Assignment 1

Name: Kim S CHY

Student ID:998008

1. Fit a binomial regression model to the O-rings data from the Challenger disaster, using a probit link. You must use R (but without using the glm function); I want you to work from first principles.

(a) Compute MLEs (maximum likelihood estimates) of the parameters in the model.

```
#Loading Data
library(faraway)
data("orings")
str(orings)
```

```
## 'data.frame':    23 obs. of  2 variables:
##  $ temp  : num  53 57 58 63 66 67 67 67 68 69 ...
##  $ damage: num  5 1 1 1 0 0 0 0 0 0 ...
```

```
#Define MLE_Function
MLE_f <- function(beta, orings) {
  eta <- cbind(1, orings$temp) %*% beta
  p <- pnorm(eta)
  return(sum( orings$damage*log(p/(1-p)) +6*log(1-p)))
}
#Calculating Beta.hat
(beta_hat <- optim(c(10,-.1),MLE_f, orings=orings, control =
                    list(fnscale=-1))$par)
```

```
## [1]  5.5917242 -0.1058008
```

```
beta_hat
```

```
## [1]  5.5917242 -0.1058008
```

(b) Compute 95% CIs for the estimates of the parameters. You should show how you derived the Fisher information.

Let $p_i = F(\eta_i) = \int_{-\infty}^{\eta_i} \frac{e^{\frac{-t^2}{2}}}{\sqrt{2\pi}} dt$, $f(t) = \frac{e^{\frac{-t^2}{2}}}{\sqrt{2\pi}}$ and $\eta_i = \beta_0 + \beta_1 x_i$.

The log-likelihood function is given by:

$$\begin{aligned} l(\beta_0, \beta_1) &= c + \sum_{i=1} [y_i \log(p_i) + (m_i - y_i) \log(1 - p_i)] \\ &= c + \sum_{i=1} [y_i \log(F(\eta_i)) + (m_i - y_i) \log(1 - F(\eta_i))] \end{aligned}$$

$$\frac{dl}{d\beta_0} = \sum_{i=1} \left[y_i \frac{f(\eta_i)}{F(\eta_i)} + (m_i - y_i) \frac{f(\eta_i)}{F(\eta_i) - 1} \right]$$

Let $u = \eta_i^2 = (\beta_0 + \beta_1 x_i)^2 \implies \frac{df}{du} = \frac{-2e^{\frac{-u}{2}}}{\sqrt{2\pi}}$ and $\frac{du}{d\beta_0} = 2\eta_i$

From the above we get $\frac{df}{d\beta_0} = -\eta_i f(\eta_i)$; hence

$$\begin{aligned}
\frac{dl^2}{d\beta_0^2} &= \sum_{i=1}\left[y_i\left(\frac{-\eta_i f(\eta_i)}{F(\eta_i)} - \frac{(f(\eta_i))^2}{(F(\eta_i))^2}\right) + (m_i - y_i)\left(\frac{-\eta_i f(\eta_i)}{F(\eta_i)-1} - \frac{(f(\eta_i))^2}{(F(\eta_i)-1)^2}\right)\right] \\
&= \sum_{i=1}\left[y_i\left(\frac{-\eta_i f(\eta_i)}{F(\eta_i)} + \frac{\eta_i f(\eta_i)}{F(\eta_i)-1} - \frac{(f(\eta_i))^2}{(F(\eta_i))^2} + \frac{(f(\eta_i)^2)}{(F(\eta_i)-1)^2}\right) - m_i\left(\frac{\eta_i f(\eta_i)}{F(\eta_i)-1} + \frac{(f(\eta_i))^2}{(F(\eta_i)-1)^2}\right)\right] \\
&= \sum_{i}\left[y_i f(\eta_i)\left(\frac{f(\eta_i)(2F(\eta_i)-1) + \eta_i F(\eta_i)(F(\eta_i)-1)}{(F(\eta_i))^2(F(\eta_i)-1)^2}\right) - m_i\left(\frac{\eta_i f(\eta_i)}{F(\eta_i)-1} + \frac{(f(\eta_i))^2}{(F(\eta_i)-1)^2}\right)\right]
\end{aligned}$$

Since $Y_i \sim \text{bin}(m_i, p_i) \implies E(Y_i) = p_i m_i$; hence:

$$\begin{aligned}
\mathbb{E}\left(-\frac{d^2 l}{d\beta_0^2}\right) &= \sum_{i} -\left[m_i f(\eta_i)\left(\frac{f(\eta_i)(2F(\eta_i)-1) + \eta_i F(\eta_i)(F(\eta_i)-1)}{(F(\eta_i))(F(\eta_i)-1)^2}\right) - m_i\left(\frac{\eta_i f(\eta_i)}{F(\eta_i)-1} + \frac{(f(\eta_i))^2}{(F(\eta_i)-1)^2}\right)\right] \\
&= \sum_{i=1}\left[\frac{m_i * f(\eta_i)}{F(\eta_i)(F(\eta)-1)}\right] \\
&= \sum_{i=1}\left[2m_i(f(\eta_i))^2\left(\frac{1}{p_i - p_i^2}\right)\right]
\end{aligned}$$

$$\frac{dl^2}{d\beta_0 d\beta_1} = \sum_{i=1}\left[x_i y_i f(\eta_i)\left(\frac{f(\eta_i)(2F(\eta_i)-1) + \eta_i F(\eta_i)(F(\eta_i)-1)}{(F(\eta_i))^2(F(\eta_i)-1)^2}\right) - m_i\left(\frac{\eta_i f(\eta_i)}{F(\eta_i)-1} + \frac{(f(\eta_i))^2}{(F(\eta_i)-1)^2}\right)\right]$$

$$\begin{aligned}
\mathbb{E}\left(-\frac{d^2 l}{d\beta_0 d\beta_1}\right) &= \sum_{i=1} -\left[x_i m_i (f(\eta_i))^2\left(\frac{1 - F(\eta_i)}{F(\eta_i)(F(\eta_i)-1)^2}\right)\right] \\
&= \sum_{i=1}\left[x_i m_i (f(\eta_i))^2\left(\frac{1}{p_i - p_i^2}\right)\right]
\end{aligned}$$

$$\frac{dl^2}{d\beta_1^2} = \sum_{i}\left[(x_i)^2 y_i f(\eta_i)\left(\frac{f(\eta_i)(2F(\eta_i)-1) + \eta_i F(\eta_i)(F(\eta_i)-1)}{(F(\eta_i))^2(F(\eta_i)-1)^2}\right) - m_i\left(\frac{\eta_i f(\eta_i)}{F(\eta_i)-1} + \frac{(f(\eta_i))^2}{(F(\eta_i)-1)^2}\right)\right]$$

$$\begin{aligned}
\mathbb{E}\left(-\frac{d^2 l}{d\beta_0 d\beta_1}\right) &= \sum_{i=1} -\left[(x_i)^2 m_i (f(\eta_i))^2\left(\frac{1 - F(\eta_i)}{F(\eta_i)(F(\eta_i)-1)^2}\right)\right] \\
&= \sum_{i=1}\left[(x_i)^2 m_i (f(\eta_i))^2\left(\frac{1}{p_i - p_i^2}\right)\right]
\end{aligned}$$

The fisher information is given by:

$$\mathcal{I}(\beta) = \begin{bmatrix} \sum_{i=1}\left[m_i(f(\eta_i))^2\left(\frac{1}{p_i - p_i^2}\right)\right] & \sum_{i=1}\left[x_i m_i(f(\eta_i))^2\left(\frac{1}{p_i - p_i^2}\right)\right] \\ \sum_{i=1}\left[x_i m_i(f(\eta_i))^2\left(\frac{1}{p_i - p_i^2}\right)\right] & \sum_{i=1}\left[(x_i)^2 m_i(f(\eta_i))^2\left(\frac{1}{p_i - p_i^2}\right)\right] \end{bmatrix}$$

Finding standard error for parameter

```r
f <- function(t){
  exp(-(t^2)/2)/(sqrt(2*pi))
}

phat <- pnorm(beta_hat[1]+orings$temp*beta_hat[2])
h <- function(t){
  1/(t-t^2)
}
I11 <- sum(6*(f(phat)^2)*h(phat))
I12 <- sum(6*orings$temp*(f(phat))^2*h(phat))
I22 <- sum(6*(orings$temp^2)*(f(phat))^2*h(phat))
Iinv <- solve(matrix(c(I11, I12, I12, I22), 2,2))
sqrt(Iinv[1,1])
```

```
## [1] 0.4065001
```

```r
sqrt(Iinv[2,2])
```

```
## [1] 0.005259742
```

The 95% Confidence Interval for β_0 and β_1 is:

```r
#Beta 0
c(beta_hat[1]-qnorm(0.025)*sqrt(Iinv[1,1]),beta_hat[1]+qnorm(0.025)*sqrt(Iinv[1,1]))
```

```
## [1] 6.388450 4.794999
```

```r
#Beta 1

c(beta_hat[2]-qnorm(0.025)*sqrt(Iinv[2,2]),beta_hat[2]+qnorm(0.025)*sqrt(Iinv[2,2]))
```

```
## [1] -0.09549185 -0.11610966
```

c. Perform a likelihood ratio test for the significance of the temperature coefficient.

```r
MaxlogL.F <- MLE_f(beta_hat,orings)
MaxlogL.F
```

```
## [1] -27.98886
```

```r
y <- orings$damage
n <- rep(6, length(y))
phatN <- sum(y)/sum(n)
MaxlogL.R = sum(orings$damage)*log(phatN) + sum(6-orings$damage)*log(1-phatN)
MaxlogL.R
```

```
## [1] -38.3724
```

LR-Test statistic and p-value for LR-test

```r
LR = -2*(MaxlogL.R - MaxlogL.F)
LR
```

```
## [1] 20.76708
```

```r
pchisq(LR, df=1,lower=FALSE)
```

```
## [1] 5.186684e-06
```

d. Compute an estimate of the probability of damage when the temperature equals 31 Fahrenheit (your estimate should come with a 95% CI, as all good estimates do).

```r
#Computing Prediction for 31 degree
phat <- pnorm(beta_hat[1]+beta_hat[2]*31)
phat
```

```
## [1] 0.9896084
```

```r
#Computing 95% CI 31 degree
si2 <- matrix(c(1, 31), 1, 2) %*% Iinv %*% matrix(c(1, 31), 2, 1)
etahat = beta_hat[1] + beta_hat[2]*31
eta_l = etahat - 2*sqrt(si2)
eta_r = etahat + 2*sqrt(si2)

c(pnorm(eta_l),pnorm(eta_r))
```

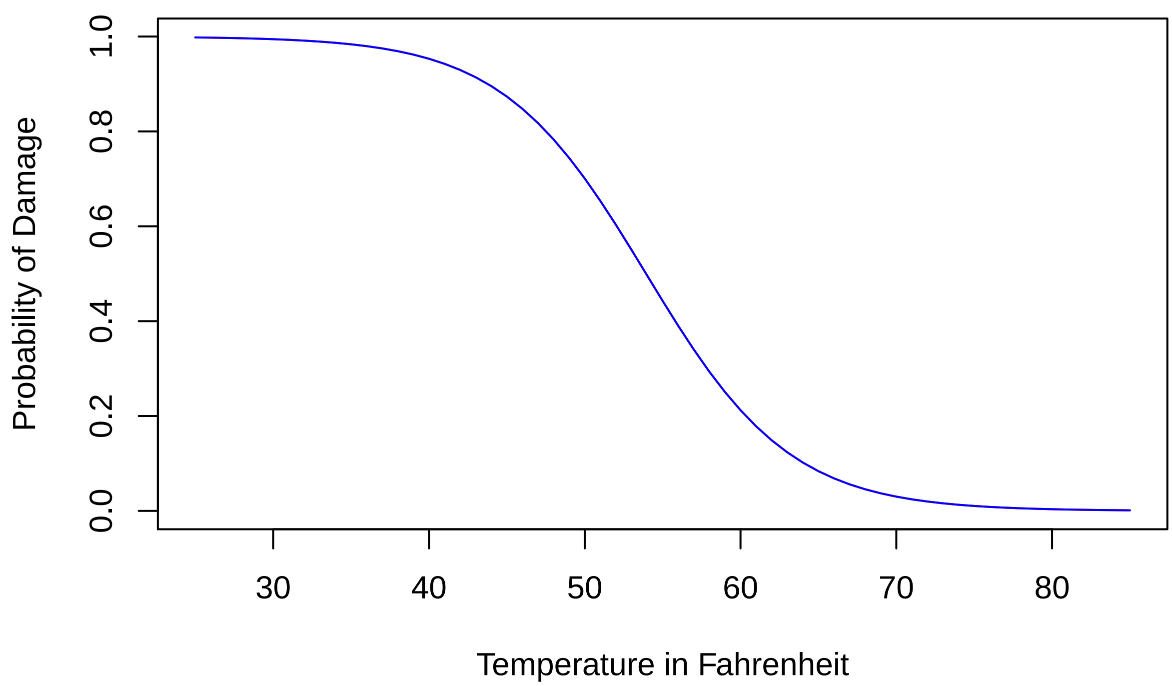
```
## [1] 0.9659392 0.9974417
```

e. Make a plot comparing the fitted probit model to the fitted logit model. To obtain the fitted logit model, you are allowed to use the glm function.

```r
#Building Model using GLM
probitmodel <-  glm(cbind(damage,6-damage) ~ temp, family=binomial(link="probit"), orings)
logitmodel <- probitmodel <-  glm(cbind(damage,6-damage) ~ temp, family=binomial, orings)
temp_rise <- seq(25, 85, 1)

#Getting sequence of response
logitresp <- predict(logitmodel,list(temp=temp_rise),type="response")
probitresp <- predict(probitmodel,list(temp=temp_rise), type = "response")

#Plotting
plot(temp_rise,logitresp, type = "l", col="red",xlab = "Temperature in Fahrenheit",
     ylab="Probability of Damage")
lines(temp_rise, probitresp, col="Blue")
```



We can see that the line perfectly overlap one another making it indifference between using either the probit or logit model.

Q2.

a.Please estimate the amount of increase in the log(odds) when the bmi increases by 7.

Loading Data and Building Model

```
#Loading Data
data()
missing <- with(pima, missing <- glucose==0 | diastolic==0 | triceps==0 | bmi == 0)
pima_subset = pima[!missing, c(6,9)]
str(pima_subset)
```

```
## 'data.frame':	532 obs. of  2 variables:
##  $ bmi : num  33.6 26.6 28.1 43.1 31 30.5 30.1 25.8 45.8 43.3 ...
##  $ test: int  1 0 0 1 1 1 1 1 1 0 ...
```

```
#Building Model using GLM
```

```
modelQ2 <- glm(cbind(test, 1-test)~bmi, family=binomial, data=pima_subset)
summary(modelQ2)
```

```
##
## Call:
## glm(formula = cbind(test, 1 - test) ~ bmi, family = binomial,
##     data = pima_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9227  -0.8920  -0.6568   1.2559   1.9560
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.03681    0.52783  -7.648 2.04e-14 ***
## bmi          0.09972    0.01528   6.524 6.84e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 676.79  on 531  degrees of freedom
## Residual deviance: 627.46  on 530  degrees of freedom
## AIC: 631.46
##
## Number of Fisher Scoring iterations: 4
```

Please Estimate the amount increase

```
x <- predict(modelQ2, newdata = list(bmi=7),type="link", se.fit=TRUE)

log_odd <- function(phat){
  exp(phat/(1-phat))
}

log_odd(x$fit)
```

```
##         1
## 0.4632348
```

b. Compute a 95% CI for the estimate.

```
log_odd(c(x$fit-2*x$se.fit, x$fit+2*x$se.fit))
```

```
##         1         1
## 0.4461340 0.4898387
```

3.The gamma distribution with shape $\nu > 0$ and rate $\lambda > 0$ has pdf:

$$f(x;\nu,\lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}$$

for $x > 0$.

a.Show that the gamma distribution is an exponential family.
Want to show that $f(x;\nu,\lambda) = \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x}$ can be written as $f(x;\theta,\phi) = \exp\left[\frac{x\theta - b(\theta)}{a(\phi)} - c(x,\phi)\right]$.

$$\begin{aligned} f(x;\nu,\lambda) &= \frac{\lambda^\nu}{\Gamma(\nu)} x^{\nu-1} e^{-\lambda x} \\ &= \exp\left[\nu \log(\lambda) - \log(\Gamma(\nu)) + (\nu-1)\log(x) - \lambda x\right] \\ &= \exp\left[-(\lambda x - \nu \log(\lambda)) + (\nu-1)\log(x) - log(\Gamma(\nu))\right] \\ &= \exp\left[\frac{\frac{\lambda x}{\nu} - \log(\lambda)}{\frac{-1}{\nu}} + (\nu-1)\log(x) - \log(\Gamma(\nu))\right] \end{aligned}$$

Let $\theta = \frac{\lambda}{\nu}$ and $\phi = \frac{1}{\nu}$.
Hence $\lambda = \theta\nu = \frac{\theta}{\phi} \implies \log(\lambda) = \log(\theta) - \log(\phi)$.

The above become:

$$f(x;\theta,\phi) = \exp\left[\frac{\theta x - \log(\theta)}{-\phi} + \frac{\log(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\log(x) - \log(\Gamma(\frac{1}{\phi}))\right)\right]$$

Now Let $b(\theta) = \log(\theta)$; $a(\phi) = -\phi$ and $c(x,\phi) = \left(\frac{1}{\phi} - 1\log(x) - \log(\Gamma(\frac{1}{\phi}))\right)$. Thus:

$$f(x;\theta,\phi) = \exp\left[\frac{x\theta - b(\theta)}{a(\phi)} - c(x,\phi)\right]$$

Hence it is part of the exponential family.

b. Obtain the canonical link and the variance function.

Since, $\mathbb{E}(X) = b'(\theta) = \frac{db}{d\theta} = \theta^{-1} = \frac{\nu}{\lambda}$ and the canonical link function is given by $[b'(\mu)]^{-1}$.
This implied the canical link function is $g(\mu) = \theta = [b'(\mu)]^{-1} = \mu^{-1}$.

Since $b''(\theta) = \frac{d^2b}{d\theta^2} = -\theta^{-2}$ and the variance function is given by: $v(\mu) = -\theta^{-2} \cdot \mu^{-1} = -\mu$.Thus the variance function is given by:

$$\text{Var}X = v(\mu)a(\phi) = \mu\phi = \frac{\mu}{\nu}$$