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 $\mathrm{E}y_i = \pi_i$, 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

$$\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - \Phi(\eta_i)}{\Phi(\eta_i)(1 - \Phi(\eta_i))} \phi(\eta_i)\boldsymbol{x}_i = X^T D \Sigma^{-1} (\boldsymbol{y} - \boldsymbol{\mu}),$$

where $D = \text{diag}(\phi(\eta_i))$ and $\Sigma = \text{diag}(\text{Var}(y_i)) = \text{diag}(\Phi(\eta_i)(1 - \Phi(\eta_i)))$. Next, we find the expected Fisher information, $F(\boldsymbol{\beta})$. We find it by using the result

$$F(\boldsymbol{\beta}) = \text{Var}(\boldsymbol{s}(\beta)) = \text{Var}\left(\sum_{i=1}^{n} \frac{y_i - \Phi(\eta_i)}{\Phi(\eta_i)(1 - \Phi(\eta_i))} \phi(\eta_i) \boldsymbol{x}_i\right)$$

$$= \sum_{i=1}^{n} \left[\frac{\phi(\eta_i)}{\Phi(\eta_i)(1 - \Phi(\eta_i))}\right]^2 \text{Var}(y_i \boldsymbol{x}_i) = \sum_{i=1}^{n} \left[\frac{\phi(\eta_i)}{\Phi(\eta_i)(1 - \Phi(\eta_i))}\right]^2 \boldsymbol{x}_i \text{Var}(y_i) \boldsymbol{x}_i^T$$

$$= \sum_{i=1}^{n} \left[\frac{\phi(\eta_i)}{\Phi(\eta_i)(1 - \Phi(\eta_i))}\right]^2 \pi_i(1 - \pi_i) \boldsymbol{x}_i \boldsymbol{x}_i^T = \sum_{i=1}^{n} \frac{\phi(\eta_i)^2}{\Phi(\eta_i)(1 - \Phi(\eta_i))} \boldsymbol{x}_i \boldsymbol{x}_i^T,$$

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

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

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

b)

The expected Fisher information is given by

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

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

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

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

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:

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


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

```r
# 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

F.inv = NULL
eta = NULL

while (rel.err > tol & iter < max_iter){
  # Calculate eta.
  eta <- X %*% beta

  # Calculate score.
  D <- diag(as.vector(phi(eta)), n, n)
  Sigma <- diag(as.vector(Phi(eta)*(1 - Phi(eta))), n, n)
  mu.vec <- as.vector(Phi(eta))
  score = t(X) %*% D %*% solve(Sigma) %*% (y - mu.vec)

  # Calculate Fisher information and its inverse.
  W <- diag(as.vector(phi(eta)^2 / (Phi(eta)*(1-Phi(eta)))), n, n)
  F <- t(X) %*% W %*% X
  F.inv <- solve(F)

  # Update beta.
  beta.new <- beta + F.inv %*% score

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

# Calculating std.errors and deviance.
```

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

```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
##   -1.541896    3.303885
```

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

```
##
## Call:
## glm(formula = y ~ x, family = binomial(link = "probit"), data = df)
##
## Deviance Residuals:
##     Min      1Q   Median       3Q      Max
## -2.4875  -0.7474   0.3070   0.7177   2.2622
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.54190    0.09963  -15.47   <2e-16 ***
## x            3.30389    0.18440   17.92   <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: 1383.16  on 999  degrees of freedom
## Residual deviance:  973.18  on 998  degrees of freedom
## AIC: 977.18
##
## Number of Fisher Scoring iterations: 4
```

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

```
##               (Intercept)           x
## (Intercept)   0.009927101 -0.01629062
```

```
## x             -0.016290618  0.03400220
```
```r
# deviance
model$deviance
```
```
## [1] 973.1849
```
```r
### fit using myglm
mymodel <- myglm(y ~ x, data = df)
# beta
mymodel$coefficients
```
```
##                     y  std.Error
## (Intercept) -1.541899 0.09964074
## x            3.303892 0.18441026
```
```r
# vcov
mymodel$vcov
```
```
##               (Intercept)            x
## (Intercept)  0.009928277 -0.01629288
## x           -0.016292881  0.03400715
```
```r
# deviance
mymodel$deviance
```
```
## [1] 973.1849
```

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
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 $T_i \sim \mathcal{N}(\mu, \sigma^2)$, where T_i is the time until menarche occurs for the i'th individual. Furthermore, let

$$\pi_i := P(y_i = 1) = P(T_i \le t_i)$$
$$= P\left(\frac{T_i - \mu}{\sigma} \le \frac{t_i - \mu}{\sigma}\right) = \Phi\left(\frac{t_i - \mu}{\sigma}\right)$$

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 $\beta_1 = 1/\sigma$. Or equivalently:

$$\sigma = \frac{1}{\beta_1}, \quad \mu = -\frac{\beta_0}{\beta_1}.$$

Due to the functional invariance of the maximum likelihood estimator, we can write the MLEs of σ and μ as

$$\hat{\sigma}(\hat{\beta}_1) = \frac{1}{\hat{\beta}_1}, \quad \hat{\mu}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\hat{\beta}_0}{\hat{\beta}_1},$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ denote the MLEs of β_0 and β_1, respectively. Thus, the maximum likelihood estimates for this data set can be computed as

```
mod.probit <- glm(menarche ~ age, family = binomial(link = 'probit'), data = juul.girl)
b <- mod.probit$coefficients
mle.mu <- -b[1]/b[2]
mle.sigma <- 1/b[2]
mle.mu
```

```
## (Intercept)
##    13.18563
```

```
mle.sigma
```

```
##      age
## 1.159653
```

The standard errors of $\hat{\mu}$ and $\hat{\sigma}$ can then be estimated using the delta method. A first order Taylor expansion of $\hat{\mu}$ gives

$$\hat{\mu} \approx \hat{\mu}(\boldsymbol{b}) + \nabla\hat{\mu}(\boldsymbol{b})^T(\hat{\boldsymbol{\beta}} - \boldsymbol{b}),$$

where we have used the notation $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ and expanded around $\hat{\boldsymbol{\beta}} = \boldsymbol{b} = (b_0, b_1)^T$. Next, we take the variance of this linear approximation, such that

$$\text{Var}(\hat{\mu}) \approx \text{Var}\left(\nabla\hat{\mu}(\boldsymbol{b})^T\hat{\boldsymbol{\beta}}\right)$$
$$= \nabla\hat{\mu}(\boldsymbol{b})^T\text{Var}(\hat{\boldsymbol{\beta}})\nabla\hat{\mu}(\boldsymbol{b}).$$

Using $\nabla\hat{\mu} = (-1/\hat{\beta}_1, \hat{\beta}_0/\hat{\beta}_1^2)^T$, we can calculate the standard error as in the code below:

```r
mod.probit <- glm(menarche ~ age, family = binomial(link = 'probit'), data = juul.girl)
b <- mod.probit$coefficients
grad.mu <- c(-1/b[2], b[1]/b[2]^2)
se.mu <- t(grad.mu) %*% vcov(mod.probit) %*% grad.mu
se.mu
```

```
##             [,1]
## [1,] 0.01404909
```

That is, $\widehat{\text{SE}(\hat{\mu})} = 0.0140491$. We follow the same procedure to estimate the standard error of $\hat{\sigma}$. First, we approximate it by a first order Taylor series expansion:

$$\hat{\sigma} \approx \hat{\sigma}(b) + \nabla\hat{\sigma}(b)^T(\hat{\beta} - b),$$

which implies that the variance can be approximated as

$$\text{Var}(\hat{\sigma}) \approx \text{Var}\left(\nabla\hat{\sigma}(b)^T\hat{\beta}\right)$$
$$= \nabla\hat{\sigma}(b)^T\text{Var}(\hat{\beta})\nabla\hat{\sigma}(b).$$

Using that $\nabla\hat{\sigma} = (0, -1/\beta_1^2)^T$, we calculate the standard error as in the code below.

```r
grad.sigma <- c(0, -1/b[2]^2)
se.sigma <- t(grad.sigma) %*% vcov(mod.probit) %*% grad.sigma
se.sigma
```

```
##             [,1]
## [1,] 0.01188363
```

That is, $\widehat{\text{SE}(\hat{\sigma})} = 0.0118836$.

Q: are we estimating the standerd error (SE) or is the SE an estimate in itself? i.e. we dont need the hat?

c)

```r
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:

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

The pdf of T_i is then given as

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

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 estimates 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 $\mathrm{E}(T_i) = -\beta_0/\beta1 \approx 13.1901147$ and an estimate of the variance is given by $\mathrm{Var}(T_i) = s^2\pi^2/3 = \pi^2/(3\beta_1^2) \approx 1.4290323$.

d)

We now assume that the latent ages follow a log-normal distribution, i.e.

$$T_i \sim \mathrm{Lognormal}(\mu, \sigma^2).$$

This is equivalent to stating that $\ln T_i \sim \mathcal{N}(\mu, \sigma^2)$. Now we can follow the same approach as in 2b):

$$\pi_i := \Pr(y_i = 1) = \Pr(T_i \le t_i) = \Pr(\ln T_i \le \ln t_i)$$
$$= \Pr\left(\frac{\ln T_i - \mu}{\sigma} \le \frac{\ln t_i - \mu}{\sigma}\right) = \Phi\left(\frac{\ln t_i - \mu}{\sigma}\right)$$

This, in turn, gives

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

where $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$. Consequently, we fit GLM with a probit link-function on $\ln t_i$:

```r
mod.lognorm <- glm(menarche ~log(age), family = binomial(link = "probit"), data = juul.girl)
mu.hat <- -mod.lognorm$coefficients[1]/mod.lognorm$coefficients[2]
sigma.hat <- 1/mod.lognorm$coefficients[2]
summary(mod.lognorm)
```

```
##
## Call:
## glm(formula = menarche ~ log(age), family = binomial(link = "probit"),
##     data = juul.girl)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.28617  -0.11683   0.00315   0.10715   2.54738
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -29.060      2.704  -10.75   <2e-16 ***
## log(age)      11.287      1.049   10.76   <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: 719.39  on 518  degrees of freedom
## Residual deviance: 198.05  on 517  degrees of freedom
## AIC: 202.05
##
## Number of Fisher Scoring iterations: 8
```

Exactly as in 2b), due to the functional invariance of MLEs, we can estimate the mean of T_i as

$$\exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) = 13.1797035,$$

and we can estimate the standard deviation as

$$\sqrt{[\exp(\hat{\sigma}^2) - 1]\exp(2\hat{\mu} + \hat{\sigma}^2)} = 0.201604.$$

The formulas for mean and standard deviation of the log-normal distribution are gathered from Wikipedia.

e)

The cloglog link function is given by

$$g(\pi) = \text{cloglog}(\pi) = \ln(-\ln(1 - \pi)).$$

Since our model is assumed to have the form "menarch $\sim$ log(age)" with the cloglog link, η becomes

$$\eta = \beta_0 + \beta_1 \ln(t),$$

where t is age. Thus the probability that menarche has occurred is given as

$$\pi = g^{-1}(\eta) = 1 - e^{-e^{\eta}} = 1 - e^{-e^{\beta_0 + \beta_1 \ln t}} = 1 - e^{-(e^{\beta_0} t^{\beta_1})},$$

where it was used that $e^{\beta_1 \ln t} = t^{\beta_1}$. Using that π is the cumulative distribution function of T, i.e. $\Pr(T \leq t) = \pi$, it follows that the distribution of T is given by

$$T \sim \frac{\partial}{\partial t}\pi = \frac{\partial}{\partial t}\left(1 - e^{-(e^{\beta_0} t^{\beta_1})}\right) = \beta_1 e^{\beta_0} \ t^{\beta_1 - 1} \ e^{-(e^{\beta_0} t^{\beta_1})}.$$

Recall that the probability density function of a Weibull distribution is given by

$$\text{Weibull}(x; \lambda, k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$$

where k is the shape parameter and λ is the scale parameter. If $k = \beta_1$ and $\lambda = e^{-\frac{\beta_0}{\beta_1}}$, then

$$\text{Weibull}\left(t; e^{-\frac{\beta_0}{\beta_1}}, \beta_1\right) = \frac{\beta_1}{e^{-\frac{\beta_0}{\beta_1}}}\left(\frac{x}{e^{-\frac{\beta_0}{\beta_1}}}\right)^{\beta_1 - 1} e^{-\left(t \ e^{\frac{\beta_0}{\beta_1}}\right)^{\beta_1}}$$

$$= \beta_1 e^{\beta_0} \ t^{\beta_1 - 1} \ e^{-(e^{\beta_0} t^{\beta_1})},$$

which shows that $T \sim \text{Weibull}\left(t; e^{-\frac{\beta_0}{\beta_1}}, \beta_1\right)$. The shape and scale parameters are given as

$$k = \beta_1, \qquad \lambda = e^{-\beta_0/\beta_1}.$$

Due to the functional invariance of the maximum likelihood estimator, we can write the MLEs of k and λ as

$$\hat{k} = \hat{\beta}_1, \qquad \hat{\lambda} = e^{-\hat{\beta}_0/\hat{\beta}_1}.$$

Fitting the appropriate model in R and finding the estimate for the shape and scale parameter:

```
mod.cloglog <- glm(menarche ~ log(age), family = binomial(link = 'cloglog'), data = juul.girl)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
beta0 <- mod.cloglog$coefficients[1]
beta1 <- mod.cloglog$coefficients[2]

k <- beta1
lam <- exp(-beta0/beta1)

k
```

```
## log(age)
## 13.57193
```

```
lam
```

```
## (Intercept)
##    13.71689
```

The estimated values for the shape parameter is $k = 13.572$ and the scale parameter is $\lambda = 13.717$.

If $W \sim \text{Weibull}(x; \lambda, k)$, the mean and variance of W are given as

$$E[W] = \lambda \, \Gamma\left(1 + \frac{1}{k}\right), \quad \text{Var}[W] = \lambda^2 \left(\Gamma\left(1 + \frac{2}{k}\right) - \Gamma^2\left(1 + \frac{1}{k}\right)\right),$$

where Γ is the gamma function. In R, we get

```
mu <- lam * gamma(1+1/k)
SE <- lam * sqrt( gamma(1+2/k) - gamma(1+1/k)^2 )

mu
```

```
## (Intercept)
##    13.20256
```

```
SE
```

```
## (Intercept)
##     1.188294
```

The mean is 13.20 and the standard error is 1.19.

f)

Dette må vel også skrives en gang...