Extrema der Luftschadstoffe Ozonextrema

Before starting, these are some useful commands:

getwd <- to show the current working directory
setwd <- to change of working directory
library <- to load the packages
install.packages <- install specific packages

OUTLINE:

In this exercise we will analyse the influence of meteorology on surface ozone concentrations.
For that, we will use regression analysis to examine the relationships between a single response variable (in our case, **ozone**) and one or more predictor variables. Daily maximum 8-hour average of ozone, and daily data of a set of meteorological variables will be used.

OBJECTIVES:

- Examine seasonal changes in surface ozone concentrations.
- Assess the relationship between ozone and meteorological variables.
- Identify the meteorological predictors that better explain the ozone variability.
- Differences between spring and summer

Exercise:

This exercise is divided in two parts:

a) Data analysis and visualization.
b) Regression analysis: Analyse the relationship between the predictand (O3) and the predictors (explanatory variables).

Packages required

MASS, stats, ggplot2, dplyr, relaimpo

A) Data:

The data contains 4018 daily values of 7 meteorological parameters and surface ozone concentrations (maximum 8-hour average) during the period of 2000-2010.
Below list of the variables included (daily values):

Predictors (Units)

- Blh: Boundary layer height (m)
- rh: Relative humidity at surface (1000 mb) (%)
- ssrd :Surface Solar Radiation downwards (W/m2)
- Tcc: Total cloud cover (0-1)
- tmax: Maximum temperature (K)
- ws :Wind speed (10m) (m/sg)
- Direction:Wind direction (Deg.)

Variable response

- O3 : Ozone (maximum 8-hour average)

#Read the file

```
mydata <-read.csv("/path/file.csv")
```

#Visualizing data:

1. Plot the distribution of the predictand (O3). For that, we could use the commands:

```
plot(mydata$o3,type="l",col="orange",xlab="day",ylab="O3",main="O3distribution")
hist(mydata$o3,main="Distribution of Ozone",xlab="O3")
```

Repeat this step for the rest of the variables (tmax, rh…)

2. Plot the data to look for multivariate outliers, non-linear relationships etc… It can be done with simple scatter plots: one variable *vs* another one (e.g. O3 *vs* tmax, O3 *vs* rh …):

```
plot(mydata$tmax,dat$o3,pch = 19, col = "blue", main = "Ozone vs Tx", xlab = "Tmax (k)",
ylab = "O3 (ppb)")
abline(lm(mydata$o3~mydata$tmax,dat),col="red") # Adding a regression line (y~x)
```

Use scatter plots to look at the relationship between the predictand to another predictor.

3. Analyse the relationships between all variables available dataset. Graphically, this can be done by constructing scatter plots matrices of all pair-wise combinations of variables in the data frame. There is useful to use **pairs** to look at the relationships:

```
pairs(mydata[,-1],panel=panel.smooth)
```

Alternatively, another function is provided:

```
# Load the function plot_panel.R

source('plot_panel.R')
pairs(mydata[,-1], cex.labels =0.9,
    lower.panel=panel.smooth, upper.panel=panel.cor,diag.panel=panel.hist)
```

Note: The pairs function plots every variable in the dataframe on the y axis against every other variable on the x axis. It only needs the name of the whole dataframe. The response variables are named in the rows and the explanatory variables are named in the columns.

4. We have examined the whole data set. Now, we will look at data by seasons, so we can see when ozone reaches the maximum values. Ozone season is usually defined from April-September.

B) Regression analysis in R:

In R the models are built by the ***lm*** function ({stats}), with returns the model object with the corresponding coefficients and the regression statistics.

The purpose of the exercise is to identify those **variables** that better explain the variability of surface ozone concentrations. Linear regression is used to identify the best fitted model, which describes the relationship between O3 and each variable. We will do the regression for summer months (**JJA**), where the meteorological influence is usually stronger. Then, you repeat the steps and fit another model for spring (MAM) to compare then models and the effect of the meteorological variables in each season.

Let's start by a simple linear regression model: (**lm(y~x,data)**), where ***y*** is the variable response and ***x*** the predictor.

B.1) Linear regression:

1. Analyse the O3 response to one single predictor. Let's start with maximum temperature. Fit a simple model:
   fit<-lm(o3~Tx, data=mydata) . # Note, mydata is assumed only for JJA (see the script)

2. Analyse the fit model: summary(fit)
   The model resulting fit will be an object of class lm for which a summary method showing the conventional regression analysis output is available. The output shows the estimates coefficients with corresponding standard errors and t-statistics as well as the F-statistic with associated p-value.

3. Exploring the object fit and the regression statistics: coefficients, residuals, fitted values, r-square. fit$coefficients.
   There are more methods for extracting the estimates, residuals or fitted:
   coef(fit)
   residuals(fit)
   fitted(fit)

4. Plot the model:

```
layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
hist(mydata$o3,main="Distribution of Ozone",xlab="O3")
hist(rstudent(fit), main="Model Student-Residuals",xlab="rstudent")
plot(residuals(fit),xlab="residuals",ylab="");
title("Simple Residual Plot")
acf(residuals(fit), main = "");
title("Residual Autocorrelation Plot ");
```

and

```
plot(fit)
```

B.2) Multiple linear regression:

To fit a multiple linear regression model with O3 as the response variable and a different set of explanatory variables, we could use the same function ***lm*** with more predictors (e.g. Tx + Gh) as:

```
fit<-lm(o3~tmax+rh,data=mydata).
```

1. Built different models with different predictors.
   If you want to modify a model (previously defined), it is useful to consider the ***update*** function, which will re-fit the model. For example, if you want to drop GH in the model:
   ```
   fit1<- update(fit,~.-rh),
   ```
   Or, to add another predictor:
   ```
   fit2<-update(fit1,~.+ssrd)
   ```
2. Include different predictors for comparing the models.

Model comparison:

3. Comparing the different models:

The anova() command computes analysis of variance (or deviance) tables. When given one model as an argument, it displays the ANOVA table. When two (or more) nested models are given, it calculates the differences between them: anova(fit,fit1)

The Akaike information criterion (AIC) is a measure of the relative quality of a statistical model for a given set of data. That is, given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Hence, AIC provides a means for model selection (Lower AIC better model) :AIC(fit,fit1)

Effect of predictors:

We are also interested in examining which predictor has a larger influence on ozone.
The R-square represents the proportion of the variance explained by a set of predictors (those selected for our model). We could assess the contribution of each predictor to the R-square. In R it can be done with the "relaimpo" package:

calc.relimp(fit, "lmg")

Then, you can see the proportion of each predictor to the total explained variance.

Summing up:

- *How is the relationship O3 and the different explanatory variables?*
- *Which is the most significant predictor?*
- *Which model seems to be better?*
- *How different are the models for each season (MAM and JJA) and the predictors?*

Exercise 2 (For extremes)

In this second exercise, we will fit another model regression for those O3 values above 50 ppb. For that, we will use *Generalized linear models (GLM)*, which are considered extensions of traditional regression models that allow the mean to depend on the explanatory variable through a link function.

The methodology will be exactly the same, but using the function **glm**.
The key of this exercise is that the predictand (O3) is converted to a binary variable. (0=not exceedance or 1=exceedance).

NOTES:

Interpreting the model output ***summary(fit)***

Name	Description
Residuals	The residuals are the difference between the actual values of the variable you're predicting and predicted values from your regression (**y-ŷ)**. For most regressions you want your residuals to look like a normal distribution when plotted. If our residuals are normally distributed, this indicates the mean of the difference between our predictions and the actual values is close to 0 (good) and that when we miss, we're missing both short and long of the actual value, and the likelihood of a miss being far from the actual value gets smaller as the distance from the actual value gets larger.
****Significance stars**	The stars are shorthand for significance levels, with the number of asterisks displayed according to the p-value computed. *** for high significance and * for low significance.
Coefficients	The estimated coefficient is the value of slope calculated by the regression. It might seem a little confusing that the Intercept also has a value, but just think of it as a slope that is always multiplied by 1
Standard error	Measure of the variability in the estimate for the coefficient. Lower means better but this number is relative to the value of the coefficient. As a rule of thumb, you'd like this value to be at least an order of magnitude less than the coefficient estimate.
t value	Score that measures whether or not the coefficient for this variable is meaningful for the model. It is used to calculate the p-value and the significance levels.

p-value	Probability the variable is *NOT* relevant. It should be as small as possible.
Significant legend	The more punctuation there is next to your variables, the better. Blank=bad, Dots=pretty good, Stars=good, More Stars=very good
Residual Std Error / Degrees of Freedom	The Residual Std Error is just the standard deviation of your residuals. The Degrees of Freedom is the difference between the number of observations included in your training sample and the number of variables used in your model (intercept counts as a variable).
R-squared	R^2 is a measure of the model's quality. Higher is better with 1 being the best. Mathematically, it is the fraction of the variance of the predictand (y) that is described by the regression model.
R-adjusted	The adjusted R-squared gives the percentage of variation described by only those independent variables that in reality affect the dependent variable. (It has a lower value than R-squared).
F-statistic & resulting p-value	Performs an F-test on the model. This takes the residual variance of the model and compares it to a model that has fewer parameters. It is expected that the model with more parameters fits better as it is more flexible. Here we test whether the model improvement is large enough to justify an additional parameter. Small p-values indicate that the model with more parameters is indeed significantly better. The DF, or degrees of freedom, pertains to how many variables are in the model.