MAST30027_Assignment1

Zi Ng (1085130)

Thursday 4.55 - 5.15, Anubhav Kaphle

Question 1

Fit a binomial regression model to the O-rings data from the Challenger disaster, using a complementary log-log link, from first principles

1a Compute the MLEs of model parameters.

First, we will load and inspect the data.

```
# load 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 ...
```

Next, derive the log-likelihood of binomial regression with complementary log-log link function, given by $\eta_i = \log(-\log(1-p))$ and $p_i = 1 - \exp(-e^{\eta_i})$.

$$
\begin{aligned}
l(\beta_0, \beta_1) &= \log L(\beta_0, \beta_1) && (1)\\
&= c + \sum_i (y_i \log p_i + (m_i - y_i) log(1 - p_i)) && (2)\\
&= c + \sum_i [y_i \log \frac{p_i}{1-p_i} + m_i \log(1 - p_i)] && (3)\\
&= c + \sum_i [y_i \log \frac{1 - \exp(-e^{\eta_i})}{1 - (1 - \exp(-e^{\eta_i}))} + m_i \log(1 - (1 - \exp(-e^{\eta_i})))] && (4)\\
&= c + \sum_i [y_i \log(\exp(e^{\eta_i}) - 1) - m_i e^{\eta_i}] && (5)
\end{aligned}
$$

We use the optim function to compute values for beta that maximize the log-likelihood.

```
# log likelihood function
logL <- function(beta, orings) {
  eta <- cbind(1, orings$temp) %*% beta
  return (sum(orings$damage * log(exp(exp(eta)) - 1) - 6*exp(eta)))
}
# determine parameter estimates
(betahat <- optim(c(10, -0.1), logL, orings=orings, control=list(fnscale=-1))$par)
```

```
## [1] 10.8585961 -0.2054664
```

Hence, the parameter estimates are $\hat{\beta}_0 = 10.8586$ and $\hat{\beta}_1 = -0.2055$.

1b 95% CIs for the parameter estimates.

From asymptotic normality of MLE, we know

$$\hat{\theta} \approx^d N(\theta^*, I(\theta^*)^{-1})$$

Firstly, we will derive the observed information, $J(\theta; Y)$. Keeping in mind that $\eta_i = \beta_0 + \beta_1 x_i$, we calculate the double derivatives.

$$\frac{\partial l(\beta_0, \beta_1)}{\partial \beta_0} = \sum_{i=1}^{n} [\frac{y_i}{\exp(e^{\eta_i}) - 1} \cdot \exp(e^{\eta_i}) \cdot e^{\eta_i} \cdot 1 - m_i e^{\eta_i} \cdot 1] \tag{6}$$

$$= \sum_{i=1}^{n} [e^{\eta_i} \cdot (\frac{y_i \exp(e^{\eta_i})}{\exp(e^{\eta_i}) - 1} - m_i)] \tag{7}$$

$$= \sum_{i=1}^{n} [e^{\eta_i} \cdot (\frac{y_i}{1 - \exp(-e^{\eta_i})} - m_i)] \tag{8}$$

$$\frac{\partial^2 l(\beta_0, \beta_1)}{\partial \beta_0^2} = \sum_{i=1}^{n} [e^{\eta_i} \cdot (\frac{y_i}{1 - \exp(-e^{\eta_i})} - m_i) + e^{\eta_i} \cdot (\frac{y_i}{(1 - \exp(-e^{\eta_i}))^2}) \cdot \exp(-e^{\eta_i}) \cdot e^{\eta_i}] \tag{9}$$

$$= \sum_{i=1}^{n} [e^{\eta_i} (\frac{y_i}{p} - m_i + \frac{y_i}{p^2}(1-p)(-\log(1-p)))] \tag{10}$$

$$= \sum_{i=1}^{n} [\log(1-p) \cdot (\frac{y_i(1-p)}{p^2} \log(1-p) + m_i - \frac{y_i}{p})] \tag{11}$$

$$\frac{\partial l(\beta_0, \beta_1)}{\partial \beta_1} = \sum_{i=1}^{n} [\frac{y_i}{\exp(e^{\eta_i}) - 1} \exp(e^{\eta_i}) \cdot x_i - m_i \cdot e^{\eta_i} \cdot x_i] \tag{12}$$

$$= \sum_{i=1}^{n} [e^{\eta_i} x_i \cdot (\frac{y_i \exp(e^{\eta_i})}{\exp(e^{\eta_i}) - 1} - m_i)] \tag{13}$$

$$= \sum_{i=1}^{n} [e^{\eta_i} x_i \cdot (\frac{y_i}{1 - \exp(-e^{\eta_i})} - m_i)] \tag{14}$$

$$\frac{\partial^2 l(\beta_0, \beta_1)}{\partial \beta_1^2} = \sum_{i=1}^{n} [x_i^2 e^{\eta_i} \cdot (\frac{y_i}{1 - \exp(-e^{\eta_i})} - m_i) + e^{\eta_i} x_i \cdot (\frac{y_i}{(1 - \exp(-e^{\eta_i}))^2} \cdot \exp(-e^{\eta_i}) \cdot e^{\eta_i} \cdot x_i)] \tag{15}$$

$$= \sum_{i=1}^{n} [x_i^2 e^{\eta_i} \cdot (\frac{y_i}{1 - \exp(-e^{\eta_i})} - m_i + \frac{y_i}{(1 - \exp(-e^{\eta_i}))^2} \cdot \exp(-e^{\eta_i}) \cdot e^{\eta_i} \cdot x_i)] \tag{16}$$

$$= \sum_{i=1}^{n} [x_i^2 \log(1-p) \cdot (\frac{y_i(1-p)}{p^2} \log(1-p) + m_i - \frac{y_i}{p})] \tag{17}$$

$$\frac{\partial^2 l(\beta_0, \beta_1)}{\partial \beta_1 \beta_0} = \sum_{i=1}^{n} [x_i e^{\eta_i} \cdot (\frac{y_i}{1 - \exp(-e^{\eta_i})} - m_i) + e^{\eta_i} x_i \cdot (\frac{y_i}{(1 - \exp(-e^{\eta_i}))^2} \cdot \exp(-e^{\eta_i}) \cdot e^{\eta_i})] \tag{18}$$

$$= \sum_{i=1}^{n} [x_i e^{\eta_i} \cdot (\frac{y_i}{1 - \exp(-e^{\eta_i})} - m_i + \frac{y_i}{(1 - \exp(-e^{\eta_i}))^2} \cdot \exp(-e^{\eta_i}) \cdot e^{\eta_i} \cdot x_i)] \tag{19}$$

$$= \sum_{i=1}^{n} [x_i \log(1-p) \cdot (\frac{y_i(1-p)}{p^2} \log(1-p) + m_i - \frac{y_i}{p})] \tag{20}$$

$$\tag{21}$$

Taking the expectations of the observed information gives us the fisher information, $I(\theta) = E[J(\theta; Y)]$. The aim is to eliminate y_i from the expression, so we note that $E(Y) = m\hat{p}$ and take the expectation of the

observed information.

$$I_{1,1} = \sum_{i=1}^{n} [\log(1-p)(\frac{m_i \cdot (1-p) \cdot \log(1-p)}{p})] \tag{22}$$

$$I_{1,2} = \sum_{i=1}^{n} [x_i \log(1-p)(\frac{m_i \cdot (1-p) \cdot \log(1-p)}{p})] \tag{23}$$

$$I_{2,1} = \sum_{i=1}^{n} [x_i \log(1-p)(\frac{m_i \cdot (1-p) \cdot \log(1-p)}{p})] \tag{24}$$

$$I_{2,2} = \sum_{i=1}^{n} [x_i^2 \log(1-p)(\frac{m_i \cdot (1-p) \cdot \log(1-p)}{p})] \tag{25}$$

$$\tag{26}$$

We now have the equations to compute the standard error of $\hat{\beta}$, which we can do in R.

```
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
##
## Attaching package: 'VGAM'
```

```
## The following objects are masked from 'package:faraway':
##
##     hormone, logit, pneumo, prplot
```

```
phat <- clogloglink(betahat[1] + orings$temp * betahat[2], inverse=TRUE)
mult <- (6 * (1-phat) * log(1-phat) / phat) * log(1-phat)
I11 <- sum(mult)
I12 <- sum(orings$temp * mult)
I22 <- sum(orings$temp^2 * mult)
Iinv <- solve(matrix(c(I11, I12, I12, I22), 2, 2))
(se.betahat1 <- sqrt(Iinv[1,1]))
```

```
## [1] 2.738166
```

```
(se.betahat2 <- sqrt(Iinv[2,2]))
```

```
## [1] 0.04563258
```

Finally, we can compute the 95% confidence intervals for $\hat{\beta}$.

```
# compute CI for betahat
betahat[1] + c(-1,1) * qnorm(0.975) * se.betahat1
```

```
## [1]  5.491889 16.225304
```

```
betahat[2] + c(-1,1) * qnorm(0.975) * se.betahat2
```

```
## [1] -0.2949046 -0.1160282
```

Hence, the 95% confidence interval for $\hat{\beta}_0$ is $[5.4919, 16.2253]$ and the confidence interval for $\hat{\beta}_1$ is $[-0.2949, -0.1160]$.

1c Perform a likelihood ratio test for the significance of the temperature coefficient.

We are testing if $\hat{\beta}_1$ is significant in the model.

$$H_0 : \beta_1 = 0 H_1 : \beta_1 \neq 0$$

First, we will compute the maximum log-likelihood for the full model, and the reduced model which does not include the temperature coefficient. We note that in the reduced model, $\hat{p}$ is the proportion of orings that were damaged in all the trials, given by $\frac{\sum y_i}{\sum m_i}$.

```
# compute the maximum log-likelihood for the full model
(mll.full <- logL(betahat, orings))
```

```
## [1] -26.93787
```

```
# compute the maximum log-likelihood for the reduced model
y <- orings$damage
n <- rep(6, length(y))
phatN <- sum(y) / sum(n)
(mll.red <- sum(y) * log(phatN) + sum(6-y) * log(1-phatN))
```

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

We then compute the Likelihood Ratio test statistic. Under the null hypothesis, we expect this test statistic to be chi-squared distributed with degrees of freedom 1 since the full model has 2 parameters, while the reduced model has 1.

```
(LR <- -2 * (mll.red - mll.full))
```

```
## [1] 22.86905
```

Hence, we can compute the p-value of the LR test statistic.

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

```
## [1] 1.734217e-06
```

Given a p-value of < 0.05, we can thus conclude that the temperature coefficient is significant at a 95% confidence level.

1d Compute an estimate of the probability of damage when the temperature equals 31 Fahrenheit, as well as the 95% confidence interval for it.

We can estimate $\hat{\eta}$ from $\hat{\beta}$ since $\eta = \beta_0 + \beta_1 t$. We use that to obtain an estimate of $\hat{p}$ since we know $p = g^{-1}(\eta)$ where g is the complementary log-log link function.

```
# estimate of probability
etahat <- betahat[1] + betahat[2] * 31
(p.31f <- clogloglink(etahat, inverse=TRUE))
```

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

From asymptotic normality of MLE, we know that $t^T\hat{\theta} \approx N(t^T\theta^*, t^T I(\hat{\theta})^{-1}t)$.To compute the confidence interval, we first need the standard error of $\hat{\eta}$.

$$se(\hat{p}) = \sqrt{\begin{bmatrix} 1 & 31 \end{bmatrix} I(\hat{\beta})^{-1} \begin{bmatrix} 1 \\ 31 \end{bmatrix}}$$

```
# se of p hat
(si2 <- matrix(c(1, 31), 1, 2) %*% Iinv %*% matrix(c(1, 31), 2, 1))
```

```
##          [,1]
## [1,] 1.799022
```

Next, we compute the CI for $\hat{eta}$, which we can plug into the inverse link function to obtain a CI for $\hat{p}$.

```
# CI for etahat
eta_l = etahat - qnorm(0.975) * sqrt(si2)
eta_r = etahat + qnorm(0.975) * sqrt(si2)
# 95% CI for p
c(clogloglink(eta_l, inverse=TRUE), clogloglink(eta_r, inverse=TRUE))
```

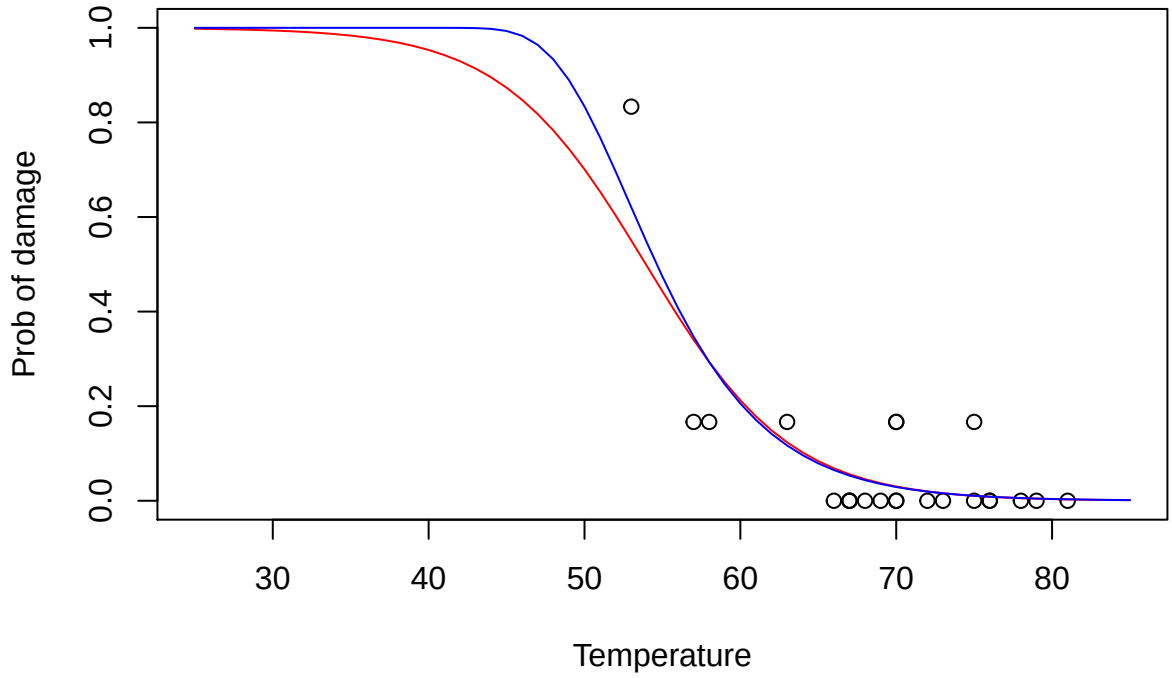
```
## [1] 0.9983804 1.0000000
```

1e Make a plot comparing the fitted complementary model to the fitted logit model.

```
# logit model
logitmodel <- glm(cbind(damage, 6-damage) ~ temp, family=stats::binomial, orings)

# plot fitted logit model
plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim=c(0,1),
     xlab="Temperature", ylab="Prob of damage")
x <- seq(25,85,1)
ilogit <- function(x) exp(x) / (1+exp(x))
lines(x, ilogit(logitmodel$coefficients[1] + logitmodel$coefficients[2]*x),
      col="red")

# plot fitted cloglog model
icloglog <- function(x) 1 - exp(-exp(x))
lines(x, icloglog(betahat[1] + betahat[2] * x), col="blue")
```



Question 2

Fit a binomial regression model with a logit link with test as a response and bmi as a predictor to the pima data set.

```
# load the data
library(faraway)
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 ...
```

```
# fit a binomial regression model with logit link
pima_model <- glm(cbind(test, 1-test) ~ bmi, family=stats::binomial, pima_subset)
summary(pima_model)
```

```
##
## Call:
## glm(formula = cbind(test, 1 - test) ~ bmi, family = stats::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
```

2a Estimate in the amount of increase in the log(odds) when the bmi increases by 5.

$$\begin{aligned}
&\Delta log(o) && (27)\\
=&log(o(t+5)) - log(o(t)) && (28)\\
=&log(\frac{p(t+5)}{1-o(t+5)}) - log(\frac{p(t)}{1-p(t)}) && (29)\\
=&\beta_0 + \beta_1(t+5) - \beta_0 - \beta_1 t && (30)\\
=&5\beta_1 && (31)
\end{aligned}$$

From the fitted model, $\hat{\beta}_1 = 0.09972$, which means $5\hat{\beta}_1 = 0.4985842 \approx 0.4986$.

2b Compute a 95% CI for the estimate.

We know that $\hat{\beta}$ is normally-distributed from the asymptotic normality. Hence, we can derive the distribution of $5\hat{\beta}_1$.

$$5\hat{\beta}_1 \approx N(5 \cdot \beta_1, 25 \cdot se(\hat{\beta}_1)^2)$$

```
betahat.1 <- pima_model$coefficients[1]
se.betahat.1 <- 0.01528 # read from summary output
5 * betahat.1 + c(-1,1) * qnorm(0.975) * se.betahat.1 * 5
```

```
## [1] -20.33379 -20.03431
```

Therefore the 95% confidence interval for the estimate of the increase in the log(odds) when bmi increases by 5 is $(-20.3338, -20.0343)$.

Question 3

The inverse Gaussian distribution.

3a Show that the inverse Gaussian distribution is an exponential family.

$$f(x;\mu,\lambda) = (\frac{\lambda}{2\pi x^3})^{1/2} \exp \frac{-\lambda(x-\mu)^2}{2\mu^2 x} \tag{32}$$

$$= \exp \frac{1}{2} \log \frac{\lambda}{2\pi x^3} - \frac{\lambda(x-\mu)^2}{2\mu^2 x} \tag{33}$$

$$= \exp \frac{-\lambda x}{2\mu^2} + \frac{\lambda}{\mu} + \frac{1}{2} \log \frac{\lambda}{2\pi x^3} - \frac{\lambda}{2x} \tag{34}$$

$$= \exp \frac{-\frac{1}{\mu^2} x - \frac{2}{\mu}}{\frac{2}{\lambda}} + \frac{1}{2} \log \frac{\lambda}{2\pi x^3} - \frac{\lambda}{2x} \tag{35}$$

The inverse Gaussian distribution is of the form $f(y;\theta,\phi) = \exp[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)]$, where

$$\theta = -\frac{1}{\mu^2} \tag{36}$$

$$b(\theta) = \frac{2}{\mu} = 2\sqrt{-\theta} \tag{37}$$

$$\phi = \lambda \tag{38}$$

$$a(\phi) = \frac{2}{\lambda} = \frac{2}{\phi} \tag{39}$$

We have shown that the inverse Gaussian distribution is an exponential family.

3b Obtain the canonical link and the variance function.

$$b'(\theta) = -\frac{1}{\sqrt{-\theta}} = -(-\theta)^{-\frac{1}{2}} \tag{40}$$

$$\therefore (b')^{-1}(\mu) = -\frac{1}{\mu^2} = \theta \tag{41}$$

Since the canonical link is given by $(b')^{-1}$, the canonical link is $g(\mu) = -\frac{1}{\mu^2}$.

$$v(x) = b''((b')^{-1}(\mu)) \tag{42}$$

$$= b''(\theta) \tag{43}$$

$$= -(-\frac{1}{2})(-\theta)^{-\frac{3}{2}} \tag{44}$$

Hence, the variance function is given by $v(\mu) = \frac{1}{2}(-\theta)^{-\frac{3}{2}}$.