     -0.0284579  0.0015703 -18.122  < 2e-16 ***
Coast1       -0.1130630  0.0076763 -14.729  < 2e-16 ***
LogRainfall   0.1726336  0.0039466  43.742  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7117 on 56455 degrees of freedom
  (30850 observations deleted due to missingness)
Multiple R-squared:  0.3745,    Adjusted R-squared:  0.3744
F-statistic:  3380 on 10 and 56455 DF,  p-value: < 2.2e-16
```

The Subset Selection was performed manually and through BSS. With respect to the full model $\mathcal{M}_{23}$, both procedures led to simpler models, preserving the fitting. The BSS model ($\mathcal{M}_{11,BSS}$ updated to $\mathcal{M}_{10,BSS}$, VIF correction) was the least complex ($\mathcal{M}_{18,M}$).

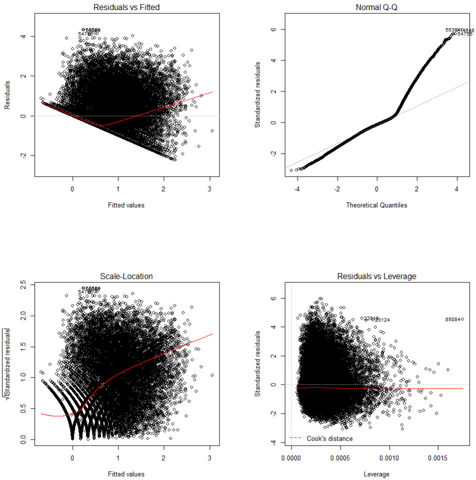
- **BIC** of $\mathcal{M}_{10,BSS}$: 121953.7 (120691.4)
- **AIC** of $\mathcal{M}_{10,BSS}$: 121846.5 (120530.5)
- R^2 of $\mathcal{M}_{10,BSS}$: 0.3743616 (0.3888385)

	GVIF
MaxTemp	2.300351
LogEvaporation	2.346201
Sunshine	3.193609
Humidity3pm	2.991694
Humidity9am	2.519413

Regression Model

UNIVERSITÀ DEGLI STUDI DI PADOVA

Assumptions and Diagnostics: Model(s) with all (subset) the explanatory variables - `WeatherV11`



- Strong evidence of a **systematic non-linear behaviour** in the residuals in a neighborhood of 1;
- **Assumption of Normality** of the residuals **not** satisfied;
- **Heteroskedasticity** of the residuals - high non-linearity in a neighborhood of zero. The outliers and the leverage point do not represent influencial points (wrt to the *Cook's Distance*)
- Presence of serious **Autocorrelation of the residuals**, according to the *Durbin-Watson test* ($\ast^2$):

  DW $= 1.8833$, p-value$< 2.2\text{x}10^{-16}$.

Regression Model

Some Considerations

Focusing on $\mathcal{M}_{10}$, we need to make some reflections about the problems encountered during the analysis.

In particular, treating the main sources of multicollinearity in the initial model the procedure led anyway to a model affected by both heteroskedasticity and autocorrelation of the residuals.

The **OLS** estimates, $\hat{\boldsymbol{\beta}}_{OLS}$, in presence of only one of the encountered issues, still lead to unbiased estimates, but not **BLUE** (i.e. not efficient), reason why reducing the multicollinearity was so important for our analysis. Furthermore, in absence of the satisfiability of such assumptions, inferential procedures for hypothesis testing and confidence interval building are no longer valid.

So, what's next?

Regression Model

A new perspective: Modeling according to `WeatherV11[LogRisk>0,]`

Characteristics of the reduced model $\mathcal{M}_{10}$

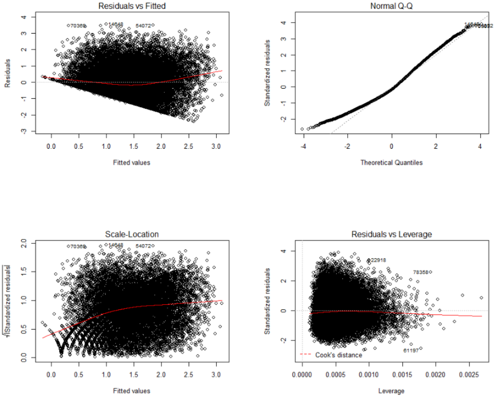
```
Call:
lm(formula = LogRISK ~ . - MinTemp - Temp9am - quarters - Temp3pm -
    Wind3pmX - Wind3pmY - Wind9amX - Wind9amY - WindGustX - Pressure3pm -
    WindGustY, data = WeatherV11[LogRISK > 0, ], na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4126 -0.6723 -0.1480  0.5962  3.5599

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    22.5404856  1.0004229  22.531  < 2e-16 ***
MaxTemp         0.0340589  0.0012734  26.746  < 2e-16 ***
LogEvaporation  0.0345192  0.0157145   2.197 0.028057 *
Sunshine       -0.0655008  0.0030724 -21.319  < 2e-16 ***
Humidity3pm     0.0167325  0.0005055  33.099  < 2e-16 ***
Humidity9am    -0.0028149  0.0005764  -4.884 1.05e-06 ***
Pressure9am    -0.0221332  0.0009749 -22.704  < 2e-16 ***
Cloud3pm        0.0210094  0.0043395   4.841 1.30e-06 ***
Cloud9am       -0.0332228  0.0038392  -8.654  < 2e-16 ***
Coast1         -0.0601403  0.0181843  -3.307 0.000944 ***
LogRainfall     0.0998348  0.0067130  14.872  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9153 on 19680 degrees of freedom
Multiple R-squared:  0.2409,    Adjusted R-squared:  0.2405
F-statistic: 624.5 on 10 and 19680 DF,  p-value: < 2.2e-16
```

A particular aspect that we already introduced was the atypical shape of the response, $\mathbb{Y}$. In fact, we noticed how the prevalent presence of zeros induced a non-linearity on its neighborhood, so we decide to restrict the analysis on the "rainy" observations, according to the threshold ($\mathbb{Y} > 0$ mm), modeling $\mathbb{Y}_{>0}$.

The BSS procedure lead to the following model.

Regression Model

Assumptions and Diagnostics: $\mathcal{M}_{10}$ - `WeatherV11[LogRisk>0,]`



- We can appreciate a consistent reduction in the non-linearity of the residuals;
- **Assumption of Normality** of the residuals is satisfied (approximately);
- **Heteroskedasticity** in the residual, present, but relatively unimportant. The outliers and the leverage point do not represent influencial points (wrt to the *Cook's Distance*)

DW=1.8822, p-value=$2.2\text{x}10^{-16}$

- **BIC** of $\mathcal{M}_{10,BSS}$: 52501.8
- **AIC** of $\mathcal{M}_{10,BSS}$: 52407.14
- R^2 of $\mathcal{M}_{10,BSS}$: 0.2405082

UNIVERSITÀ DEGLI STUDI DI PADOVA

Characteristics of the reduced model $\mathcal{M}_{10}$

```
Call:
lm(formula = LogRISK ~ . - MinTemp - MaxTemp - Temp9am - LogEvaporation -
    Cloud3pm - Cloud9am - Pressure9am - LogRainfall - Wind3pmX -
    Wind9amX, data = WeatherV17Positive, na.action = na.omit)

Residuals:
     Min       1Q   Median       3Q      Max
-2.29815 -0.60932 -0.02128  0.54119  3.07768

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.972849   5.480944  15.686  < 2e-16 ***
Temp3pm     -0.032153   0.011090  -2.899 0.003839 **
Sunshine    -0.031739   0.011051  -2.872 0.004183 **
Wind3pmY     0.015598   0.003151   4.951 8.94e-07 ***
Wind9amY     0.012362   0.003392   3.645 0.000284 ***
WindGustX    0.002363   0.001106   2.136 0.032964 *
WindGustY    0.002832   0.001302   2.176 0.029830 *
Humidity3pm  0.007244   0.002409   3.008 0.002712 **
Humidity9am -0.003385   0.001965  -1.723 0.085248 .
Pressure3pm -0.083031   0.005363 -15.484  < 2e-16 ***
quarters2    0.455500   0.119776   3.803 0.000153 ***
quarters3    0.547042   0.127811   4.280 2.08e-05 ***
quarters4    0.188898   0.117084   1.613 0.107042
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7818 on 843 degrees of freedom
  (44 observations deleted due to missingness)
Multiple R-squared:  0.3886,    Adjusted R-squared:  0.3799
F-statistic: 44.64 on 12 and 843 DF,  p-value: < 2.2e-16
```

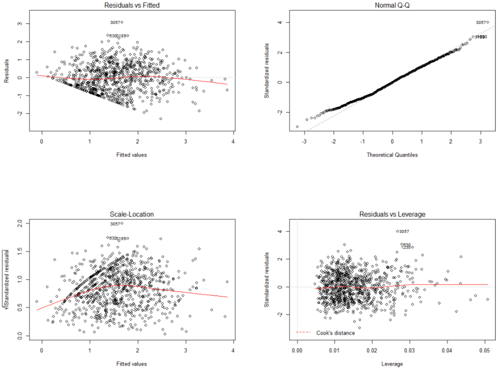
The problem is that the models are quite inaccurate if they take into account all the various microclimates that are present in Australia. In fact, considering one city the performance of the Multiple Linear Regression Model improves

- **BIC** of $\mathcal{M}_{10,BSS}$: 2089.25
- **AIC** of $\mathcal{M}_{10,BSS}$: 2022.718
- R^2 of $\mathcal{M}_{10,BSS}$: 0.37986

But still, residuals were autocorrelated.

Regression Model

Assumptions and Diagnostics: $\mathcal{M}_{10}$ - `WeatherV17`



- We can appreciate a consistent reduction in the non-linearity of the residuals;
- **Assumption of Normality** of the residuals is satisfied (approximately);
- **Heteroskedasticity** in the residual, present, but unimportant. The outliers and the leverage point do not represent influencial points (wrt to the *Cook's Distance*)

Presence of **Autocorrelation of the residuals**, according to the *Durbin-Watson test*

DW=1.897, p-value=0.05117

- **BIC** of $\mathcal{M}_{10,BSS}$: 2089.25
- **AIC** of $\mathcal{M}_{10,BSS}$: 2022.718
- R^2 of $\mathcal{M}_{10,BSS}$: 0.37986

Regression Model

Residuals' Modeling: Using AR(1) according to Cochrane-Orcutt

Serially autocorrelated residuals could lead to inexact inferences, making the model pretty unuseful.

Still, we wanted to look for possible ways to model them, since they present a systematic pattern (in a specific interval of time), so we applied the **Cochrane-Orcutt**, removing the autocorrelation of the residuals, expecting a 1^{st} order autocorrelation pattern (also more complex patterns are plausible).

`Weather17`	$\mathcal{M}_{10,OC}$	$\mathcal{M}_{10}$
Intercept	86.3123354	85.972849
Temp3pm	-0.0306172	-0.032153
Sunshine	-0.0312010	-0.031739
Wind3pmY	0.0160168	0.015598
Wind9amY	0.0129635	0.012362
WindGustX	0.0024170	0.002363
WindGustY	0.0027932	0.002832
Humidity3pm	0.0074308	0.007244
Humidity9am	-0.0038863	-0.003385
Pressure3pm	-0.0833846	-0.083031
quarters2	0.4663489	0.455500
quarters3	0.565489	0.547042
quarters4	0.1977468	0.188898
DW	$\mathcal{M}_{10,OC}$	$\mathcal{M}_{10}$
	1.99721, $4.41x10^{-1}$	1.89696, $5x10^{-2}$

- **BIC** of $\mathcal{M}_{10,BSS}$: 2089.25
- **AIC** of $\mathcal{M}_{10,BSS}$: 2022.718
- R^2 of $\mathcal{M}_{10,BSS}$: 0.37986

The aim of this section is to build a classifier able to predict whether tomorrow will rain or not. We performed three different procedures for this task.

Procedure 1:

- selection of the dataset with the 13 most promising variables and one observation every 3 days;
- division in training and test set (80% and 20%);
- construction of different **Logistic Regression** models with different number of variables;
- selection of the best model (the one with 9 variables), according to the **deviance difference test**;
- selection of the threshold that provides the best **Sensitivity** and **Specificity**.

Results 1:

- List of the 9 variables of the final model: Sunshine, Wind3pmX, Wind9amX, Wind9amY, Humidity3pm, Humidity9am, Pressure3pm, quarters, LogRainfall;
- VIF values are quite close to 1. The largest are those corresponding to Humidity3pm/9am.

Performance measures on the training set:

Accuracy	Specificity	Sensitivity
0.780	0.774	0.8

Performance measures on the test set:

Accuracy	Specificity	Sensitivity
0.773	0.756	0.827

Procedure 2:

- selection of the dataset with all the variables and one observation every 3 days;
- **under-sampling** to balance the dataset;
- division in training and test set (80% and 20%);
- **Principal Component Analysis** and **Logistic Regression**;

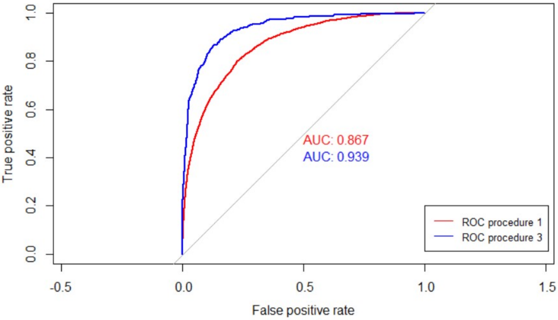
Accuracy of the resulting model:

Training set accuracy	**Test set accuracy**
0.777	0.769

UNIVERSITÀ DEGLI STUDI DI PADOVA

Procedure 3:
to conclude, we considered the dataset concerning just **one city**, and we performed the same steps described in the first procedure.

In this case, the classifier is able to model the phenomenon more precisely, as can be seen from these results.



Performance measures on the test set:

Accuracy	Specificity	Sensitivity
0.848	0.851	0.836

Regression Analysis

- Fitting a **Truncated Normal Model**, incorporating the information of the `Date` in the model;
- Adapt a **Semi-Parametrical appoach**, with Regression Splines (according to the `termplot()` function's output);
- Apply the latter models considering Clusters of cities, taking into account their specific microclimate;

Thank you for the attention

```
# Define a function that counts the number of outliers and
# their percentage.

NumberOfOutliers <- function(variable) {
  first_quartile = quantile(variable, na.rm = TRUE)[2]
  third_quartile = quantile(variable, na.rm = TRUE)[4]
  iqr = IQR(variable, na.rm = TRUE)

  lower = first_quartile - 1.5*iqr
  upper = third_quartile + 1.5*iqr

  num_outliers = sum(variable < lower | variable > upper,
                                           na.rm = TRUE)
  perc_outliers = 100*num_outliers/length(variable)

  return(data.frame(num_outliers,perc_outliers))
}
```

```r
# Define a function that convert compass directions into cartesian coordinates,
# to keep track of the cyclic behaviour of the directions.

CardinalToNumbers <- function(variable) {
        l = length(variable)
        NS = rep(0, l)
        WE = rep(0, l)

        directions = c("N", "NNE", "NE", "ENE", "E", "ESE", "SE", "SSE",
                       "S", "SSW", "SW", "WSW", "W", "WNW", "NW", "NNW")
        cosine = c(0, 0.5, sqrt(2)/2, sqrt(3)/2, 1, sqrt(3)/2, sqrt(2)/2, 0.5,
                   0, -0.5, -sqrt(2)/2, -sqrt(3)/2, -1, -sqrt(3)/2, -sqrt(2)/2, -0.5)
        sine = c(1, sqrt(3)/2, sqrt(2)/2, 0.5, 0, -0.5, -sqrt(2)/2, -sqrt(3)/2,
                 -1, -sqrt(3)/2, -sqrt(2)/2, -0.5, 0, 0.5, sqrt(2)/2, sqrt(3)/2)

        conversor = matrix(c(cosine,sine), nrow = 2, byrow = TRUE)
        colnames(conversor) = directions

        for (i in seq(1,l,length.out = l)){
                if (is.na(variable[i])){
                        NS[i] = NA
                        WE[i] = NA
                }
                else {
                        NS[i] = conversor[2, variable[i]]
                        WE[i] = conversor[1, variable[i]]
                }
        }
        return(data.frame(NS, WE))
}
```

- The VIF ($\ast^1$) coefficient is a multiple that determines the increase of the variance of β_j due to the correlation between X_j and other explanatory variables. More precisely, it is shown (e.g. Greene, 2011) that:

$$\mathbb{V}\left(\hat{\beta}_j\right) = VIF_j \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_{ij} - \bar{x}_j\right)^2}$$

So, if $R_j^2 = 0$ and therefore $VIF_j = 1$, we have that $\mathbb{V}(\beta_j)$ coincides with the variance that we would have if X_j was the only one explanatory variable in the model. As R_j^2 increases, $\mathbb{V}(\beta_j)$ will also increase: the *multicollinearity leads to instability in estimates*. "Generalized collinearity diagnostics", Monette et Al.

- The DW test is used to test the hypothesis that the residuals are serially correlated at lag 1 , i.e. that in the following model:

$$\epsilon_i = \rho\epsilon_{i-1} + v_i$$

the hypothesis being tested is:

$$\begin{cases} H_0 : \rho = 0 & \text{(no serial correlation)} \\ H_1 : \rho \neq 0 & \text{(serial correlation at lag 1)} \end{cases}$$

since we do not observe the true error terms, we use the OLS residuals $\hat{\epsilon}_i$ and calculate the Durbin-Watson statistic as:

$$DW = \frac{\sum_{i=2}^{N} \left(\hat{\epsilon}_i - \hat{\epsilon}_{i-1}\right)^2}{\sum_{i=1}^{N} \hat{\epsilon}_i^2}$$

Furthermore, its distribution no longer holds, when the equation of Y_i contains a lagged dependent variable, Y_{i-1}. As a quick rule of thumb, if the DW statistic is near 2 , then we do not reject the null hypothesis of no serial correlation.