TMA4315: Project 1

Jim Totland, Martin Gudahl Tufte

9/8/2021

Problem 1

a)

Since the response variables $y_i \sim \text{Bernoulli}(\pi_i)$, where $\pi_i = \Pr(y_i = 1 \mid \boldsymbol{x}_i)$. The conditional mean is given by $\text{E}y_i = \pi_i$, which is connected to the covariates via the following relationship:

$$\boldsymbol{x}_i^T \boldsymbol{\beta} =: \eta_i = \Phi^{-1}(\pi_i),$$

or equivalently: $\pi_i = \Phi(\eta_i)$. This results in the likelihood function

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

Thus, the log-likelihood becomes

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

To find the score function, we calculate

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

Consequently, the score function is given by

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

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

$$
\begin{aligned}
F(\boldsymbol{\beta}) &= \mathrm{Var}(\boldsymbol{s}(\beta)) = \mathrm{Var}\left(\sum_{i=1}^{n} \frac{y_i - \Phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))}\phi(\eta_i)\boldsymbol{x}_i\right) \\
&= \sum_{i=1}^{n}\left[\frac{\phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))}\right]^2 \mathrm{Var}(y_i\boldsymbol{x}_i) = \sum_{i=1}^{n}\left[\frac{\phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))}\right]^2 \boldsymbol{x}_i \mathrm{Var}(y_i)\boldsymbol{x}_i^T \\
&= \sum_{i=1}^{n}\left[\frac{\phi(\eta_i)}{\Phi(\eta_i)(1-\Phi(\eta_i))}\right]^2 \pi_i(1-\pi_i)\boldsymbol{x}_i\boldsymbol{x}_i^T = \sum_{i=1}^{n} \frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1-\Phi(\eta_i))}\boldsymbol{x}_i\boldsymbol{x}_i^T,
\end{aligned}
$$

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 general relationship

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

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

b)

The expected Fisher information is given by

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

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

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

Inserting the expected Fisher information and the score function we get Jeg tror det du beskriver under er Iterated Reweighted Least Squares, som er en variant av Fisher Scoring algo. Tror bi kun trenger formelen over for å implementere myglm.

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

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

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

litt om deviance her. Litt ullent for meg hva en saturated model er, men dette gir riktig svar i c). We also need the deviance, which is defined as

$$
D = 2(l_{\text{saturated}} - l(\hat{\boldsymbol{\beta}})).
$$

When we fit a parameter for each data point (which is the case for the saturated model), the result for the Bernoulli distribution is that $\hat{\pi}_i = y_i$. This means that the likelihood function of the saturated model is given by

$$
L_{\text{saturated}} = \prod_{i=1}^{n} \hat{\pi}_i^{y_i}(1-\hat{\pi}_i)^{1-y_i} = \prod_{i=1}^{n} y_i^{y_i}(1+y_i)^{1-y_i} = 1,
$$

Where we have used $0^0 = 1$. Consequently, the log-likelihood $l_{\text{saturated}} = \ln(1) = 0$ and the deviance becomes $-2l(\hat{\boldsymbol{\beta}})$. Next follows the Implementation of `myglm` in 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
  }

  # Calculating vcov and deviance.
  F.inv <- solve(A)
```

```
  std.Error <- sqrt(diag(F.inv))

  eta = X %*% beta
  deviance = -2 * sum(y*log(pnorm(eta)) + (1 - y)*log(1 -pnorm(eta)))

  return (list("coefficients" = data.frame(beta, std.Error),
            "deviance" = deviance,
            "vcov" = F.inv))
}
```

c)

Simulation of 1000 Bernoulli draws with a random probability.

```
# 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
##   -1.490684    2.877741
```

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

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "probit"), data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1992  -0.7996  -0.4045   0.7900   2.2247
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.49068    0.09543  -15.62   <2e-16 ***
## x            2.87774    0.17017   16.91   <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: 1383.2  on 999  degrees of freedom
## Residual deviance: 1042.8  on 998  degrees of freedom
## AIC: 1046.8
##
## Number of Fisher Scoring iterations: 4
```

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

```
##              (Intercept)           x
## (Intercept)  0.009107042 -0.01435117
## x           -0.014351166  0.02895826
```

```
# deviance
model$deviance
```

```
## [1] 1042.767
```

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

```
##                  beta  std.Error
## (Intercept) -1.490685 0.09543177
## x            2.877743 0.17017302
```

```
# vcov
mymodel$vcov
```

```
##              (Intercept)           x
## (Intercept)  0.009107223 -0.01435148
## x           -0.014351476  0.02895886
```

```
# deviance
mymodel$deviance
```

```
## [1] 1042.767
```

Problem 2

a)

```
#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))
```

```
mod.probit <- glm(menarche ~ age, family=binomial(link="probit"), data= juul.girl)
anova(mod.probit, 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 $\beta_1 = 1/\sigma$.

c)

```
mod.logit <- glm(menarche ~ age, family = binomial(link = 'logit'), data = juul.girl)
mod.logit$coefficients[2]
```

```
##      age
## 1.517289
```

To show find the distribution of the T_i's, we start with the cumulative distribution:

$$\Pr(T_i \leq t_i) = \Pr(y_i = 1 \mid t_i) = \pi_i = \frac{1}{1 + e^{-\eta_i}}.$$

The pdf of T_i is then given as

$$\begin{aligned} f_{T_i}(t_i) &= \frac{\mathrm{d}}{\mathrm{d}t_i}\left(\frac{1}{1 + e^{-\eta_i}}\right) = \frac{\beta_1 e^{-\beta_0 - \beta_1 t_i}}{(1 + e^{-\beta_0 - \beta_1 t_i})^2} \\ &= \frac{e^{-(t_i - (-\beta_0/\beta_1))/(1/\beta_1)}}{1/\beta_1 (1 + e^{-(t_i - (-\beta_0/\beta_1))/(1/\beta_1)})^2} = \frac{e^{-(t_i - \mu)/s}}{s(1 + e^{-(t_i - \mu)/s})^2}. \end{aligned}$$

This is the logistic distribution, with parameters $\mu = -\beta_0/\beta_1$ and $s = 1/\beta_1$, where we have used the parametrization from Wikipedia. We compute estiamtes of the mean and variance from the estimates of β_0 and β_1 in the output above. An estimate of the mean is then given by $\mathrm{E}(T_i) = -\beta_0/\beta 1 \approx 13.1901147$ and an estimate of the variance is given by $\mathrm{Var}(T_i) = s^2\pi^2/3 = \pi^2/(3\beta_1^2) \approx 1.4290323$.