MASM22/FMSN30/FMSN40 SPRING 2020
PROJECT 1: LINEAR REGRESSION
MONTHLY PRECIPITATION IN SWEDEN

Short instructions

Follow the instructions on the website under *Project 1: Instructions*. There you will also find the data and updated versions of this text.

Remember: The report must be written in English and submitted twice.

First in a preliminary version for peer assessment on 11.00 on Monday 27 April (Assignment: *Report1 - peer version*). At 11.30 you will get a random report for peer assessment. Read it and give feedback to its authors by 10.00 on Tuesday 28 April. And get corresponding feedback from whoever read your report. Then you must submit a final version for the teachers' eyes at 16.00 on Wednesday 29 April (Assignment: *Report1 - final version*).

Introduction

Various weather data have been collected in Sweden since the middle of the 18:th century. For instance, the precipitation has been measured in Lund since 1748 and the temperature since 1753. Swedish Meteorological and Hydrological Institute, `www.smhi.se`, has a large collection of data from the Swedish meteorological stations. A small part of that data is used in this project were you will try to model the total monthly precipitation (rain, snow, hail etc), as a function of the monthly average temperature and air pressure, and well as the location of the station (Lund in the South, Uppsala North of Stockholm, and Abisko above the Polar circle).

The data is located in the file `weather.rda` located on the *Project 1: Instructions* page. Download it to your computor and load it into R. It consists of the following variables:

`month`	a text (factor) variable of the format `"yyyy-mm"`
`year`	the year (numeric)
`monthnr`	the month of the year (numeric)
`location`	the location of the weather station (text): Lund, Uppsala or Abisko.
`rain`	total monthly precipitation (mm)
`temp`	average monthly temperature (°C)
`pressure`	average monthly air pressure (hPa)

Since not all stations have measured all variables all the time, I have deleted any uncomplete cases. Also note that the different stations have different time periods.

1 Precipitation as a function of temperature

Uses the techniques in Lab 1.

We want to find a suitable simple linear regression model for describing how the total monthly amount of precipitaion (`rain`) varies as a function of the temperature (`temp`), ignoring all the other variables.

1.(a). Start by plotting the data and fitting a linear model, without any transformations. Then perform a basic residual analysis and comment on the results. Is it a good model?

1.(b). The residual analysis in 1.(a) may (=should!) indicate that it would be better (but maybe not perfect) to transform some variable(s) in order to get a more linear relationship. Do that and present the corresponding results, as before.

1.(c). Compare the results and plots for the two models and explain what features in the plots caused you to decide which model is the better (less bad) one.

1.(d). Present the better model (from 1.(b)), including the (transformed) linear model, as well as how this translates back to the original, untransformed, scale.

Report the β-estimates with 95 % confidence intervals. If the final model uses some transform of β, present those estimates as well, together with confidence intervals.

Explain how, according to the model, the precipitation changes when the temperature increases by 1 °C, on average.

1.(e). Plot precipitation against temperature together with model 1.(b), its confidence interval and prediction interval.

1.(f). For months where the average temperature is 5 °C, how much would the total monthly amount of precipitation be expected to vary? Answer with a suitable interval that would be expected to contain 95 % of the months.

2 Precipitation as a function of temperature and more

Uses the techniques of Lab 2 as well

2.1 Temperature again

2.(a). Using your better model from 1.(b), perform a suitable test to determine whether the temperature has a significant effect on the amount of precipitation.

2.2 Temperature and pressure

We would expect that the monthly amount of precipitation depends not only on the average temperature but also on the average air pressure. When the pressure is low the weather is more unstable and it is more likely to rain.

2.(b). Plot all three variables, `rain`, `temp` and `pressure`, against each of the others.

Does is seem like there might be a linear relationship between precipitation and air pressure? Does it improve if we use the same transformations as in 1.(b)?

Does it look like there might be a strong linear relationship between temperature and pressure, that might cause problems?

2.(c). Fit a linear model (using the same transform as in 1.(b)) of precipitation as a function of both temperature and pressure. Report the β-estimates together with their confidence intervals. Also test whether the inclusion of pressure was statistically significant.

2.(d). Perform a basic residual analysis, plotting the residuals against the fitted values, temperature, and pressure. Also do a Q-Q-plot. Make a visual comparison between these residuals and those from 1.(b). Have they improved?

2.(e). Explain how, according to this model, the precipitation changes when the temperature increases by 1 °C, on average. Compare with the result in 1.(d). Are the results substantially different?

2.(f). Explain how, according to this model, the precipitation changes when the pressure increases by 20 hPa, on average.

2.(g). For months where the average temperature is 5 °C, how much would the total monthly amount of precipitation be expected to vary? Answer with a suitable interval that would be expected to contain 95 % of the months when the average air pressure is 1000 hPa and another interval for when the average air pressure is 1020 hPa. Explain how the difference between these to intervals relates to 2.(f).

2.3 Temperature and pressure with interaction

We could also expect that there is an interaction between teperature and pressure so that the effect of pressure when it is cold is different from the effect when it is warm.

2.(h). Examine this by fitting a model with an added interaction term. Report all β-parameters together with their confidence intervals. Also test if the interaction term is statistically significant.

2.(i). Did any of the other parameters change substantially when we included the interaction? Why might that be? Hint; the average of the air pressure values is 1012 hPa. Refit the model but with `I(pressure - 1012)` instead of `pressure`.

2.(j). Perform a basic residual analysis, plotting the residuals against the fitted values, temperature, and pressure. Also do a Q-Q-plot. Make a visual comparison between these residuals and those from 1.(b) and 2.(d). Have they improved?

2.(k). Explain how, according to this model, the precipitation changes when the temperature increases by 1 °C, on average, as a function of the air pressure. Also examplify by estimating the change when the air pressure is 1000 hPa, and when it is 1020 hPa. Compare with the results in 1.(d) and 2.(e). Comment on any interesting differences.

2.(l). Use this model to estimate the expected monthly amount of precipitation for the four different combinations of temperature and pressure, where the temperature is -10 °C or 10 °C and the pressure is 1000 hPa or 1020 hPa. Also report 95 % confidence intervals for the estimates.

2.4 Temperature, pressure and location

It is reasonable to expect that the location of the measurment station also has an effect on the amount of precipitation, due to different climates.

2.(m). Make a table over the number of observations of each of the locations and chose a suitable reference category.

2.(n). Add the location to the model from 2.(h). Report all β-parameters together with their confidence intervals. Also test if the addition of the locations is statistically significant. Explain what type of test you are using, the p-value and the conclusion.

2.(o). Did any of the other parameters change substantially when you included the location?

2.(p). Perform a basic residual analysis, plotting the residuals against the fitted values, temperature, and pressure, each separately for the different locations. Also do Q-Q-plots, both for all residuals and separately for the different locations. Make a visual comparison between these residuals and those from 1.(b), 2.(d) and 2.(j). Have they improved further?

2.(q). For any given combination of temperature and pressure, which location has the highest expected amount of rain?

3 Precipitation — which variables are needed?

Uses the techniques of Lab 3 as well

3.1 Outliers and influential observations

3.(a). Use the model from 2.(n). Calculate the leverages and plot them against temperature and against pressure, separately for each location. Explain why the smallest leverage in Lund and in Abisko are larger than the smallest leverage in Uppsala. Hint: why did we pick Uppsala as reference category?

3.(b). Identify any observations with unusually high leverage, e.g., larger than 0.026. Plot temperature against pressure, separately for each location, and highlight these problematic observations. Explain what makes the leverage high in these observations.

3.(c). Plot the studentized residuals against the fitted values and highlight the observations from 3.(b). Are their residuals problematic? Are there any other problematic observations with large residuals?

3.(d). Identify the largest residual in each of the three locations and plot log-precipitation against temperature and against pressure, separately for each location, highlighting both the high-leverage obervations from 3.(b), and the observations with the largest residuals. What type if weather is causing the problem with the residuals?

3.(e). Calculate Cook's D and plot it against, e.g., pressure, separately for each location, and highlighting the observations identified in 3.(b) and 3.(d). Did any of the observations with high leverage have a large influence on the estimates? Did the observations with the largest residuals have a large influence? Are there other observations that had a large influence?

3.(f). Create a new data set without the problematic observations you have identified (but keep the observations with high leverage but low Cook's D) and re-fit the model using the reduced data set. Plot the new studentized residuals and the new Cook's D and compare with the plots in 3.(c) and 3.(e).

3.2 Model comparisons

We will now use the reduced data set for the further analysis.

3.(g). Re-fit model 1.(b), 2.(d) and 2.(h) as well, and calculate R^2, R^2_{adj}, AIC and BIC for all four re-fitted models. Which of the models is "best" and how much of the variability in log-precipitation can it explain?

3.(h). Maybe the combination of temperature and air pressure has different effects on the precipitation for different locations. Investigate this by fitting a model with the three-way interaction `temp*pressure*location` and test whether this is a statistically significant improvement.

3.(i). Perform a backward elimination, using BIC as criterion in order to reduce the number of parameters in model 3.(h). Which model do we end up with?

3.(j). Perform a forward selection, using BIC as criterion, starting with a model with only intercept, `log(rain) ~ 1`, and with model 3.(h) as the largest model allowed. Which variable is the first to be selected? Which model do we end up with?

3.(k). It is not unreasonable to think that there might be seasonal differences in the effect of temperature, air pressure and location on the amount of precipitation.

Create a new categorical variable grouping the months, `monthnr`, into, e.g., Spring (March – May), Summer (June – August), Autumn (September – November), and Winter (December – February). As simple way to create the new variable is

```
my.data$season <- "winter"
my.data4season[my.data$monthnr == 3] <- "spring"
my.data4season[my.data$monthnr == 4] <- "spring"
etc
```

Since you have a lot of data you can try a forward selection where the largest model allowed is a model with interactions between all four variables: `temp * pressure * location * season`.

Present the final model. Are there seasonal differences? Does a change in air pressure affect the precipitation in the same way regardless of the season? How much of the variability can this final model explain?