MSG500/MVE190
Linear Models - MiniAnalysis 2
Due: upload on Canvas by 12.00 on Friday 27 November

When and where?

Connect at 13.15 on 27 November on zoom: https://chalmers.zoom.us/j/69509322860 with passcode: 519431

How to submit mini-analysis 2

- If you haven't done this yet, BEFORE submitting you must register your group on Canvas, by going to "People → groups". Otherwise Canvas only tells me the name of the one student who submitted, not the other one.
- Go to Canvas, "Assignments→mini analysis 2". Only one person in each group should upload the slides on Canvas (powerpoint, or pdf slides made with Beamer-LaTeX or some other presentation tool). **Strict deadline for submission: 12:00 on Friday 27 November**.

You must attend mini-analyses presentations and be ready to possibly present. If no slides are uploaded, the group will have to jointly write a more detailed report (instead of slides) and send it to picchini@chalmers.se **by Tuesday 1 December**. If slides are uploaded and only one student attends, it is only the absent group-member that has to write a report.

Format

.
Important addition: make sure the first slide reports the name of all students in the group. Besides this, also look at the section "Format" the document for mini-presentation1.

While in zoom...

Please, your zoom should show your full name (I take presences), not just initials or nicknames. Set it up so that it shows your full name.

1 More analysis for bikeshared data

We keep studying the bikesharing data, where a lot of the most useful information has stayed uncovered so far, especially with regard to categorical covariates, which in the previous analysis you have probably found to be informative to explain the variation in the count response variable.

Build up on what you have already discovered and move forward.

As usual you might decide to analyze only a subset of the data, depending on what you are looking for.
Here follow some ideas for exploration. None of them are necessary, these are just examples.

Be creative, you can follow your own path. For example, sometimes you may want to analyze all data available, but you may also decide that you do not want to consider everything at the same time. For example you may analyze data for a specific season. Or analyze data for all seasons, and instead consider only a specific range of hourly times. This is up to you.

Some example is below **but again your are not required to strictly follow what is below:**

- for example, you know that you can extract the `hours` when a bike has been used/shared by appropriately using the package ``chron'' on variable `timestamp`, as discussed in section 3 of the previous mini1 document. Then last week some of your colleagues decided to create a categorical variable `is_rushhour`, having "levels" `rush_hour` if the hour was between 6am-9am or between 4pm-7pm, and level `not_rush_hour` for the remaining times. You may decide to use such categorical variable in a linear regression model (possibly multivariate).

- another alternative to the above is, again starting from the hours, to create another categorical variable with more than 2 "levels". Example, if you make `boxplot(cnt~hrs)` you realize that it makes sense to create a categorical variable `hrs_cat` starting from the original `hours`. You can do this by appropriately using `which()` (see also the mini1 document). For example as

```
hrs_cat <- hours              # copy the original vector of hours
id_0_5 <- which(hrs_cat<=5)  # finds indeces of hours between 0 and 5am
hrs_cat[id_0_5] <- "hrs_0_5" # put in "value" "hrs_0_5"
id_6_9 <- which(hrs_cat>=6 & hrs_cat<=9)  # finds indeces of hours between 6am and 9am
#etc
```

  The above has created the level `"hrs_0_5"`, and you can do similarly to create other groups of hours. But then once you are finished you have to remember to use the `factor()` function, before using `hrs_cat` in a model.

- Remember: it is **important** to try to give an interpretation (when possible, that is unless transformations of variables make this hard) of the estimated parameters. Not just report their values. Parameters of dummy-variables (levels of categorical covariates) are easy to interpret.

- you could use a Partial F test to check whether adding some covariate to a previously constructed model looks like a good idea.

- Then again, it *may be* that if the data (or subset thereof) you are fitting is large enough we could incur in the problem of always having a small p-value, which makes conclusions not so reliable (see pages 9-10 in slides_6.pdf, and minute 9 in the video L6_2.mp4). So you could you perhaps also check for the existence of an effect-size, to see if there is some noticeable difference in the expected response when you increase by 1 the value of a covariate (if the variable is quantitative; if not, it is the same to check what happens to the prediction conditionally on a certain level of a categorical covariate, instead of the baseline/reference level).

- Now that you have learned how to fit models also incorporating categorical covariates, plot residuals vs fitted response. How does this look? How about standardised residuals vs fitted responses?

- ...you can do more or something a bot different, it's up to your imagination.

Comment on your findings and considerations. Notice: some of the concepts above may be introduced in next lectures, so you can revisit your preliminary analyses before the deadline.

Variables:

- timestamp: day of the year and hour

- "cnt" - the number of new bike shares in the considered timestamp

- "t1" - real temperature in C
- "t2" - perceived temperature in C
- "hum" - humidity in percentage
- "wind_speed" - wind speed in km/h
- "weather_code" (see below for a description)
- "is_holiday" - 1 means holiday / 0 for non holiday
- "is_weekend" - 1 if the day is weekend
- "season" - meteorological seasons: 0-spring ; 1-summer; 2-fall; 3-winter.

"weather_code": 1 = Clear ; 2 = scattered clouds / few clouds; 3 = Broken clouds; 4 = Cloudy; 7 = Rain; 10 = rain with thunderstorm; 26 = snowfall; 94 = Freezing Fog.