MAST30027: Modern Applied Statistics

Assignment 1 Solution 2023

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.

   Your report should include the following:

   (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 probit model to the fitted logit model. To obtain the fitted logit model, you are allowed to use the `glm` function.

Solution

For a binomial regression with a probit link we have $y_i \sim \text{bin}(m_i, \Phi(\eta_i))$, where ϕ is the density of the standard normal, Φ is its cdf, and $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$, so

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_i \left[y_i \log\Phi(\eta_i) + (m_i - y_i)\log(1-\Phi(\eta_i))\right] \\
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_i \left[\frac{y_i\phi(\eta_i)}{\Phi(\eta_i)} - \frac{(m_i-y_i)\phi(\eta_i)}{1-\Phi(\eta_i)}\right] \\
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1} &= \sum_i \left[\frac{y_i\phi(\eta_i)x_{i1}}{\Phi(\eta_i)} - \frac{(m_i-y_i)\phi(\eta_i)x_{i1}}{1-\Phi(\eta_i)}\right] \\
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0^2} &= \sum_i \left[\frac{-y_i\phi(\eta_i)^2}{\Phi(\eta_i)^2} + \frac{-y_i\phi(\eta_i)\eta_i}{\Phi(\eta_i)}\right. \\
&\qquad \left. - \frac{(m_i-y_i)\phi(\eta_i)^2}{(1-\Phi(\eta_i))^2} - \frac{-(m_i-y_i)\phi(\eta_i)\eta_i}{1-\Phi(\eta_i)}\right] \\
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_1^2} &= \sum_i x_{i1}^2 \left[\frac{-y_i\phi(\eta_i)^2}{\Phi(\eta_i)^2} + \frac{-y_i\phi(\eta_i)\eta_i}{\Phi(\eta_i)}\right. \\
&\qquad \left. - \frac{(m_i-y_i)\phi(\eta_i)^2}{(1-\Phi(\eta_i))^2} - \frac{-(m_i-y_i)\phi(\eta_i)\eta_i}{1-\Phi(\eta_i)}\right] \\
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_0 \partial \beta_1} &= \sum_i x_{i1} \left[\frac{-y_i\phi(\eta_i)^2}{\Phi(\eta_i)^2} + \frac{-y_i\phi(\eta_i)\eta_i}{\Phi(\eta_i)}\right. \\
&\qquad \left. - \frac{(m_i-y_i)\phi(\eta_i)^2}{(1-\Phi(\eta_i))^2} - \frac{-(m_i-y_i)\phi(\eta_i)\eta_i}{1-\Phi(\eta_i)}\right]
\end{aligned}
$$

Using $\mathbb{E}(y_i) = m_i\Phi(\eta_i)$, we can get

$$
-\mathbb{E}\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \sum_i m_i\phi(\eta_i)^2 \left[\frac{1}{\Phi(\eta_i)} + \frac{1}{1-\Phi(\eta_i)}\right] \begin{bmatrix} 1 & x_{i1} \\ x_{i1} & x_{i1}^2 \end{bmatrix}
$$

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

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

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

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

```
> X <- cbind(1, orings$temp)
> zetahat <- X %*% betahat
> a <- dnorm(zetahat)^2*(1/pnorm(zetahat) + 1/(1-pnorm(zetahat)))
> I11 <- sum(6*X[,1]^2*a)
> I12 <- sum(6*X[,1]*X[,2]*a)
> I22 <- sum(6*X[,2]^2*a)
> Iinv <- solve(matrix(c(I11, I12, I12, I22), 2, 2))
> c(betahat[1] - 1.96*sqrt(Iinv[1,1]), betahat[1] + 1.96*sqrt(Iinv[1,1]))
```

```
[1] 2.239700 8.943748
```

```
> c(betahat[2] - 1.96*sqrt(Iinv[2,2]), betahat[2] + 1.96*sqrt(Iinv[2,2]))
```

```
[1] -0.15784765 -0.05375385
```

Comparing with `glm` output, we see that the estimates and standard errors agree with ours to four significant figures.

```
> probitmod <- glm(cbind(damage,6-damage) ~ temp, family=binomial(link=probit), orings)
> summary(probitmod)
```

```
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial(link = probit),
    data = orings)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0134  -0.7761  -0.4467  -0.1581   1.9983

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.59145    1.71055   3.269  0.00108 **
temp        -0.10580    0.02656  -3.984 6.79e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 18.131  on 21  degrees of freedom
AIC: 34.893

Number of Fisher Scoring iterations: 6
```

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

First we calculate the deviance for the model including temperature.

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

[1] 18.13058

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

[1] 21
```

Next we fit the null model and use a likelihood ratio test.

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

[1] 0.07971014

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

[1] 38.89766

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

[1] 22

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

[1] 5.186684e-06
```

We have very strong evidence that $\beta_1 \neq 0$.

Note that our deviance calculations agree with the output from `glm`.

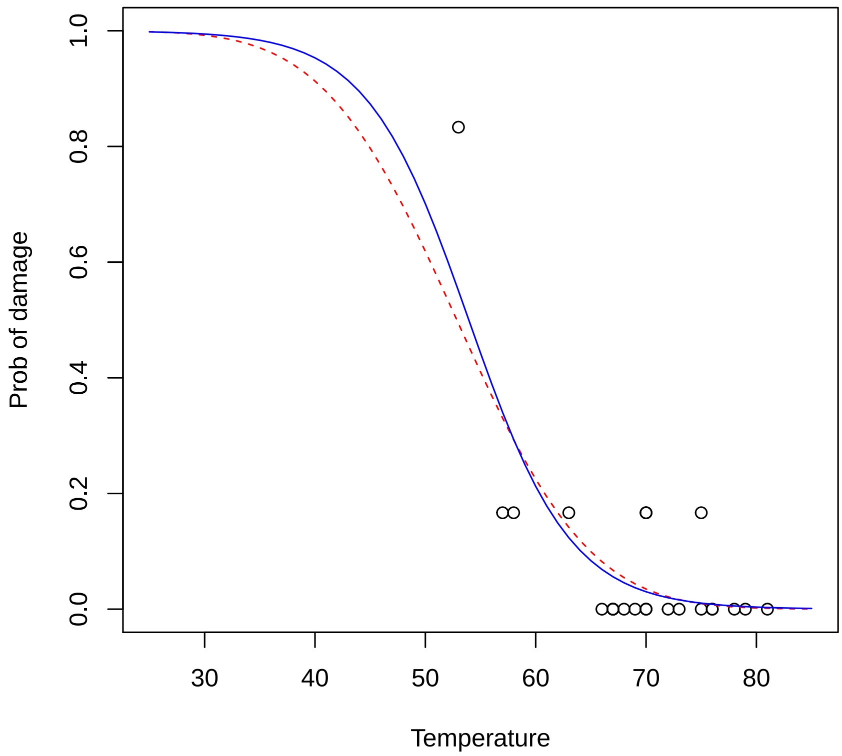
(d) (3 marks) Compute an estimate of the probability of damage when the temperature equals 31 Fahrenheit (your estimate should come with a 95% CI, as all good estimates do).

```
> si2 <- matrix(c(1, 31), 1, 2) %*% Iinv %*% matrix(c(1, 31), 2, 1)
> (p31 <- pnorm(betahat[1] + betahat[2]*31))

[1] 0.9896084

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

[1] 0.7108118

> pnorm(betahat[1] + betahat[2]*31 + 1.96*sqrt(si2))[1]

[1] 0.9999763
```

(e) (2 marks) 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.

They are very close, but the probit model puts a little more weight in the tails.

```
> 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, pnorm(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 binomial regression with a logit link 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 the log(odds) when the bmi increases by 7.

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

You are allowed to use the `glm` function.

Solution

(a) (3 marks) Please estimate the amount of increase in the log(odds) when the bmi increases by 7.

Let o_x, η_x, o_{x+7}, η_{x+7} be the odds and linear response for a woman with `bmi` at x and $x+7$ respectively. Then, for binomial regression with logit link,

$$\begin{aligned} \log(o_{x+7}) - \log(o_x) &= \eta_{x+7} - \eta_x \\ &= 7\beta_{bmi} \end{aligned}$$

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      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
```

A point estimate for $7\beta_{bmi}$ is

$$7 \times 0.09972 = 0.69804.$$

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

The standard error of the estimate for $7\beta_{bmi}$ is $7 \times$ standard error of the estimate for β_{bmi}. 95% CI for the estimate is

$$7(0.09972 \pm 1.959964 \times 0.01528) = (0.48840, 0.90768)$$

3. The gamma distribution with shape $\nu > 0$ and rate $\lambda > 0$ has p.d.f.

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

for $x > 0$.

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

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

Solution

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

The gamma distribution with shape $\nu > 0$ and rate $\lambda > 0$ has log density

$$\begin{aligned}\log f(x;\nu,\lambda) &= (\nu-1)\log(x) - \lambda x + \nu\log(\lambda) - \log(\Gamma(\nu)) \\ &= \frac{x(-\lambda/\nu) + \log(\lambda/\nu)}{1/\nu} - \nu\log(1/\nu) + (\nu-1)\log(x) - \log(\Gamma(\nu))\end{aligned}$$

Put $\theta = -\lambda/\nu$ and $\phi = 1/\nu$ then we have

$$\log f(x;\nu,\lambda) = \frac{x\theta - \log(-1/\theta)}{\phi} - \frac{\log(\phi)}{\phi} + \left(\frac{1}{\phi} - 1\right)\log(x) - \log(\Gamma(1/\phi))$$

This is in the form of an exponential family, with

$$\begin{aligned}b(\theta) &= \log(-1/\theta) \\ a(\phi) &= \phi \\ c(x,\phi) &= \frac{-\log(\phi) + (1-\phi)\log(x) - \phi\log(\Gamma(1/\phi))}{\phi}\end{aligned}$$

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

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

For the canonical link g we have $g(\mu) = \theta$. Here $\mu = \nu/\lambda = -1/\theta$, so $g(x) = -1/x$. (Note that in practice people tend to use the inverse link $x \mapsto 1/x$ rather than $x \mapsto -1/x$, because it is convenient to keep things positive.) The variance is $\nu/\lambda^2 = \phi\mu^2 = a(\phi)v(\mu)$. That is, the variance function is $v(\mu) = \mu^2$.