Interactive lecture - first week

Theoretical questions - with and without use of R

Problem 1: Model assumptions

a) What are the model assumptions for a binary regression?
b) Which link function and response function is used for the logit model?
c) What is the difference between the logit model and a logistic regression?

↑ you say this is the same in the module pages! ("logit model AKA logistic regression)

Problem 2: Log-likelihood.

a) What is the definition of the log-likelihood?

b) For the logit model the log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{G} [\ln \binom{n_j}{y_j} + y_j \ln \pi_j - y_j \ln(1-\pi_j) + n_j \ln(1-\pi_j)]$$

for grouped data. Explain. Explain what? Same as in 2d?

c) Write the version of the loglikelihood for individual data (i.e. $n_j = 1$ and $G = n$).

d) Where is $\boldsymbol{\beta}$ in thi~~s~~e expression? Rewrite this from 2c to be a function of $\boldsymbol{\beta}$.

e) Why can we ignore the ~~normalization~~ normalizing constant (what is the ~~normalization~~ constant?) in the case of $n_j = 1 \; \forall j$? Considering what the log-likelihood is used for, why can we ignore the normalization constant in all cases (i.e., also when $n_j \neq 1$)?

f) What does this graph of l look like as a function of $\boldsymbol{\beta}$ for the beetle data? To aid you in answering this we look at the loglikelihood for the beetle data: - what are these doing in this exercise now?

The beetle data has only one covariate (in addition to the intercept) - so this means that we have $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Look at the following code and explain what is done - remark: we have used the $n_i = 1$ version of the loglikelihood here.

```
library(investr)
library(ggplot2)
ldose = rep(investr::beetle$ldose, investr::beetle$n)
y = NULL
for (i in 1:8) y = c(y, rep(0, investr::beetle$n[i] - investr::beetle$y[i]),
    rep(1, investr::beetle$y[i]))
beetleds = data.frame(killed = y, ldose = ldose)
dim(beetleds)
```

```
## [1] 481   2
```

```
loglik <- function(par, args) {
    y <- args$y
    x <- args$x
    n <- args$n
    res <- sum(y * x %*% par - n * log(1 + exp(x %*% par)))
    return(res)
}

loglik(c(1, 1), args = list(y = beetleds$killed, x = cbind(rep(1, nrow(beetleds)),
    beetleds$ldose), n = rep(1, nrow(beetleds))))
```

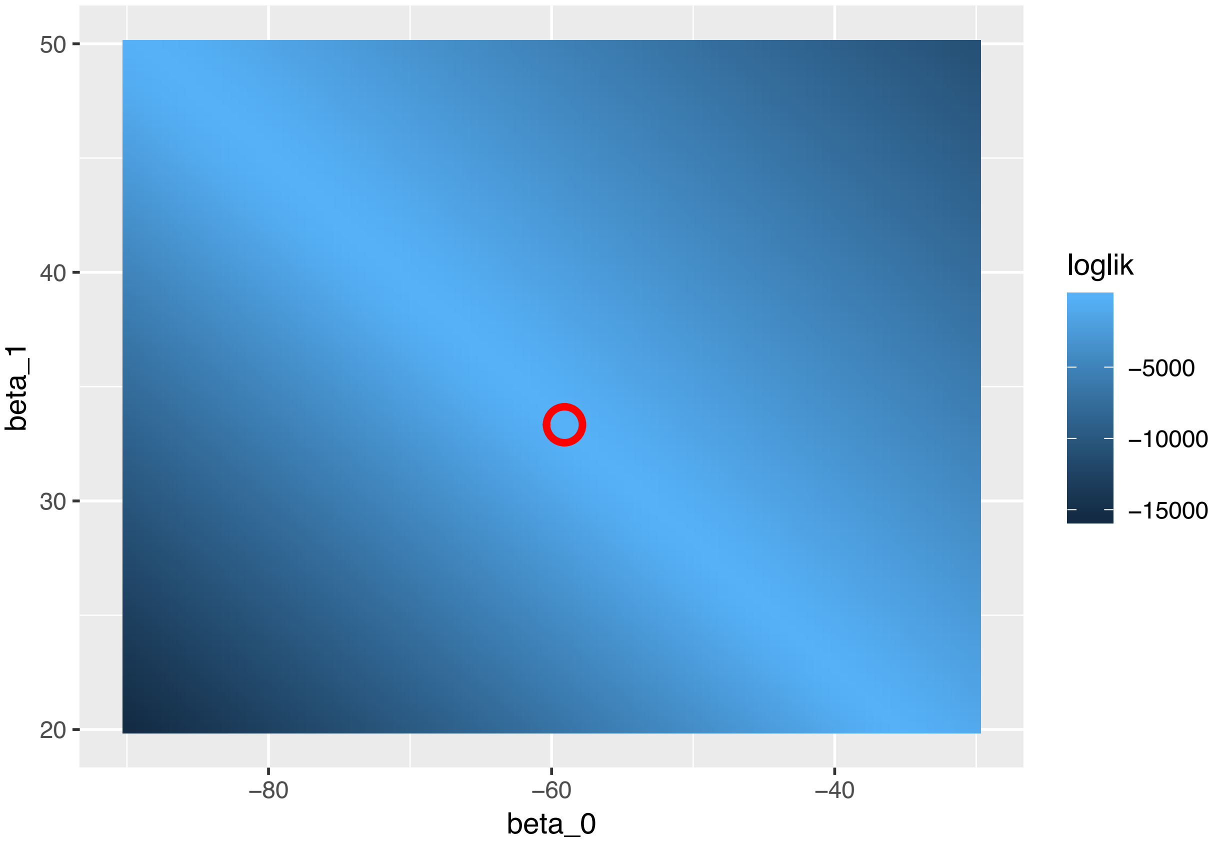
```
## [1] -549.2543
```

```
loglikmat <- matrix(NA, nrow = 100, ncol = 100)
loglikframe <- data.frame()
beta_0 <- seq(-90, -30, length.out = 100)
beta_1 <- seq(20, 50, length.out = 100)

for (i in 1:length(beta_0)) {
    for (j in 1:length(beta_1)) {
        loglikmat[i, j] <- loglik(c(beta_0[i], beta_1[j]), args = list(y = beetleds$killed,
            x = cbind(rep(1, nrow(beetleds)), beetleds$ldose), n = rep(1, nrow(beetleds))))
        loglikframe <- rbind(loglikframe, c(beta_0[i], beta_1[j], loglikmat[i,
            j]))
    }
}
names(loglikframe) <- c("beta_0", "beta_1", "loglik")
head(loglikframe)
```

```
##   beta_0   beta_1    loglik
## 1    -90 20.00000 -15545.83
## 2    -90 20.30303 -15384.56
## 3    -90 20.60606 -15223.28
## 4    -90 20.90909 -15062.01
## 5    -90 21.21212 -14900.73
## 6    -90 21.51515 -14739.46
```

```
ggplot(data = loglikframe, mapping = aes(x = beta_0, y = beta_1, z = loglik)) +
    geom_raster(aes(fill = loglik)) + geom_point(data = loglikframe[which.max(loglikframe$loglik),
    ], mapping = aes(x = beta_0, y = beta_1), size = 5, col = "red", shape = 21,
    stroke = 2) + scale_shape(solid = FALSE)
```



Comments to the code: for the loglik function we have two arguments: par= the parameters to be estimated, and args=a list with data. The reason for only having these two arguments is that it is easier to use in the optimization of the loglikelihood to find the ML estimates.

function

Problem 3: Score function

a) What is the definition of the score function? What is the dimension of the score function?

b) Derive the score function for the logit model (individual data). The result should be

$$s(\boldsymbol{\beta}) = \sum_{i=1}^{n} \mathbf{x}_i(y_i - \pi_i) = \sum_{i=1}^{n} \mathbf{x}_i(y_i - \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1+\exp(\mathbf{x}_i^T \boldsymbol{\beta})})$$

c) What do we need the score function for?

Problem 4: Fisher information.

a) What is the definition of the expected and the observed Fisher information matrix?

b) What is the role of these matrices in ML estimation?

c) For the logit model with grouped data the expected and the observed Fisher information matrix are equal and given as

$$F(\boldsymbol{\beta}) = \sum_{j=1}^{G} \mathbf{x}_j \mathbf{x}_j^T n_j \pi_j (1-\pi_j)$$

Where is $\boldsymbol{\beta}$ in this expression? d) Write the version of the loglikelihood for individual data (i.e. $n_j = 1$ and $G = n$).

Fisher information you mean?

Problem 5: Maximum likelihood

To find the ML estimate for $\boldsymbol{\beta}$ we may either use the function `glm` or optimize the log-likelihood manually. We will do both.

a) First we use the `glm` function in R, and we also check that the individual and the grouped data give the same parameter estimates for the $\boldsymbol{\beta}$. What is the exercise here?

```
# the beetle.ds was made above
fitind = glm(killed ~ ldose, family = "binomial", data = beetleds)  # individual data
summary(fitind)
```

```
##
## Call:
## glm(formula = killed ~ ldose, family = "binomial", data = beetleds)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4922  -0.5986   0.2058   0.4512   2.3820
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## ldose         34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 645.44  on 480  degrees of freedom
## Residual deviance: 372.47  on 479  degrees of freedom
## AIC: 376.47
##
## Number of Fisher Scoring iterations: 5
```

```
fitgrouped = glm(cbind(y, n - y) ~ ldose, family = "binomial", data = investr::beetle)  # grouped data.
summary(fitgrouped)
```

```
##
## Call:
## glm(formula = cbind(y, n - y) ~ ldose, family = "binomial", data = investr::beetle)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5941  -0.3944   0.8329   1.2592   1.5940
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -60.717      5.181  -11.72   <2e-16 ***
## ldose         34.270      2.912   11.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 284.202  on 7  degrees of freedom
## Residual deviance:  11.232  on 6  degrees of freedom
## AIC: 41.43
##
## Number of Fisher Scoring iterations: 4
```

b) What is the default convergence criterion for the glm? (Note: IRWLS used in `glm` - more in Module 5.)

c) We implemented the log-likelihood as a function in item 2 above. Now we will use this together with the `optim` function on the beetle data set to optimize the loglikelihood. what is the exercise here?

Har lagt til kode for å regne ut stderrs, også se LF

```r
loglik_gr <- function(par, args) {

    y <- args$y
    x <- args$x
    n <- args$n

    res <- y %*% x - t(t(n * x) %*% ((1 + exp(-x %*% par))^(-1)))
    return(res)
}

optim(c(-60, 30), fn = loglik, gr = loglik_gr, args = list(y = beetleds$killed,
    x = cbind(rep(1, nrow(beetleds)), beetleds$ldose), n = rep(1, nrow(beetleds))),
    control = list(fnscale = -1))
```

```
## $par
## [1] -60.73001  34.27737
##
## $value
## [1] -186.2354
##
## $counts
## function gradient
##       57       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

Problem 6: Interpreting results

a) Interpret the estimated $\boldsymbol{\beta}$´s. Odds ratio is useful for this.
b) Plot the predicted probability of a beetle dying against the dosage and discuss what you see. (Yes, since this is the last question you may try to program by yourself!)