1 Instructions

You must write a report, in English. Work in groups of two. Discussion between groups is permitted (and encouraged), as long as your report reflects your own work. Write a clear report, presenting your approach to the assignment, discussing methods and results. Results discussion and interpretation is important. Just reporting results is not enough! It should be noted that for some questions there isn’t a unique “right” answer and there are a myriad of different issues that you could discuss, so use your imagination. In addition to the text, use as many figures and tables as necessary, with explanatory captions.

The report should be readable, not a random disorganized collection of thoughts, plots and tables (see also the Peer Review Guidelines at the end of this document). For example, it should be possible for the reader to understand what you are doing without having access to your code. Also, key information may be better summarized in tables than by including the R printouts (e.g. it may be enough to give regression coefficients and p-values without all the accompanying information provided by R). There is no need to include your R code in the report, but you can include some of the R output.

1.1 Peer review

Bring a printed version of your report to the peer assessment (16 May, 815), or email the report to anna@maths.lth.se at least 14.00 on Tuesday, 15 May! so I can print a copy.

1.2 Final submission

E-mail the final version (a single PDF document) to one of (depending on your course) the following addresses by the deadline 16.00 at Thursday 17 May. Also attach to the same message your R-files (or implementation in other language), in a file named proj1.R that can be used to run your analyses.

- MASM22/FMSN30 students: email to fmsn30@matstat.lu.se
- FMSN40 students: email to fmsn40@matstat.lu.se

Subject field of the email: write “Project2 by studid1 and studid2” where studid1 and studid2 are the id numbers for two students in a given group.

Example: Project2 by d08xhj and d08fh
2 Atmospheric particles in Oslo

2.1 Summary

The data is a random subsample of 500 observations from a data set that originates in a study where air pollution at a road is related to traffic volume and meteorological variables, collected by the Norwegian Public Roads Administration. The data is hourly measurements at Alnabru in Oslo, Norway, between October 2001 and August 2003. In order to get rid of the strong correlation between successive measurements a random sample of the original, larger, data has been taken.

The objective is to model the probability that the concentration of atmospheric particles with a diameter between 2.5 and 10 μm, PM\textsubscript{10}, exceeds the limit 50 μg/m\textsuperscript{3}. This is the Swedish limit for the daily average but since we do not have access to the daily averages we will compare our hourly values to this limit\textsuperscript{1}.

2.2 Variables:

The data file pm10.txt can be downloaded from the course home page.

- **highpm10**: the concentration of PM\textsubscript{10} particles (categorical),
  - 0 = PM\textsubscript{10} \leq 50 μg/m\textsuperscript{3},
  - 1 = PM\textsubscript{10} > 50 μg/m\textsuperscript{3}
- **cars**: the number of cars per hour
- **temp2m**: temperature 2 meters above ground (degree C)
- **windspeed**: wind speed (meters/second)
- **winddirection**: wind direction (categorical),
  - 1 = NE,
  - 2 = SE,
  - 3 = SW,
  - 4 = NW
- **time**: time of day (categorical),
  - 1 = 01–06,
  - 2 = 06–12,
  - 3 = 12–18,
  - 4 = 18–24

3 Questions

3.1 High particle levels and the time of day

(a). Estimate the probability of the concentration of PM\textsubscript{10} exceeding 50 μg/m\textsuperscript{3} on a randomly chosen hour. Also calculate a 95 % confidence interval for this probability.

(b). Fit a logistic regression model where the probability of PM\textsubscript{10} exceeding 50 μg/m\textsuperscript{3} depends on the time of day. Report the estimated odds ratios for the three non-reference categories and calculate 95 % confidence intervals of these odds ratios.

(c). Use the model to estimate the probability of PM\textsubscript{10} exceeding 50 μg/m\textsuperscript{3} at 3 o’clock in the morning, at 11 o’clock in the morning, at 15 o’clock in the afternoon and at 23 o’clock in the evening. Also calculate 95 % confidence intervals for the probabilities.

\textsuperscript{1}For more on atmospheric particles, see https://en.wikipedia.org/wiki/Particulates and https://www.naturvardsverket.se/Stod-i-miljoarbetet/Vagledningar/Luft-och-klimat/Miljokvalitetsnormer-for-utomhusluft/Gransvarden-malvarden-utvarderingstrosklar/
3.2 High particle levels and the number of cars

(a). Plot the number of cars against the time of day. Which variable would you expect to influence the PM$_{10}$ concentration, time of day or number of cars?

(b). Fit a model where the probability of PM$_{10}$ exceeding 50 μg/m$^3$ depends on the number of cars. Report the odds ratio for cars and its 95% confidence interval. How much does the odds change if the number of cars is increased from 1000 to 1100?

(c). Use the model from (b) and estimate the probability of PM$_{10}$ exceeding 50 μg/m$^3$ when the number of cars is 300 and when it is 3000. Also calculate 95% confidence intervals for the probabilities.

(d). We might suspect that it would be better to use the logarithm of the number of cars instead. Fit a model where the probability of PM$_{10}$ exceeding 50 μg/m$^3$ depends on the logarithm of number of cars. Report the odd ratio for log cars and its 95% confidence interval. How much does the odds change if the number of cars is increased from 1000 to 1100?

(e). Use the model from (d) to estimate the probability of PM$_{10}$ exceeding 50 μg/m$^3$ when the number of cars is 300 and when it is 3000. Also calculate 95% confidence intervals for the probabilities. Compare with the result in (c). Any major differences?

(f). Use some suitable measures to determine whether model (b) or model (d) is best.

(g). Plot the estimated probability as a function of the number of cars using the two different models in the same plot. Also add the corresponding 95% confidence intervals for the probability. Any major differences? Add a kernel smoother regression curve to the plot (use a large bandwidth, e.g. 2000). Which of the two models seems closest to the data?

(h). If we want the probability of PM$_{10}$ exceeding 50 μg/m$^3$ to be less than 10 %, how many cars per hour we can allow, according to model (d)?

(i). Determine whether there are any observations that has had a large influence on the estimates.

3.3 High particle levels and . . .

(a). Starting with model (d) use a suitable test to determine whether adding the time of day would improve the model.

(b). Now use all the variables to build an even better model. Which variables have a significant effect on the probability of PM$_{10}$ exceeding 50 μg/m$^3$? Does the model seem reasonable, in particular the sign of the β’s?

---

End of Project 2
4 Useful R-commands for logistic regression

- Fit a logistic regression, get the estimated odds (for the intercept) and odds ratios (for the x-variables), and their confidence intervals:
  
  ```r
  mymodel <- glm(y ~ x, data=mydata, family="binomial")
  or <- exp(mymodel$coefficients)
  ci.or <- exp(confint(mymodel))
  ```

- Predicted probabilities
  
  ```r
  p.hat <- predict(mymodel, x0, type="response")
  ```

- Predicted log odds and their standard errors for computing confidence intervals for the probabilities:
  
  ```r
  logit.hat <- predict(mymodel, x0, se.fit=TRUE)
  logit.lo <- logit.hat$fit - 1.96*logit.hat$se.fit
  logit.hi <- logit.hat$fit + 1.96*logit.hat$se.fit
  ci.odds <- cbind(lo = exp(logit.lo), hi=exp(logit.hi))
  ci.p <- ci.odds/(1+ci.odds)
  ```

- AIC/BIC for a logistic regression
  
  ```r
  AIC(mymodel)
  AIC(mymodel, k=log(nrow(mydata)))
  ```

- Likelihood ratio test for nested models and \( \chi^2 \)-quantile and P-value:
  
  ```r
  lrtest <- anova(mymodel.reduced, mymodel.full)
  qchisq(1-alpha, lrtest$Df)
  pchisq(lrtest$Deviance, lrtest$Df, lower.tail=FALSE)
  ```

- Leverage, Pearson residuals and Cook's distance:
  
  ```r
  influence(mymodel)$hat
  influence(mymodel)$pear.res
  cooks.distance(mymodel)
  ```

- Pseudo \( R^2 \)-values require the package pscl. Install it once on your computer. Then activate it (once every R-session) with `library(pscl)`. Among the output: r2ML = Cox-Snell, r2CU = Nagelkerke:
  
  ```r
  pR2(mymodel)
  ```

- Calculating a kernel smoother regression of a 0/1 y variable as a function of one continuous x-variable and adding it to a plot:
  
  ```r
  ksm <- ksmooth(x, y, bandwidth = w)
  plot(y~x, ylim=c(0,1))
  lines(ksm$x, ksm$y) or lines(ksm)
  ```
5 Peer review guidelines

The following guidelines for peer-review should be followed.

5.1 Questions regarding the content

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Have all the tasks in the assignment been completed?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2. Does the report contain relevant figures and tables?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>3. Has all notation been properly introduced and/or explained?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>4. Has the model been properly introduced?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>5. Are the results properly presented and discussed?</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>

5.2 Questions regarding the report presentation

<table>
<thead>
<tr>
<th></th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Does the report have:</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>• Title, authors, and date?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>• Page numbers?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>• Introduction?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>• Results and/or conclusions?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>2. Has the report been proofread? Have language and spelling mistakes been corrected?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>3. Are figures and tables:</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>• Numbered?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>• Equipped with suitable captions?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>• Referred to in the text?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>4. Is the text divided into paragraphs and well structured with clear and suitable section headings?</td>
<td>☐</td>
<td>☐</td>
</tr>
<tr>
<td>5. Is the report easy to read, and understandable without access to the project description?</td>
<td>☐</td>
<td>☐</td>
</tr>
</tbody>
</table>