TMA4315: Project 1

Jim Totland, Martin Gudahl Tufte

9/8/2021

Litt usikker på hva slags notasjon vi skal bruke, f. eks. boldface for vektorer eller ikke? Bare si ifra hvis du vil ha noe spesifikit:)

Problem 1

a)

Since the response variables $y_i \sim \text{Bernoulli}(p_i)$, where $p_i = \Pr(y_i = 1 \mid x_i) = \Phi(x_i^T \beta)$, the conditional mean is given by $\mathrm{E} y_i = p_i$, which is connected to the covariates via the following relationship:

$$x_i^T \beta =: \eta_i = \Phi^{-1}(p_i),$$

which implies that $p_i = \Phi(\eta_i)$. This results in the likelihood function

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \Phi(\eta_i)^{y_i}(1-\Phi(\eta_i))^{1-y_i}. \end{aligned}$$

Thus, the log-likelihood becomes

$$l(\beta) := \ln(L(\beta)) = \sum_{i=1}^{n} y_i \ln(\Phi(\eta_i)) + (1-y_i)\ln(1-\Phi(\eta_i)) = \sum_{i=1}^{n} l_i(\beta).$$

To find the score function, we calculate

$$\begin{aligned} \frac{\partial l_i(\beta)}{\partial \beta} &= \frac{y_i}{\Phi(\eta_i)} \frac{\partial \Phi(\eta_i)}{\partial \beta} - \frac{1-y_i}{1-\Phi(\eta_i)} \frac{\partial \Phi(\eta_i)}{\partial \beta} \\ &= \frac{y_i}{\Phi(\eta_i)} \phi(\eta_i) x_i - \frac{1-y_i}{1-\Phi(\eta_i)} \phi(\eta_i) x_i \\ &= \frac{y_i(1-\Phi(\eta_i)) - (1-y_i)\Phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))} \phi(\eta_i) x_i \\ &= \frac{y_i - \Phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))} \phi(\eta_i) x_i. \end{aligned}$$

Consequently, the score function is given by

$$s(\beta) = \sum_{i=1}^{n} \frac{y_i - \Phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))} \phi(\eta_i) x_i.$$

Next, we find the expected Fisher information, $F(\beta)$. We find it by using the result

$$\begin{aligned}
F(\beta) &= \mathrm{Var}(s(\beta)) = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{y_i - \Phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))}\phi(\eta_i)x_i\right) \\
&= \sum_{i=1}^{n} \underbrace{\left[\frac{\phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))}\right]^2}_{=:\xi_i} \mathrm{Var}(y_i x_i) = \sum_{i=1}^{n} \xi_i x_i \mathrm{Var}(y_i) x_i^T \\
&= \sum_{i=1}^{n} \xi_i p_i (1-p_i) x_i x_i^T \\
&= \sum_{i=1}^{n} \frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))} x_i x_i^T,
\end{aligned}$$

Where in the third equality we have used that the y_i's are independent. The expected Fisher information can also be verified to have this expression by the relationship

$$F(\beta) = \sum_{i=1}^{n} \frac{h'(\eta_i)^2}{\mathrm{Var}(y_i)} x_i x_i^T,$$

where $h'(\eta_i) = \Phi'(\eta_i) = \phi(\eta_i)$ and $\mathrm{Var}(y_i) = p_i(1-p_i) = \Phi(\eta_i)(1-\Phi(\eta_i))$.

b) The expected Fisher information is given by

$$F(\beta) = \sum_{i=1}^{n} \frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))} x_i x_i^T = x^T W x,$$

where $W = \mathrm{diag}\left(\frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))}\right)$.

The Fisher scoring algorithm states that the next iterate is given by

$$\beta^{(t+1)} = \beta^{(t)} + F(\beta(t))^{-1} s(\beta(t)).$$

Inserting the expected Fisher information and the score function we get

$$\beta^{(t+1)} = (x^T W^{(t)} x)^{-1} x^T W^{(t)} \tilde{y}^{(t)},$$

where the working response vector $\tilde{y}^{(t)}$ has element i given by

$$\tilde{y}_i^{(t)} = x_i^T \beta^{(t)} + \frac{y_i - h(x_i^T \beta(t))}{h'(x_i^T \beta(t))} = \eta_i^{(t)} + \frac{y_i - \Phi(\eta_i^{(t)})}{\phi(\eta_i^{(t)})}.$$

The deviance is defined as

$$D = -2\ l(\hat{\beta}) + 2\ l(\text{saturated model})$$

$D = -2\ l(\hat{\beta}) + 2\ l(\text{sturated model})$.

Implementing myglm in R:

```r
Phi <- function(x) return (pnorm(x))
phi <- function(x) return (dnorm(x))


myglm <- function(formula, data, start = NULL){

  # response variable
  resp <- all.vars(formula)[1]
  y <- as.matrix( data[resp] )


  # model matrix
  X <- model.matrix(formula, data)
  n <- dim(X)[1]
  p <- dim(X)[2]


  # starting beta
  if (is.null(start)){
    beta = rep(0, p)
  }
  else {
    beta = start
  }


  # Fisher scoring algorithm
  max_iter <- 50
  tol <- 1e-10
  iter <- 0
  rel.err <- Inf

  while (rel.err > tol & iter < max_iter){
    # calculate eta, y tilde, W
    eta <- X %*% beta
    y.tilde <- eta + (y - Phi(eta)) / (phi(eta))
    W <- diag( as.vector(phi(eta)^2 / (Phi(eta)*(1-Phi(eta)))), n, n )


    # update beta
    A <- t(X) %*% W %*% X
    b <- t(X) %*% W %*% y.tilde
    beta.new <- solve(A, b)

    iter <- iter + 1
    rel.err <- max(abs(beta.new - beta) / abs(beta.new))
    beta <- beta.new
  }




  # remains to find the coefficients matrix, deviance and estimated variance matrix
```

```r
  F.inv <- solve(A)

  Estimate <- beta
  Std.Error <- sqrt(diag(F.inv))

  return (list("coefficients" = data.frame(Estimate, Std.Error),
               "deviance" = NULL,
               "vcov" = F.inv)
         )
}
```

c)

Simulation of 1000 bernoulli draws with a random probability.

```r
# probability
x = runif(1000, 0, 1)
# draw n bernoulli with prob x
y <- rbinom(1000, 1, x)

df <- data.frame(y, x)



### fit using glm
model <- glm(y ~ x, family = binomial(link = "probit"), data = df)

# beta
model$coefficients
```

```
## (Intercept)           x
##   -1.509700    2.802501
```

```r
# se for beta
summary(model)
```

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "probit"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0901  -0.8503  -0.3966   0.8311   2.2361
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.50970    0.09915  -15.23   <2e-16 ***
## x            2.80250    0.17157   16.33   <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: 1382.2  on 999  degrees of freedom
## Residual deviance: 1068.5  on 998  degrees of freedom
```

```
## AIC: 1072.5
##
## Number of Fisher Scoring iterations: 4
```

```r
# vcov
vcov(model)
```

```
##              (Intercept)           x
## (Intercept)  0.009831593 -0.01522603
## x           -0.015226030  0.02943661
```

```r
# deviance
model$deviance
```

```
## [1] 1068.518
```

```r
### fit using myglm
mymodel <- myglm(y ~ x, data = df)

# beta
mymodel$coefficients
```

```
##              Estimate  Std.Error
## (Intercept) -1.509701 0.09915554
## x            2.802503 0.17157328
```

```r
# vcov
mymodel$vcov
```

```
##             (Intercept)           x
## (Intercept)  0.00983182 -0.01522643
## x           -0.01522643  0.02943739
```

```r
# deviance
mymodel$deviance
```

```
## NULL
```

Problem 2

a)

```r
#install.packages("ISwR")
library(ISwR) # Install the package if needed
data(juul)
juul$menarche <- juul$menarche - 1
juul.girl <- subset(juul, age>8 & age<20 & complete.cases(menarche))
```

```r
?juul
```

```
## starting httpd help server ... done
```

```r
head(juul.girl)
```

```
##       age menarche sex igf1 tanner testvol
## 167  8.96        0   2   NA      1      NA
## 343 13.01        0   2  682      2      NA
## 743  8.03        0   2   NA      1      NA
## 744  8.08        0   2   NA      1      NA
```

```
## 745  8.13        0   2  210      1      NA
## 746  8.17        0   2  564     NA      NA
```

```r
model <- glm(menarche ~ age, family=binomial(link="probit"), data= juul.girl)
anova(model, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: probit
##
## Response: menarche
##
## Terms added sequentially (first to last)
##
##
##      Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                   518     719.39
## age   1      522       517     197.39 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The low p-value suggests that age has an effect on the response variable.

b)

Relating to the `juul` data set, we define for each observation/individual

$$y_i = \begin{cases} 0, & \text{if menarche has occured.} \\ 1, & \text{if menarche has not occured.} \end{cases}$$

and t_i as the age at the time of examination, which corresponds to `age` in the data set. Let $T_i \sim N(\mu, \sigma)$, where T_i is the time until menarche occurs for the i'th individual. Furthermore, let

$$\begin{aligned} \pi_i &:= P(y_i = 1) = P(T_i \leq t_i) \\ &= P\left(\frac{T_i - \mu}{\sigma} \leq \frac{t_i - \mu}{\sigma}\right) = \Phi\left(\frac{t_i - \mu}{\sigma}\right) \end{aligned}$$

This, in turn, gives

$$\Phi^{-1}(pi_i) = -\frac{\mu}{\sigma} + \frac{1}{\sigma} t_i = \beta_0 + \beta_1 t_i,$$

where $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$.