# Series 8

**1.** The dataset `heart.dat` contains data for 99 people grouped by age. In each age group the total number of individuals ($m_i$) is known, as well the number of those with symptoms of heart disease ($N_i$). The goal of this exercise is to estimate the probability of having such symptoms as a function of age using logistic regression.

The data is located at `http://stat.ethz.ch/Teaching/Datasets/heart.dat`.

**a)** In contrast to the binary classification example in the lecture notes (page 54), the response variable $N$ has not a Bernoulli, but a binomial distribution: $N_1, \ldots, N_n$ independent, $N_i \sim \text{Binomial}(m_i, \pi(x_i))$.
Show that the log-likelihood is in this case

$$\ell(\beta; (x_1, m_1, N_1), \ldots, (x_n, m_n, N_n)) = \sum_{i=1}^{n} \left[ \log \binom{m_i}{N_i} + N_i g(\beta; x_i) - m_i \log \left(1 + e^{g(\beta; x_i)}\right) \right],$$

where $g(\beta; x) = \beta_0 + \beta_1 x$ is the model function for the logistic transform of $\pi(x)$ (see section 6.4 of the lecture notes).

**b)** Write an R function `neg.ll(beta, data)` that calculates the *negative* log-likelihood

$$-\ell(\beta; (x_1, m_1, N_1), \ldots, (x_n, m_n, N_n))$$

that you derived in task a). `beta` is a vector with two entries $\beta_0$ and $\beta_1$, and `data` is a data frame with columns `age`, `m` and `N` (as in `heart.dat`).
Make a contour plot of the negative log-likelihood of the `heart` dataset in the range $-10 \le \beta_0 \le 10$, $-1 \le \beta_1 \le 1$.
**R hint:** use a function call like

```
> contour(beta0.grid, beta1.grid, neg.ll.values)
```

`beta0.grid` and `beta1.grid` are equidistant grids of values of $\beta_0$ and $\beta_1$ in the region of interest; use, e.g.

```
> beta0.grid <- seq(-10, 10, length = 101)
```

`neg.ll.values` is a matrix of negative log-likelihood values for the different values of $\beta_0$ and $\beta_1$.

**c)** Estimate the parameters $\beta_0$ and $\beta_1$ of the model function (see task a)) using the R function `glm`. Does age influence this probability in a significant way? How do you interpret the sign of the coefficient of `age`?
Compare the estimates from `glm` with estimates you get when minimizing the negative log-likelihood function you implemented in task b).
**R hint:** the logistic regression model can be fitted by using the function call

```
> fit <- glm(cbind(N, m - N) ~ age, family = binomial, data = heart)
```

Binomial responses $N_i \sim \text{Bin}(m_i, \pi_i)$ for $m_i > 1$ should be entered as a (two-column) matrix, with the number of "successes" ($N_i$) in the first column and the number of "failures" ($m_i - N_i$) in the second.
To minimize your function `neg.ll` from task b), use

```
> optim(c(0, 0), neg.ll, data = heart)
```

The first argument is the start value used for numerical optimization.

**d)** Plot the probability estimate against age. At what age would you expect 10%, 20%, ..., 90% of people to have symptoms of heart disease? Discuss your results.
**R hint:** you can obtain probability estimates at arbitrary ages `new.age` by using the function call

```
> predict(fit, newdata = data.frame(age = new.age), type = "response")
```

**2.** In this exercise we are investigating the ozone dataset which you have already seen in the lecture. The dataset `ozone` is available in numerous R-packages, e.g. in the package `gss`. You can load it with `data(ozone, package = "gss")`. If you do not have access to the package, you can get the data from `http://stat.ethz.ch/Teaching/Datasets/ozone.dat`. A short description of the variables is available at `help(ozone, package = "gss")`.

**R-Hints:**
Use the R-skeleton available on the course's webpage.

**a)** Get an overview of the data with `pairs()`. You should take the log of the response (`upo3`) and remove the outlier in the predictor `wdsp`. Explain why you should do this.
**R-Hints:**
The transformation can be done using:

```
ozone$logupo3 <- log(ozone$upo3)
d.ozone <- subset(ozone, select=-upo3)
```

**b)** We want to compare linear regression with an additive model (7.2 in the lecture notes).
In order to better fit the model, we allow polynomial fits of the data in the predictive variables up to degree d for linear regression, i.e.
$y_i = \beta_0 + \beta_{1,1}x_{i1} + \beta_{1,2}x_{i1}^2 + ...\beta_{1,d}x_{i1}^d + ... + \beta_{p,1}x_{ip} + \beta_{p,2}x_{ip}^2 + ...\beta_{p,d}x_{ip}^d + \epsilon_i$.
This means instead of having $p+1$ predictive variables we have $pd+1$.
Here we fit the 5 models with linear regression for the degrees up to 5. We also fit the data with an additive model as programmed in R.

**c)** Plot the respective fits. Use function `termplot()` for the linear models. What do you observe? Which polynomial degree would you choose? Use `mult.fig()` and the default `plot()` function for the additive model. Compare the additive model plot to the linear model ones. What do you observe?

**d)** Now we want to perform model selection on our models. We fix $\sigma$ as the estimated scale from the linear fit of degree 5. Then we calculate Mallows' $C_p$ statistic for all 6 models. Which of the linear models do we prefer? And which of all the models?

**R-Hints:**
The skeleton for task 2 can be found on the webpage of the course.
Use the function `poly()` in the formula when working with `lm()`, `gam()` (to fit an additive model) can be found in the package `mgcv`. Usage: `gam(formula, data)`. `formula` must be of the form `logupo3 ~ s(vdht) + s(wdsp) + ...`
Make use of `summary()` to get an overview of your `gam()` output.

**Preliminary discussion:** Friday, May 06.

**Deadline:** Friday, May 13.