PROJECT 2: LOGISTIC REGRESSION

MASM22/FMSN30/FMSN40: LINEAR AND LOGISTIC REGRESSION (WITH DATA GATHERING), 2019

Peer assessment: **Wednesday 15 May (8^{15}–9)**

Final deadline: **Friday 17 May, 16.00**

---

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 (15 May, 8^{15}), or email the report to anna.lindgren@matstat.lu.se at least 1 hr in advance 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 Friday 17 May**. Also attach to the same message your R-files (or implementation in other language), in a file named `proj2.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 (forgot your id? Go to the link in the "Form groups" section).

Example: Project2 by d08xhj and d08fjh

2 U.S. county demographic information (contd.): Unemployment

We will use the same data as for computer lab 3, providing some county demographic information (CDI) for 440 of the most populous counties in the United States in years 1990–92. Each line of the dataset provides information on 14 variables for a single county. Counties with missing data were deleted from the dataset. See next page for further information. Here are the definitions for the variables considered in the population for which county demographic information (CDI) are available.

Variable	Description
id	identification number, 1–440
county	county name
state	state abbreviation
area	land area (square miles)
popul	estimated 1990 population
pop1834	percent of 1990 CDI population aged 18–34
pop65plus	percent of 1990 CDI population aged 65 years old or older
phys	number of professionally active nonfederal physicians during 1990
beds	total number of beds, cribs and bassinets during 1990
crimes	total number of serious crimes in 1990 (including murder, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft)
higrads	percent of adults (25 yrs old or older) who completed at least 12 years of school
bachelors	percent of adults (25 yrs old or older) with bachelor's degree
poors	Percent of 1990 CDI population with income below poverty level
unemployed	percent of 1990 CDI labor force which is unemployed
percapitaincome	per capita income of 1990 CDI population (dollars)
totalincome	total personal income of 1990 CDI population (in millions of dollars)
region	Geographic region classification used by the U.S. Bureau of the Census, where 1 = Northeast, 2 = Midwest, 3 = South, 4 = West

The data is available as a tab-separated txt-file, `CDI.txt`, on the course web page. Download and save it to your R working directory and then read it into R. Since region is a categorical variable, we should also turn it into a factor in R. We will also use the number of serious crimes per 1000 inhabitants:

```
cdi <- read.delim("CDI.txt")
cdi$region <- factor(cdi$region, levels = c(1, 2, 3, 4),
                     labels = c("Northeast", "Midwest", "South", "West"))
cdi$crm1000 <- 1000 * cdi$crimes / cdi$popul
```

3 High or low crime rate

We will try to determine which covariates can be used to predict if a county has a low or high crime rate (per 1000 inhabitants). Do a summary of `crm1000` in order to find the median, which has 50 % of the values below and 50 % above it. Then create a new variable, `cdi$hicrm`, taking the value 1 or TRUE if the county's crime rate lies above the median, and 0 or FALSE otherwise. We will use `hicrm` as response variable, Y, in the models.

3.1 Adults with 12 years in school

Plot `hicrm` against `higrads`, add a kernel smoother (see the lecture R-code for an example), and determine visually whether there might be a relationship. Then fit a logistic regression using `higrads` as covariate and add the estimated probabilities to the plot. Does it seem reasonable?

Report the β-estimates together with their confidence intervals and test whether the amount of adults with 12 years in school has a significant effect on the probability of having a higher than median crime rate.

Estimate the relative change in the odds (odds ratio) of having a high crime rate, with confidence interval, when the amount of `higrads` is increased by 1 (percent), and when it is increased by 10 (percent).

Use the model to predict the probability, with confidence interval, of having a high crime rate in a county where the amount of `higrads` is 65 (percent), and where it is 85 (percent).

Use the model to predict, for each of the counties, whether it would be expected to have a low or a high crime rate (predicted probability below or above 0.5) and calculate the sensitivity and specificity[1] for this model.

3.2 Region

Make a cross-tabulation between `region` and `hicrm`. Choose as reference region in your regression models the one that has the largest number of counties in it's smallest low/high category. As a tie-breaker, use the other low/high category. Why is this a good idea? *Hint*: look at how the standard errors for the log odds (ratios) are calculated in this situation.

Fit a logistic regression using `region` as (categorical) covariate and report the β-estimates together with their confidence intervals. Test whether there are any significant differences between the regions in the probability of having a high crime rate.

Estimate the odds ratios for having a high crime rate, with confidence interval, for the different regions, compared to the reference region. Also estimate the probability of having a high crime rate, with confidence interval, for the different regions, including the reference region.

Calculate the sensitivity and specificity for this model. If we are allowed to have *either* `higrads` *or* `region` as covariate, which one should we choose?

3.3 Adults with 12 years in school and Region

Fit a third model using both `higrads` and `region`. For all three models, report their AIC, BIC, Nagelkerke pseudo R^2, sensitivity and specificity. Which model is best? Use the third model and plot the squared standardized Pearson residuals and the standardized deviance residuals against the linear predictor $x\hat{\beta}$. Is there anything alarming here? Plot Cook's distance against the linear predictor, as well as against `higrads` and against `region`. Any interesting finds?

One might suspect that the effect of `higrads` is different in different regions. Fit at fourth model adding the interaction `higrads * region` to the third model, and use a Likelihood ratio test in order to determine whether this is significantly better. Also calculate AIC, BIC, Nagelkerke, sensitivity and specificity and plot the residuals and Cook's distance. Do the interaction terms improve the model?

3.4 Other variables

Find a better model using combinations of the variables `higrads`, `region`, `poors` and `phys1000 = 1000*phys/popul` (see Lab 3). You may ignore interactions. Motivate why your model is better.

[1] The proportion of high (sensitivity) and low (specificity) crime rate counties that have been correctly classified, see Lecture 9.

4 Peer review guidelines

The following guidelines for peer-review should be followed.

4.1 Questions regarding the content

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

4.2 Questions regarding the report presentation

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