Exercises for Day 3

Used datasets and R scripts can be downloaded in a ZIP archive from the Absalon *page (Applied Statistics) or from*

`https://datalab.science.ku.dk/english/course/smb-1/material/day3.zip`

Exercise 3.1: *Data representation & Validation of a probit analysis*

The purpose of this exercise is to compare three different representations of the same dataset. Read the three text files `beetle.txt`, `beetle_long.txt`, and `beetle_verylong.txt` into R using the commands

```
beetle_A <- read.delim("beetle.txt")
beetle_B <- read.delim("beetle_long.txt")
beetle_C <- read.delim("beetle_verylong.txt")
```

Have a look at them, e.g. by clicking on them in the *Environment* window, in order to verify that they encode the same dataset. Suppose we want to fit the probit model presented in the lecture to the dataset. Replace the *?*-signs by the appropriate data frame (either `beetle_A`, `beetle_B` or `beetle_C`) in the following calls to glm() in order to achieve this:

```
glm(cbind(y,n-y)~x,data=?,family=binomial(link="probit"))
glm(status~x,weights=count,data=?,family=binomial(link="probit"))
glm(status~x,data=?,family=binomial(link="probit"))
```

An additional question: What is the purpose of the *weights*-option?

Now something completely different: Above you fitted a probit analysis to three different organizations of the same dataset. Answer and discuss the following questions:

- Suppose that the above models are called `m_A`, `m_B`, and `m_C`. Verify that the same parameter estimates (and associated p-values) indeed are found in all three models, e.g. by executing the R code:

  ```
  summary(m_A)
  summary(m_B)
  summary(m_C)
  ```

- Validate the three models using cumulative residuals. This may be done using the R code:

```
library(gof)
plot(cumres(m_A))
plot(cumres(m_B))
plot(cumres(m_C))
```

  Does the validity of the model depend on the organization of the dataset?

Remark: Previously version 0.9.1 of the `gof`-package was available on the CRAN. In that version the validation was only correctly implemented for the "very long" dataset, that is `beetle_C`. In version 0.9.2 the bug was fixed for the "binomial" representation, that is `beetle_A`, and if you use the `cumres()` function on models with a `weight` option, as needed for `beetle_B`, then you get an error message stating that the weight-option is not supported.

Presently the status of the package is as follows:

- The `gof`-package has been removed from the CRAN,
- Version 1.0.1 of the `gof`-package is available from GitHub.
- Apparently the `cumres()` function gives an error if the used dataset is too large. This must be due to a bug in the program, which is rather unfortunate.

To install packages from GitHub you must have the `devtools`-package (may be installed from CRAN) and Windows user also need Rtools (available from `https://cran.r-project.org/bin/windows/Rtools/`). Thereafter you may install from GitHub like this:

```
library(devtools)
install_github("kkholst/gof")
```

Try to see if this works on your laptop. And if it doesn't, then don't despair. Then you simply don't use this methodology in your work. Actually, not using cumulated residuals was the state of the art until 15 years ago (the paper that introduced cumulated residuals for categorical regression models is from 2002).

Exercise 3.2: *Proportional odds model for the chicken gait score example.*

The purpose of this exercise is to reanalyse the chicken gait score example from Day 2 (see Day 2 lecture slides 34–39) using the proportional odds model (see Day 3 lecture slides 38–41). The R script `exercise3_2.R` contains the dataset (lines 15–23), the preparation of the dataset and repetition from Day 2 (lines 29-88), and the exercise questions (lines 91–122). Execute the lines in R script one by one and answer the questions.

Exercise 3.3: *Logistic regression.*

The data table below shows the result of the glutaraldehyde coagulation test (GLA) on 420 cows from 10 Danish dairy herds. The GLA-test is positive or negative and it is believed that the test may reflect inflammatory disease in the animal. To investigate this hypothesis the result of the test was compared with a score from a clinical examination of the animal concentrating on utter, limbs and external physical injuries. This score was 0, 1, 2 or 3 increasing with severity (infection status). The data table shows the number of cows from the different herds with the eight possible combinations of GLA-test result and clinical score.

Clinical score	0		1		2		3	
GLA	−	+	−	+	−	+	−	+
Herd 1	26	6	4	4	2	1	1	0
2	23	9	5	1	1	1	1	2
3	3	3	5	11	4	0	4	13
4	20	18	4	1	3	1	0	4
5	24	8	3	1	0	0	0	1
6	0	0	5	5	10	7	5	9
7	0	8	5	9	10	6	5	1
8	2	0	4	1	12	4	8	6
9	1	0	0	0	14	7	11	7
10	0	0	1	2	4	16	1	16

Use a logistic linear model to investigate how the GLA-test result relates to the clinical score and to the herd (dataset is available in the text file `GLA.txt`). Estimate an odds-ratio for being GLA-positive for animals with clinical score 1 (respectively 2 and 3) relative to those with clinical score 0.

See hints on the next page!

Hint 1: The explanatory variables *herd* and *clin* may be used as categorical variables (instead of continuous variables) by appropriately using the `factor()` inside the model formula.

Hint 2: If you use `factor(clin)` as a main effect in the model, then the associated parameters will be log(odds ratio) against clinical score=0.

Hint 3: Remember to backtransform the parameter estimates by the exponential function! Why?

(Data from project report by Trine Tølbøll (1999): Use of Glutaraldehyde test as an indicator of inflammatory diseases in dairy herds, KVL.)

Exercise 3.4: *Proportional odds model.*

In a survey of the usage of *Olsomarka* 365 persons were classified by how often they walk in the forest and how long they walk. The dataset listed below is taken from (Haakenstad, 1975):

	Walking distance in km					
Frequency of walks	≤ 2.5	2.5–5	5–10	10–20	≥ 20	Total
F1: Each week	14	29	80	56	16	195
F2: Each month	9	22	30	17	4	82
F3: Sometimes during the season	24	23	30	9	2	88
Total	47	74	140	82	22	365

The dataset is available in the text file `walks.txt`. Please do the following statistical analyses and present their conclusions:

- An ordinal logistic regression of distance on frequency. Is the proportional odds assumption valid? Provide estimates and confidence intervals for the odds-ratio of walking shorter distances relative to frequency group F1.

  Remark: If use want to use *distance* as the response in an ordinal regression, then this variable should be encoded as a factor. This can be done directly in the call to `clm()`:

  ```
  clm(factor(distance)~frequency,data=walks,weights=count)
  ```

- An ordinal logistic regression of frequency on distance. Try to use distance both as a factor and as a continuous covariate[1]. Is the proportional odds assumption valid? Provide estimates and confidence intervals for the odds-ratio of walking less frequently when walking distance is increased.

  Remark: The variable *frequency* is already encoded as a factor in the data frame, so it can be used as a response in `clm()` without any further ado. However, the following code doing a *multinomial regression* of *frequency* on the numerical variable *distance* used as a categorical factor[2] does not work on my laptop:

  ```
  clm(frequency~1,nominal=~factor(distance),data=walks,weights=count)
  ```

  In my opinion this must be seen as a bug in the ordinal-package. Luckily there is a fix, namely

```
clm(frequency~1,nominal=~factor(walks$distance),data=walks,weights=count)
```

- What are the differences between the three statistical analysis done above, e.g. in their interpretation and in their power to falsify the hypothesis of no association?

Remark: See script `solution3_2.R` for comments on the usage of `distance` as a continuous explanatory variable for the response `frequency`.

Exercise 3.5: *Poisson regression.*

In an investigation of the earth profile at *Mejlbjerg Hoved* the number of fine gravel particles categorized in the 4 groups *crystalline*, *sediment grains*, *chalcedony chert*, and *quartz grains* were counted in three depth levels at two different locations. The dataset listed below was taken from (Blæsild and Granfeldt, 1995):

depth:	4–11 meters		11–14 meters		18–20 meters	
location:	A	B	A	B	A	B
Crystalline	207	205	87	104	159	142
Sediment grains	48	61	23	14	19	12
Chalcedony chert	15	12	46	64	94	100
Quartz grains	30	28	133	127	41	51

[1] Possibly on logarithmic scale! The reason that it might be a good idea to use the continuous covariate on the log scale is that the odds also are modelled on a log scale!

[2] The multinomial regression is needed as the reference model in the Lack-of-Fit test for the proportional odds assumption.

The two locations (A and B) serve as blocks and should not be used in interactions. But the interaction of particle type and depth may be relevant. Perform a Poisson regression of the dataset (available in `particles.txt`) and report the estimate and its confidence interval of the relative risk of *sediment grains* against *quartz grains* at 11–14 meters. Consider also the following questions:

- What is the interpretation of the relative risk requested above? Could you think of a more easy way of computing this estimate?
- If the estimate for the relative risk can be computed directly from the dataset, what has then been achieved by the Poisson regression?

If you are unsatisfied with the model validity you might try the option `family=quasipoisson` in the glm()-call, cf. Section 10.1.2 in the R guide.

(This exercise was made on the basis of exercise 10.1 in Bo Martin Bibby: "Noter til Regressionsanalyse" (in danish).)

End of exercises.