MAST30027: Modern Applied Statistics

Assignment 1 Solution 2020

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

   (a) (3 marks) Compute MLEs (maximum likelihood estimates) of the parameters in the model.

   (b) (7 marks) Compute 95% CIs for the estimates of the parameters. You should show how you derived the fisher information.

   (c) (3 marks) Perform a likelihood ratio test for the significance of the temperature coefficient.

   (d) (3 marks) 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).

   (e) (2 marks) Make a plot comparing the fitted complementary log-log model to the fitted logit model. To obtain the fitted logit model, you are allowed to use the `glm` function.

   Show your working including the R code you use and detailed derivation.

Solution

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

For a binomial regression with a c-log-log link we have $y_i \sim \text{bin}(m_i, p_i)$, where $p_i = 1 - \exp(-e^{\eta_i})$ and $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, so

$$\begin{aligned} l(\boldsymbol{\beta}) &= c + \sum_i [y_i \log p_i - (m_i - y_i)\log(1 - p_i)] \\ &= c + \sum_i [y_i \log(1 - \exp(-e^{\eta_i})) - (m_i - y_i)e^{\eta_i}] \end{aligned}$$

```
> library(faraway)
> data(orings)
> logL <- function(beta, orings) {
+   y <- orings$damage
+   X <- cbind(1, orings$temp)
+   zeta <- X %*% beta
+   p <- 1 - exp(-exp(zeta))
+   return(sum(y*log(p) + (6 - y)*log(1 - p)))
+ }
> (betahat <- optim(c(10, -.2), logL, orings=orings, control=list(fnscale=-1))$par)
[1] 10.8622281 -0.2054973
```

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

$$-\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} = -\sum_{i=1}^{n}\left[-(m_i-y_i)\frac{e^{\beta_0+\beta_1 t_i}e^{-e^{\beta_0+\beta_1 t_i}}}{e^{-e^{\beta_0+\beta_1 t_i}}}+y_i\frac{e^{\beta_0+\beta_1 t_i}e^{-e^{\beta_0+\beta_1 t_i}}}{1-e^{-e^{\beta_0+\beta_1 t_i}}}\right]$$

$$= -\sum_{i=1}^{n}\left[-(m_i-y_i)\,e^{\beta_0+\beta_1 t_i}+y_i\frac{e^{\beta_0+\beta_1 t_i}}{e^{e^{\beta_0+\beta_1 t_i}-1}}\right]$$

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0^2} = -\sum_{i=1}^{n}\left[-(m_i-y_i)\,e^{\beta_0+\beta_1 t_i}+y_i\frac{e^{\beta_0+\beta_1 t_i}\left(e^{e^{\beta_0+\beta_1 t_i}}-1\right)-e^{2(\beta_0+\beta_1 t_i)}e^{e^{\beta_0+\beta_1 t_i}}}{\left(e^{e^{\beta_0+\beta_1 t_i}-1}\right)^2}\right]$$

$$E\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0^2}\right) = -\sum_{i=1}^{n}\left[-(m_i-m_ip_i)\log\left(\frac{1}{1-p_i}\right)+m_ip_i\frac{\log\left(\frac{1}{1-p_i}\right)\left(\frac{p_i}{1-p_i}\right)-\log\left(\frac{1}{1-p_i}\right)^2\left(\frac{1}{1-p_i}\right)}{\left(\frac{p_i}{1-p_i}\right)^2}\right]$$

$$= -\sum_{i=1}^{n} m_i\left[\frac{-(1-p_i)\log\left(\frac{1}{1-p_i}\right)p_i}{p_i}+\frac{\log\left(\frac{1}{1-p_i}\right)(1-p_i)\,p_i-\log\left(\frac{1}{1-p_i}\right)^2(1-p_i)}{p_i}\right]$$

$$= \sum_{i=1}^{n}\frac{m_i\,(1-p_i)\,(\log(1-p_i))^2}{p_i}$$

$$-\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = -\sum_{i=1}^{n} t_i\left[-(m_i-y_i)\frac{e^{\beta_0+\beta_1 t_i}e^{-e^{\beta_0+\beta_1 t_i}}}{e^{-e^{\beta_0+\beta_1 t_i}}}+y_i\frac{e^{\beta_0+\beta_1 t_i}e^{-e^{\beta_0+\beta_1 t_i}}}{1-e^{-e^{\beta_0+\beta_1 t_i}}}\right]$$

$$= -\sum_{i=1}^{n} t_i\left[-(m_i-y_i)\,e^{\beta_0+\beta_1 t_i}+y_i\frac{e^{\beta_0+\beta_1 t_i}}{e^{\beta_0+\beta_1 t_i}-1}\right]$$

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1^2} = -\sum_{i=1}^{n} t_i^2\left[-(m_i-y_i)\,e^{\beta_0+\beta_1 t_i}+y_i\frac{e^{\beta_0+\beta_1 t_i}\left(e^{e^{\beta_0+\beta_1 t_i}}-1\right)-e^{2(\beta_0+\beta_1 t_i)}e^{e^{\beta_0+\beta_1 t_i}}}{\left(e^{\beta_0+\beta_1 t_i}-1\right)^2}\right]$$

$$E\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1^2}\right) = -\sum_{i=1}^{n} t_i^2\left[-(m_i-m_ip_i)\log\left(\frac{1}{1-p_i}\right)+m_ip_i\frac{\log\left(\frac{1}{1-p_i}\right)\left(\frac{p_i}{1-p_i}\right)-\log\left(\frac{1}{1-p_i}\right)^2\left(\frac{1}{1-p_i}\right)}{\left(\frac{p_i}{1-p_i}\right)^2}\right]$$

$$= -\sum_{i=1}^{n} m_it_i^2\left[\frac{-(1-p_i)\log\left(\frac{1}{1-p_i}\right)p_i}{p_i}+\frac{\log\left(\frac{1}{1-p_i}\right)(1-p_i)\,p_i-\log\left(\frac{1}{1-p_i}\right)^2(1-p_i)}{p_i}\right]$$

$$= \sum_{i=1}^{n} t_i^2\frac{m_i\,(1-p_i)\,(\log(1-p_i))^2}{p_i}$$

$$-\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} = -\sum_{i=1}^{n} t_i\left[-(m_i-y_i)\frac{e^{\beta_0+\beta_1 t_i}e^{-e^{\beta_0+\beta_1 t_i}}}{e^{-e^{\beta_0+\beta_1 t_i}}}+y_i\frac{e^{\beta_0+\beta_1 t_i}e^{-e^{\beta_0+\beta_1 t_i}}}{1-e^{-e^{\beta_0+\beta_1 t_i}}}\right]$$

$$= -\sum_{i=1}^{n} t_i\left[-(m_i-y_i)\,e^{\beta_0+\beta_1 t_i}+y_i\frac{e^{\beta_0+\beta_1 t_i}}{e^{e^{\beta_0+\beta_1 t_i}-1}}\right]$$

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0\partial \beta_1} = -\sum_{i=1}^{n} t_i\left[-(m_i-y_i)\,e^{\beta_0+\beta_1 t_i}+y_i\frac{e^{\beta_0+\beta_1 t_i}\left(e^{e^{\beta_0+\beta_1 t_i}}-1\right)-e^{2(\beta_0+\beta_1 t_i)}e^{e^{\beta_0+\beta_1 t_i}}}{\left(e^{\beta_0+\beta_1 t_i}-1\right)^2}\right]$$

$$
\begin{aligned}
E\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_0\partial\beta_1}\right) &= -\sum_{i=1}^{n} t_i \left[-\left(m_i - m_i p_i\right)\log\left(\frac{1}{1-p_i}\right) + m_i p_i \frac{\log\left(\frac{1}{1-p_i}\right)\left(\frac{p_i}{1-p_i}\right) - \log\left(\frac{1}{1-p_i}\right)^2\left(\frac{1}{1-p_i}\right)}{\left(\frac{p_i}{1-p_i}\right)^2}\right] \\
&= -\sum_{i=1}^{n} m_i t_i \left[\frac{-\left(1-p_i\right)\log\left(\frac{1}{1-p_i}\right)p_i}{p_i} + \frac{\log\left(\frac{1}{1-p_i}\right)\left(1-p_i\right)p_i - \log\left(\frac{1}{1-p_i}\right)^2\left(1-p_i\right)}{p_i}\right] \\
&= \sum_{i=1}^{n} t_i \frac{m_i\left(1-p_i\right)\left(\log\left(1-p_i\right)\right)^2}{p_i} \\
&= E\left(-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_0}\right)
\end{aligned}
$$

```
> X <- cbind(1, orings$temp)
> zetahat <- X %*% betahat
> phat <- 1 - exp(-exp(zetahat))
> a <- 6*(1 - phat)*(log(1-phat))^2/phat
> I11 <- sum(X[,1]^2*a)
> I12 <- sum(X[,1]*X[,2]*a)
> I22 <- sum(X[,2]^2*a)
> Iinv <- solve(matrix(c(I11, I12, I12, I22), 2, 2))
> (si_1 <- sqrt(Iinv[1,1]))

[1] 2.736517

> c(betahat[1] - 1.96*si_1, betahat[1] + 1.96*si_1)

[1]  5.498654 16.225802

> (si_2 <- sqrt(Iinv[2,2]))

[1] 0.04560421

> c(betahat[2] - 1.96*si_2, betahat[2] + 1.96*si_2)

[1] -0.2948815 -0.1161130
```

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

First we calculate the deviance for the model including temperature (full model) using the fomula on the page 11 of "Binomial Regression II" slides. We use $\hat{p}_i$ with the complementary log-log link.

```
> y <- orings$damage
> m <- rep(6, length(y))
> ylogxy <- function(x, y) ifelse(y == 0, 0, y*log(x/y))
> (D <- -2*sum(ylogxy(m*phat, y) + ylogxy(m*(1-phat), m - y)))

[1] 16.02857

> (df <- length(y) - length(betahat))

[1] 21
```

Next we calculate the deviance for the model without temperature (reduced model) using the same fomula on the page 11, but with $\hat{p}_i$ for this reduced model. This reduced model is equivalent with the reduced model we considered in "Likelihood Ratio test" of Challenger.pdf. So they will have the same $\hat{p}_i$.

```
> (phatN <- sum(y)/sum(m))

[1] 0.07971014

> (DN <- -2*sum(ylogxy(m*phatN, y) + ylogxy(m*(1-phatN), m - y)))
```

```
[1] 38.89766
```

```
> (dfN <- length(y) - 1)
```

```
[1] 22
```

Perform a likelihood ratio test.

```
> pchisq(DN - D, dfN - df, lower=FALSE) # p-value
```

```
[1] 1.734185e-06
```

We have very strong evidence that coefficient of the temperature $\neq 0$. So we prefer the model with temperature.

(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).

We follow “Confidence Interval for p” in Challenger.pdf using a complementary log-log link. MLEs and their standard errors should be obtained with the complementary log-log link.

```
> options(digits=16)
> si2 <- matrix(c(1, 31), 1, 2) %*% Iinv %*% matrix(c(1, 31), 2, 1)
> (p31 <- 1 - exp(-exp(betahat[1] + betahat[2]*31)))
```

```
[1] 1
```

```
> 1 - exp(-exp(betahat[1] + betahat[2]*31 - 1.96*sqrt(si2)))[1]
```

```
[1] 0.9984231914898988
```

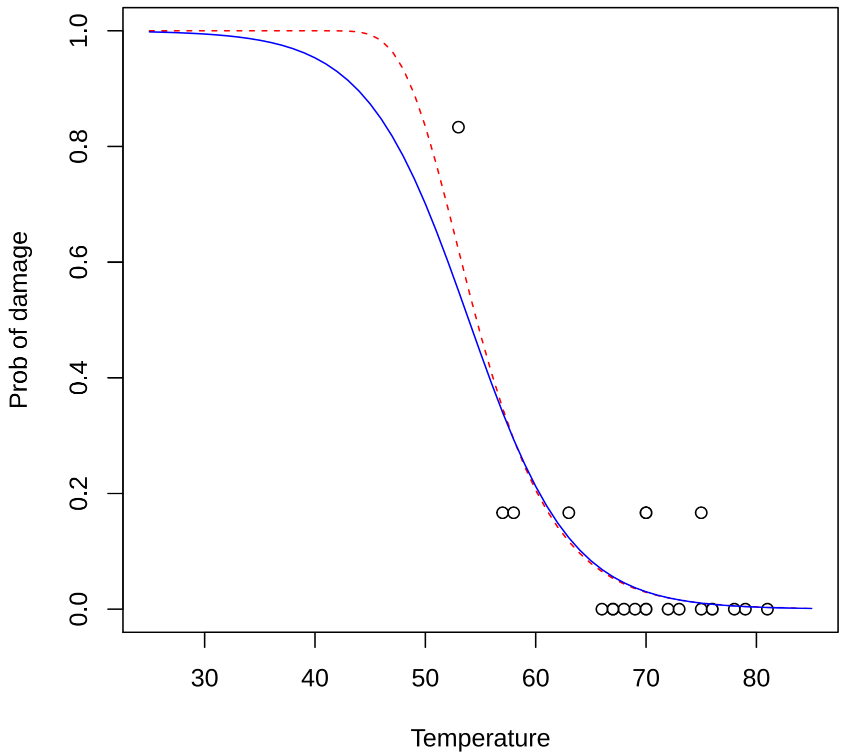
```
> 1 - exp(-exp(betahat[1] + betahat[2]*31 + 1.96*sqrt(si2)))[1]
```

```
[1] 1
```

(e) Make a plot comparing the fitted complementary log-log model to the fitted logit model. To obtain the fitted logit model, you are allowed to use the `glm` function.

Plot of the fitted c-log-log model (dashed line) and logit (solid line) models. They are very close for the observed data points, but the c-log-log model puts much less weight in the left tail, giving a notably larger fit when temperature equals 31°.

```
> plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim=c(0,1),
+      xlab="Temperature", ylab="Prob of damage")
> x <- seq(25,85,1)
> lines(x, 1 - exp(-exp(betahat[1] + betahat[2]*x)), col="red", lty=2)
> betalogit <- glm(cbind(damage,6-damage) ~ temp, family=binomial, orings)$coefficients
> lines(x, ilogit(betalogit[1] + betalogit[2]*x), col="blue")
```



2. The data frame 'pima_subset' contains a subset of the `pima` data set. For details of the `pima` data set, please see the practical problem 2 for the week 2. You can obtain 'pima_subset' using the commands:

```
> 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 ...
```

Using the 'pima_subset' data set, we will fit a model with `test` as a response and `bmi` as a predictor to see the relationship between the odds of a patient showing signs of diabetes and his/her bmi. The odds o and probability p are related by

$$o = \frac{p}{1-p} \quad p = \frac{o}{1+o}.$$

(a) (3 marks) Please estimate the amount of increase in log(odds) when bmi increases by 5.

(b) (3 marks) Give a 95% CI for the estimate.

You are allowed to use the `glm` function.

Solution

(a) Please estimate the amount of increase in log(odds) when bmi increases by 5.
Let o_x, η_x, o_{x+5}, η_{x+5} be the odds and linear response for a woman with `bmi` at x and $x+5$ respectively. Then, for binomial regression with logit link,

$$\begin{aligned}\log(o_{x+5}) - \log(o_x) &= \eta_{x+5} - \eta_x \\ &= \beta_{bmi}5\end{aligned}$$

We fit a binomial regression.

```
> 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 ...
> model <- glm(cbind(test, 1-test)~., family=binomial, data=pima_subset)
> summary(model)
Call:
glm(formula = cbind(test, 1 - test) ~ ., family = binomial, data = pima_subset)

Deviance Residuals:
             Min                1Q            Median                3Q
-1.9226771148525  -0.8919941100309  -0.6567660905735   1.2559424452913
             Max
 1.9560325992250

Coefficients:
                    Estimate        Std. Error  z value   Pr(>|z|)
(Intercept) -4.03681034866537  0.52783416246102 -7.64788 2.0433e-14 ***
bmi          0.09971684048953  0.01528411411199  6.52421 6.8359e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 676.78803680083  on 531  degrees of freedom
Residual deviance: 627.45577307249  on 530  degrees of freedom
AIC: 631.45577307249

Number of Fisher Scoring iterations: 4
```

A point estimate for $\beta_{bmi}5$ is

$$0.09972 \times 5 = 0.4986.$$

(b) Give a 95% CI for the estimate.
The estimate for $\beta_{bmi}5$ is a linear combination of normally distributed random variable that follows a normal distribution. The mean standard error of the estimate for $\beta_{bmi}5$ is 5 $\times$ standard error of the estimate for β_{bmi}.
95% CI for the estimate is

$$5(0.09972 \pm 1.959964 \times 0.01528) = (0.3488588, 0.6483412)$$

3. The inverse Gaussian distribution has p.d.f.

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

for $x > 0$, where $\mu > 0$ and $\lambda > 0$.

(a) (5 marks) Show that the inverse Gaussian distribution is an exponential family.

(b) (5 marks) Obtain the canonical link and the variance function.

[hint: you could consider $\theta = -1/\mu^2$.]

Solution

(a) Show that the inverse Gaussian distribution is an exponential family.
The inverse Gaussian has log density (for $\lambda > 0$ and $x > 0$)

$$\log f(x; \mu, \lambda) = \frac{1}{2} \log \frac{\lambda}{2\pi x^3} - \frac{\lambda x}{2\mu^2} + \frac{\lambda}{\mu} - \frac{\lambda}{2x}$$

Put $\theta = -1/\mu^2$ and $\phi = 2/\lambda$ then we have

$$\log f(x; \nu, \lambda) = \frac{x\theta + 2\sqrt{-\theta}}{\phi} + \frac{1}{2} \log \frac{1}{\phi \pi x^3} - \frac{1}{\phi x}$$

This is in the form of an exponential family, with

$$\begin{aligned} b(\theta) &= -2\sqrt{-\theta} \\ a(\phi) &= \phi \\ c(x, \phi) &= \frac{1}{2} \log \frac{1}{\phi \pi x^3} - \frac{1}{\phi x} \end{aligned}$$

Note that with this parameterisation we have $\theta < 0$ and $\phi > 0$.

(b) Obtain the canonical link and the variance function.
For the canonical link g we have $g(\mu) = \theta = -1/\mu^2$, so $g(x) = -1/x^2$.
$b'(\theta) = \mu$ is the mean. As $b'(\theta) = (-\theta)^{-\frac{1}{2}}$ and $b''(\theta) = \frac{1}{2}(-\theta)^{-\frac{3}{2}}$, the variance is $b''(\theta)a(\phi) = \frac{1}{2}(\frac{1}{\mu^2})^{-\frac{3}{2}}\frac{2}{\lambda} = \mu^3/\lambda = v(\mu)a(\phi) = v(\mu)2/\lambda$. So $v(\mu) = \mu^3/2$.