I4-TD3
Generalized Linear Models

Problem 1

Mroz's Data on Women's Labour-Force Participation: The data in the data frame `Mroz` in the `car` package, originally employed by Mroz (1987), were used by Long (1997) to illustrate the method of logistic regression. The variables in the dataset are described in the following table, adapted from Long (and using Long's variable names). Most of the variables are self-explanatory. The variable `lwg` is the log of the wife's actual wage if she is working outside the home. For women not working outside the home, Mroz proceeded as follows: He regressed the log-wages of the working women on the other variables, and used the resulting regression equation to predict the wages of those not working outside the home.

- `lfp`: 1 if the wife is in the paid labour force, 0 otherwise
- `k5`: number of children ages 5 and younger
- `k618`: number of children ages 6 to 18
- `age`: wife's age in years
- `wc`: 1 if wife attended college, 0 otherwise
- `hc`: 1 if husband attended college, 0 otherwise
- `lwg`: log of wife's estimated wage rate
- `inc`: family income excluding wife's wages, $1000s

Following Long, perform a logistic regression of `lfp` on the other variables. Briefly (i.e., in a paragraph) summarize the results of this regression. Offer two concrete interpretations of the coefficient of `inc` in the logistic regression.

Problem 2

Powers and Xie's Data on High-School Graduation: Employing a sample of 1643 men between the ages of 20 and 24 from the U. S. National Longitudinal Survey of Youth, Powers and Xie (2000) investigate the relationship between high-school graduation and parents' education, race, family income, number of siblings, family structure, and a test of academic ability. The data set, in the file `Powers.txt` on the course web site, contains the following variables (using Powers and Xie's variable names):

- `hsgrad`: whether the respondent was graduated from high school by 1985 (`Yes` or `No`)
- `nonwhite`: whether the respondent is black or Hispanic (`Yes` or `No`)
- `mhs`: whether the respondent's mother is a high-school graduate (`Yes` or `No`)
- `fhs`: whether the respondent's father is a high-school graduate (`Yes` or `No`)
- `income`: Family income in 1979 (in $1000s) adjusted for family size
- `asvab`: standardized score on the Armed Services Vocational Aptitude Battery test
- `nsibs`: number of siblings
- `intact`: whether the respondent lived with both biological parents at age 14 (`Yes` or `No`)

The data file also contains respondent ID numbers, which are not contiguous.

a. Following Powers and Xie perform a logistic regression of `hsgrad` on the other variables in the data set. Compute a likelihood-ratio test of the omnibus null hypothesis that **none** of the explanatory variables influences high-school graduation. Then construct 95-percent confidence intervals for the coefficients of the seven explanatory variables. What conclusions can you draw from these results? Finally, offer **two** brief, but concrete, interpretations of each of the estimated coefficients of `income` and `intact`.

b. The logistic regression in the previous problem assumes that the partial relationship between the log-odds of high-school graduation and number of siblings is linear. Test for nonlinearity by fitting a model that treats `nsibs` as a factor, performing an appropriate likelihood-ratio test. In the course of working this problem, you should discover two errors in the data. Deal with the errors in a reasonable manner. Does the result of the test change?

Problem 3

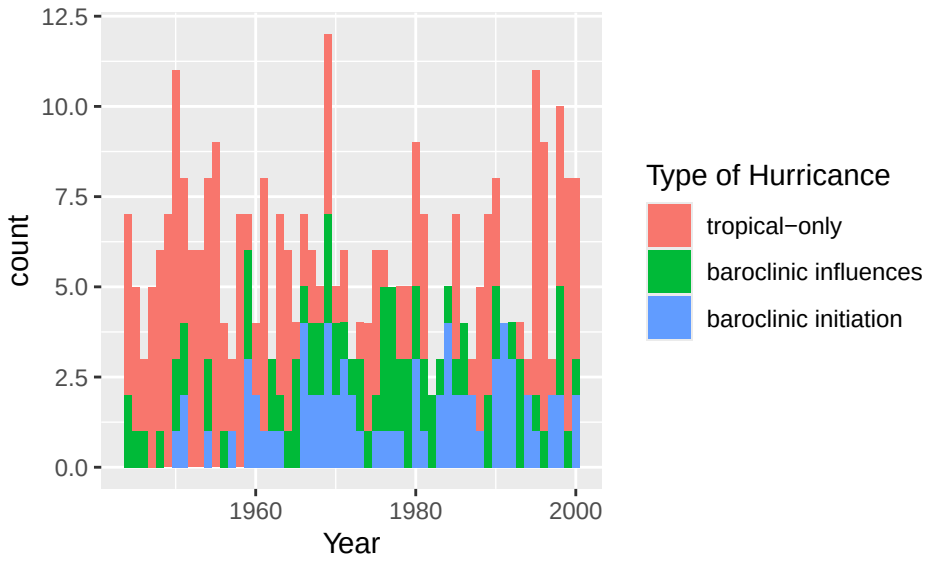
North Atlantic Hurricanes: the goal of the following exercise is to build a model that predicts the group membership of a hurricanes, either tropical or non-tropical, based on the latitude of formation.

```
library(openxlsx)
hurricanes <- read.xlsx("https://userpage.fu-berlin.de/soga/data/raw-data/hurricanes.xlsx")
str(hurricanes)
```

```
## 'data.frame':    337 obs. of  12 variables:
##  $ RowNames: chr  "1" "2" "3" "4" ...
##  $ Number  : num  430 432 433 436 437 438 440 441 445 449 ...
##  $ Name    : chr  "NOTNAMED" "NOTNAMED" "NOTNAMED" "NOTNAMED" ...
##  $ Year    : num  1944 1944 1944 1944 1944 ...
##  $ Type    : num  1 0 0 0 0 1 0 1 0 0 ...
##  $ FirstLat: num  30.2 25.6 14.2 20.8 20 29.2 16.1 27.6 21.6 19 ...
##  $ FirstLon: num  -76.1 -74.9 -65.2 -58 -84.2 -55.8 -80.8 -85.6 -95.2 -56.6 ...
##  $ MaxLat  : num  32.1 31 16.6 26.3 20.6 38 21.9 27.6 28.6 24.9 ...
##  $ MaxLon  : num  -74.8 -78.1 -72.2 -72.3 -84.9 -53.2 -82.9 -85.6 -96.1 -79.6 ...
##  $ LastLat : num  35.1 32.6 20.6 42.1 19.1 50 28.4 31.7 29.5 28.9 ...
##  $ LastLon : num  -69.2 -78.2 -88.5 -71.5 -93.9 -46.5 -82.1 -79.1 -96 -81.8 ...
##  $ MaxInt  : num  80 80 105 120 70 85 105 100 120 120 ...
```

There are 337 observations and 12 variables in the dataset. We are primarily interested in the variable `type`, which is our response variable, and the variable `FirstLat`, which corresponds to the latitude of formation, and thus is our predictor variable.

a. Produce the following plot and give the interpretation. You may consider using the function `ggplot` in 'ggplot2" package.



b. In class 0, tropical hurricanes, there are 187 observations, in class 1, baroclinic influences, there arr 77 observations and in class 3, baroclinic initiation, there are 73 observations. Since we can only deal with dichotomous (binary) data in logistic regression, please re-code the classes and assign class 1 and 3, both are being influenced by the outer tropics, the label 1 and name the new variable `Type_new`.

```
table(hurricanes$Type)
```

```
##
##   0   1   3
## 187  77  73
```

c. Fit a logistic model with `Type_new` as a response variable and `Firstlat` has predictor. Interpret the estimated coefficient of `Firstlat`. Obtain the 95% confident interval of the log-dd and the odd.

d. Predict if a hurricane is a tropical hurricane or non-tropical hurricane with the first latitudes are $10°, 23.5°$ and $30°$. You may use the `predtic()` function. Please note, that if we add the argument `type = "response"` to the function call, the `predict()` function returns the probability and not the log-odds.

e. Now consider all the variables in the data. Choose the best model (with only main effects).

f. Check if there is overdispersion in the model. Take it into account if there is.

Problem 4

Given the dataset `Default` with 4 variables. The goal is to build a logistic regression model to predict `defualt`.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
```

```
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ----------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ISLR)
data(Default)
head(Default)
```

```
##   default student   balance    income
## 1      No      No  729.5265 44361.625
## 2      No     Yes  817.1804 12106.135
## 3      No      No 1073.5492 31767.139
## 4      No      No  529.2506 35704.494
## 5      No      No  785.6559 38463.496
## 6      No     Yes  919.5885  7491.559
```

a. Explore the relationship between `defualt` and other variables (`income`, `balance`, `student`). You may create some scatterplots, and boxplots.

b. Check if two-ways and three-way interactions are significant.

c. Choose the best model (including interactions if they are significant). Interpret the coefficient estimate of model.

Problem 5

A researcher is interested in how variables, such as GRE (Graduate Record Exam scores), GPA (grade point average) and prestige of the undergraduate institution, effect admission into graduate school. The response variable, admit/don't admit, is a binary variable.

```
# loading data
admit <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")
## view the first few rows of the data
head(admit)
```

```
##   admit gre  gpa rank
## 1     0 380 3.61    3
## 2     1 660 3.67    3
## 3     1 800 4.00    1
## 4     1 640 3.19    4
## 5     0 520 2.93    4
## 6     1 760 3.00    2
```

a. Analyze the data with main effects by using logit, probit and complementary log-log link-functions. Compare the models and indicate the best one that fits with the data the most.

b. In case of the logit model, interpret the coefficient estimate of `gre`, `gpa` and `rank`.

c. Predict the probability of admission at each value of rank hold `gre` and `gpa` at their mean.

Problem 6

The data set looks at how many warp breaks occurred for different types of looms per loom, per fixed length of yarn.

```
library(datasets)
data = warpbreaks
head(data)
```

```
##   breaks wool tension
## 1     26    A       L
## 2     30    A       L
## 3     54    A       L
## 4     25    A       L
## 5     70    A       L
## 6     52    A       L
```

- `breaks`: the number of breaks
- `wool`: the type of wool (A or B)
- `tension`: the level of tension (L, M, H)

a. We are interested in modelling `breaks` on other variables. Do EDA on the dataset.
b. Modelling `breaks` and obtain the best model. Is there any sign of overdispersion? If so, account for it. Interpret the coefficient estimates.

Problem 7

The data contains the number of cancer in a variable named `cases`. We are interested in modelling the number of cases.

```
library(ISwR)
data(eba1977)
head(eba1977)
```

```
##         city   age  pop cases
## 1 Fredericia 40-54 3059    11
## 2    Horsens 40-54 2879    13
## 3    Kolding 40-54 3142     4
## 4      Vejle 40-54 2520     5
## 5 Fredericia 55-59  800    11
## 6    Horsens 55-59 1083     6
```

a. Do EDA on the dataset.
b. Modeling `cases` and obtain the best model with proper transformation of variable `pop`. Is there any sign of overdipersion? Interpret coefficient estimates of `age`.
c. Predict a future observation where he/she is from Kolding, 42 years old, with the number of inhabitant of 1000.

Problem 8

There are 8 variables in the dataset. The variable 'affairs' is the number of extramarital affairs in the past year and is our response variable. We will include as covariates the variables 'gender', 'age', 'yearsmarried', 'children', 'religiousness', 'education' and 'rating' in our analysis. 'religiousness' ranges from 1 (anti) to 5 (very) and 'rating' is a self rating of the marriage, ranging from 1 (very unhappy) to 5 (very happy).

```
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survival'
```

```
## The following object is masked from 'package:ISwR':
##
##     lung
```

```
data(Affairs)
data = Affairs[ ,-8]
str(data)
```

```
## 'data.frame':    601 obs. of  8 variables:
##  $ affairs      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ gender       : Factor w/ 2 levels "female","male": 2 1 1 2 2 1 1 2 1 2 ...
##  $ age          : num  37 27 32 57 22 32 22 57 32 22 ...
##  $ yearsmarried : num  10 4 15 15 0.75 1.5 0.75 15 15 1.5 ...
##  $ children     : Factor w/ 2 levels "no","yes": 1 1 2 2 1 1 1 2 2 1 ...
##  $ religiousness: int  3 4 1 5 2 2 2 2 4 4 ...
##  $ education    : num  18 14 12 18 17 17 12 14 16 14 ...
##  $ rating       : int  4 4 4 5 3 5 3 4 2 5 ...
```

a. Fitted a Poisson model and perform a model specification.
b. Check if there is over dispersion in the model. If there is, fitted a model that account for overdispersion. Interpret the coefficient estimate.
c. Check if model assumptions **count response**, **independent events** and **constant variance** are satisfied by using `plot(model, which = 1:6)`.