Project 1: Linear Regression
MASM22/FMSN30/FMSN40: Linear and Logistic Regression (with Data Gathering), 2023

Peer assessment version: **12.30 on Wednesday 26 April**

Peer assessment comments: **13.00 on Thursday 27 April**

Final version: **17.00 on Friday 28 April**

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 air pressure, temperature, wind speed, location of the station (Lund in the South, Uppsala North of Stockholm, and Katterjåkk on Malmbanan[1] upp in the mountains close to the Norwegian border in the far North), and the season.

The data is located in the comma-separated file `weather.csv` on the *Project 1: Instructions* page. Save it in the `Data` subfolder in your RStudio project folder and then read it into R:

```
weather <- read.csv("Data/weather.csv")
```

It consists of the following variables:

`location`	the location of the weather station (text): Lund, Uppsala or Katterjåkk.
`month`	a text variable of the format `"yyyy-mm"`
`rain`	total monthly precipitation (mm)
`pressure`	average monthly air pressure (hPa)
`temp`	average monthly temperature (°C)
`speed`	average monthly wind speed (m/s)

The different stations have different time periods as well as gaps in their series.

Our goal is to model how the total monthly precipitation varies as a function of one or several of the other variables, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$. We will use a linear regression model $Y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i$ where the random errors ε_i are assumed to be pairwise independent and $N(0, \sigma^2)$. In order to fulfill these model assumptions we may have to let Y be a suitable transformation of the total monthly precipitation.

[1] `https://en.wikipedia.org/wiki/Iron_Ore_Line`

Part 1. Finding a suitable transformation

According to Stidd, C.K., "Cube-root-normal precipitation distributions", *Transactions, American Geophysical Union*, Volume 34, Issue 1 (1953), p.31–35[2]:

> "It is observed that the cube roots of precipitation amounts often are distributed normally. This is found to apply to data for stations ranging from arctic to tropic regions, from extremely wet to extremely dry regions, and for precipitation periods of one day to one year."

We will choose between three possible transformations of the precipitation where the dependent variable Y will be either

- the precipitation, $Y = \texttt{rain}$,
- the logarithm of the precipitation, $Y = \texttt{log(rain)}$,
- the cube root of the precipitation, $Y = \sqrt[3]{\texttt{rain}} = \texttt{rain\^{}(1/3)}$.

It might be assumed that the air pressure is an important variable so we start with that.

1(a). Determine and motivate which of the three transformations is best, i.e., the one where the model assumptions are most reasonably fulfilled.

In addition to your residual analysis plots, you should also present five additional plots: one for each of the transformations (log and cube root), with the transformed precipitations against air pressure, with the estimated linear relationship, its confidence interval, and a prediction interval for new observations. Also present the three corresponding plots for the resulting relationships between precipitation (mm) and air pressure.

Your report should also include how the linear relationship between the transformed precipitation and the air pressure translates into a relationship between the original precipitation (mm) and the air pressure (hPa), for each of the three models, as well as the estimates of the β-parameters, with 95 % confidence intervals.

The model with the best transformation will be referred to as *Model.1(a)*.

In the rest of the project you should use the best of the three transformations.

1(b). Express how the rate of change in the precipitation varies as a function of the air pressure and the β-parameters in the three models by taking the derivatives of the relationships with respect to the air pressure.

Calculate the estimated change rates for air pressures 990, 1010, and 1030 hPa for all three models and comment on the differences between the rates, and relate them to the plots in 1(a).

[2] https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/TR034i001p00031

Part 2. Adding more explanatory variables

2(a). Test if there is a significant linear relationship between the transformed precipitation and the air pressure, according to *Model 1(a)*. Report the type of test you use, the null hypothesis, the value of the test statistic, its distribution, the p-value and the conclusion.

2(b). **Pressure and location** Turn the `location` variable into a factor variable, using the text as labels,

```
weather$location <- as.factor(weather$location)
```

and determine which location would be best to use as reference, and why.

Plot the residuals from *Model 1(a)* against `location` using boxplots. Does this look good or are there systematic differences between the locations?

Account for the systematic differences by adding `location` to *Model 1(a)* and state the resulting model, the β-estimates and their 95 % confidence intervals. This model will be referred to as *Model 2(b)*.

Test if this gives a significant improvement compared to *Model 1(a)*. Also test whether `pressure` is still significant in the model when we have added `location`. Report the types of test you use, the null hypotheses, the values of the test statistics, their distributions, the p-values and the conclusions.

Plot the residuals from *Model 2(b)* against `location` and compare with the earlier plot of the *Model 1(a)*-residuals. Comment on any differences between the plots.

Plot the transformed precipitation against air pressure together with the estimated linear relationships from *Model 2(b)*, using colour to separate the locations, and relate the differences between the relationships to the estimated model parameters.

Plot the observed precipitation against the air pressure, separately for each location, and add the estimated relationships. Explain how the addition of location to the model changes these relationships.

2(c). **Pressure, location and wind speed** Start by ploting the (transformed) precipitation against the average monthly wind speed, `speed`, the wind speed against the air pressure, and also make a boxplot of the average monthly wind speed separated by `location`. Does there seem to be any relationship between precipitation and wind speed? Does there seem to be any strong linear, possibly problematic, relationship between wind speed and air pressure? Are there systematic differences in wind speed between the locations that could, potentially, explain the systematic differences i precipitation?

Add `speed` to *Model 2(b)* and state the resulting model, the β-estmates and their 95 % confidence intervals. This model will be referred to as *Model 2(c)*.

Test if this gives a significant improvement compared to *Model 2(b)*. Also test whether `location` is still significant in the model when we have added `speed`. Report the types of test you use, the null hypotheses, the values of the test statistics, their distributions, the p-values and the conclusions.

Compare the value of the β-parameter for `pressure` to the corresponding value in *Model 2(b)*. Did it change substantially?

Compare the values for the β-parameters for the dummy variables for the two non-reference locations to the corresponding values in *Model 2(b)*. Explain why one of them changed so much more than the other when you added wind speed to the model. *Hint:* your boxplot for the wind speeds and the sign of the β-parameter for `speed`.

2(d). **. . . and interaction** Plot the residuals from *Model 2(c)* against `speed`, separated by `location`. Use `facet_wrap(~location, scales = "free_x")` to get separate subplots with different scales on the x-axes. You may also want to add a `geom_smooth()`. Does this look good or are there systematic differences between the locations? Explain what those differences mean.

Add the interaction between `speed` and `location` to *Model 2(c)* and state the resulting model, the β-estimates and their 95 % confidence intervals. This model will be referred to as *Model 2(d)*.

Test if the addition of the interactions gives a significant improvement compared to *Model 2(c)*. Report the type of test you use, the null hypothesis, the value of the test statistic, its distributions, the p-value and the conclusion.

Plot the residuals from *Model 2(d)* against `speed`, separated by `location` and compare with the previous plot. Comment on any changes between the plots. Did the addition of the interaction have any effect on the systematic behaviour of the residuals?

Calculate the slopes for `speed` in the linear model for the transformed precipitation, for the three locations and compare them to the corresponding slope from *Model 2(c)*. Does the precipitation change in the same direction in all three locations?

2(e). Calculate the estimated expected precipitation (mm) in each of the three locations for a month where the air pressure is 1011 hPa, both for wind speed 2 m/s (the average in Uppsala) and for 5 m/s (the average in Lund), for *Model 2(b)*, *Model 2(c)*, and *Model 2(d)*.

Relate the differences in the predictions to the differences between the models. For instance, how did the change in the β-parameters for the dummy variables between *Model 2(b)* and *Model 2(c)* change the predictions, and how did the interactions between location and wind speed in *Model 2(d)* change the effect of the windspeed for the different locations?

The observed average precipitations under these conditions were

	Katterjåkk	Lund	Uppsala
2 m/s	66	74	53
5 m/2	76	69	21

Which of the three models seems best (least bad) at predicting this?

Part 3. Model validation and selection

3(a). **Influential observations**. Calculate the leverage from *Model 2(d)* and identify (report the location and year-month) the seven observations with the highest leverage. (*Hint:* you can plot them against something to find a suitable cut-off value) Also calculate Cook's distance and identify the observation with the largest Cook's distance.

Plot air pressure against wind speed, separately for the three locations, and highlight the seven high-leverage observations and the high Cook's distance observation. Explain why the seven observations have the highest leverage. Also explain why the high Cook's distance observation has a lower leverage.

Investigate the DFBETAS to find the β-parameter that the observation with the highest Cook's distance has had the highest influence on. Then plot the (transformed) precipitation against the corresponding variable, highlighting the observation, and explain why it had a large influence on that parameter.

3(b). **Advanced residual analysis.** Calculate the studentized residuals, r_i^*, for *Model 2(d)* and plot them against the linear predictor, adding suitable reference lines. Also plot $\sqrt{|r_i^*|}$ in a similar manner. Comment on the result. Specifically, is the any non-linear behaviour left in the residuals? Does the variance appear to be constant?

3(c). **Variable selection.** We will now take the temperature, `temp`, into account and also add a seasonal variation. Pick out the month number from the `month` variable (`substr()`) and turn it into a factor variable:

```
weather$monthnr <- as.factor(substr(weather$month, 6, 7))
```

Then fit the null model using only an intercept, as well as a full model using all five variables and all their interactions:

```
lm(rain^(1/3) ~ pressure*location*speed*temp*monthnr, data = weather)
```

Perform a stepwise selection using BIC as criterion, starting with the null model, using the full model as upper scope. For each step, report which variable or interaction is included in or excluded from in the model.

Construct a table containing the number of β-parameters, the R^2, adjusted R^2, AIC and BIC for the following seven models: the null model, *Model 1(a)*, *Model 2(b)*, *Modle 2(c)*, *Model 2(d)*, the stepwise BIC model, and the full model. Motivate which model you find best, and in what sense.

How much of the variability of the transformed precipitation does this model explain?

Finally, plot the studentized residuals for the best model and compare them to the corresponding residual plots for *Model 2(d)* in 3(b). Has there been any improvement in the behaviour of the residuals?

End of Project 1