MATH38172 Generalised Linear Models

Computer Lab 2

Inference for logistic and Poisson regression

Wald inference

For logistic and Poisson regression, it is easy to obtain Wald tests and intervals from the output of the R command `summary()`, passing the fitted model object as an argument. As an example consider the Challenger data.

Recall that the fitted model is

$$6Y_i \sim \text{Binomial}(6, \mu_i)$$
$$\log \frac{\mu_i}{1-\mu_i} = \beta_0 + \beta_1 x_i$$

where Y_i is the proportion of O-rings damaged on the ith launch, μ_i is the probability of an individual O-ring being damaged, and β_0, β_1 are unknown parameters.

Now consider the following R code (assuming you have already loaded the orings data):

```
chal1 <- glm( damage/6 ~ temp, weights=rep(6,nrow(orings)), family=binomial, data=orings)
summary(chal1)
```

The following table is given as part of the output:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
 (Intercept) 11.66299    3.29626   3.538 0.000403 ***
 temp        -0.21623    0.05318  -4.066 4.78e-05 ***
 ---
 Signif. codes:
 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The output shows for each parameter

- the maximum likelihood estimate $\hat{\beta}_j$
- the standard error, $\text{s.e.}(\hat{\beta}_j) = [\mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}]_{jj}$
- the test statistic (z-value) and p-value for a Wald test of the hypothesis

$$H_0 : \beta_j = 0 \qquad \text{vs} \qquad H_1 : \beta_j \neq 0\,.$$

The MLEs and standard errors can easily be used to conduct other Wald inferences about the parameter β_j. For example, the end-points of a 95% Wald confidence interval for β_{Temp} can be found using the formula $\hat{\beta}_j \pm z_{0.025}\,\text{s.e.}(\hat{\beta}_j)$ as follows

```
-0.21623 + c(-1,1)*qnorm(0.975)*0.05318
```

```
## [1] -0.3204609 -0.1119991
```

Recall that the `qnorm` function gives quantiles of a normal distribution. For example:

(i) to compute the 0.95 quantile of the standard normal (i.e. $N(0,1)$) distribution use `qnorm(0.95)`.

(ii) to compute the 0.95 quantile of a normal distribution with mean 3 and variance 4 (i.e. $N(3,2^2)$), use `qnorm(0.95,mean=3,sd=2)`.

Profile likelihood intervals

Profile likelihood intervals for each parameter β_j can be easily obtained by passing the fitted model object as an argument to the `confint()` command, e.g for the Challenger model:

```
confint(chal1, level=0.9)
```

```
## Waiting for profiling to be done...
```

```
##                    5 %       95 %
## (Intercept)  6.5165681 17.5027878
## temp        -0.3120442 -0.1347199
```

Note that we can sometimes find profile likelihood intervals for functions of the parameter by fitting a GLM with different explanatory variables. For example, suppose we want to find a 95% profile likelihood confidence interval for the probability of O-ring damage at $30°F$, which we denote μ^*. Note that

$$\log \frac{\mu^*}{1-\mu^*} = \beta_0 + \beta_1.30 = \lambda$$

The logistic model can be rewritten equivalently (i.e. reparameterised) as

$$\begin{aligned} 6Y_i &\sim \text{Binomial}(6, \mu_i) \\ \log \frac{\mu_i}{1-\mu_i} &= \beta_0 + \beta_1 x_i = \beta_0 + \beta_1 30 + \beta_1(x_i - 30) \\ &= \lambda + \beta_1(x_i - 30) \end{aligned}$$

This is now a Binomial GLM with explanatory variables 1 and $(x-30)$, and parameters λ and β_1. This reparameterised model can be fitted as follows:

```
chalR <- glm(damage/6~I(temp-30), family=binomial, weights=rep(6,23),data=orings)
coef(chalR)
```

```
##  (Intercept) I(temp - 30)
##    5.1759798   -0.2162337
```

This gives $\hat{\lambda} = 5.176$. We can obtain a profile likelihood CI for λ via:

```
confint(chalR)
```

```
## Waiting for profiling to be done...
```

```
##                  2.5 %    97.5 %
## (Intercept)   1.938259  8.799512
## I(temp - 30) -0.332657 -0.120179
```

Using the fact that $\mu^* = e^\lambda/(1+e^\lambda)$, we can transform the interval to get a 95% profile likelihood confidence interval for μ^*:

```
cil <- c( 1.938259,  8.799512 )
exp(cil)/(1+exp(cil))
```

```
## [1] 0.8741608 0.9998492
```

Estimated variance-covariance matrix

Recall that for a model with vector parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ the approximate distribution of the maximum likelihood estimator is $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}, \mathcal{I}(\boldsymbol{\beta})^{-1})$, with asymptotic variance matrix $\mathcal{I}(\boldsymbol{\beta})^{-1}$ depending on the true values of the parameters.

The estimated asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ is obtained by plugging in the MLE of $\boldsymbol{\beta}$:

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1}\,.$$

For logistic and Poisson regression, the estimated asymptotic variance matrix can easily be obtained in R by passing the fitted model object to the command `vcov()`. For example, for the Challenger model:

```
vcov(chal1)
```

```
##             (Intercept)         temp
## (Intercept)   10.865351 -0.174240974
## temp          -0.174241  0.002827797
```

Delta method

Recall that the estimated asymptotic variance matrix can be used in the delta method. In the lecture notes we showed that $h(\boldsymbol{\beta})$ can be estimated by $h(\hat{\boldsymbol{\beta}})$, and a $100\kappa\%$ confidence interval for $h(\boldsymbol{\beta})$ is given by the end points

$$h(\hat{\boldsymbol{\beta}}) \pm z_{(1-\kappa)/2}\,\mathrm{s.e.}(\hat{\boldsymbol{\beta}}) = h(\hat{\boldsymbol{\beta}}) \pm z_{(1-\kappa)/2}\sqrt{\nabla h(\hat{\boldsymbol{\beta}})^T \mathcal{I}(\hat{\boldsymbol{\beta}})^{-1} \nabla h(\hat{\boldsymbol{\beta}})}$$

For example suppose that x^* is the temperature at which there is a 50% probability of O-ring damage. Note that $x^* = h(\boldsymbol{\beta}) = -\beta_0/\beta_1$, and this can be estimated by $h(\hat{\boldsymbol{\beta}}) = -\hat{\beta}_0/\hat{\beta}_1$:

```
betah <- coef(chal1)
betah
```

```
## (Intercept)        temp
##  11.6629897  -0.2162337
```

```
xh <- -betah[1]/betah[2]
xh
```

```
## (Intercept)
##    53.93697
```

Note that $\hat{\beta}_0$ is `betah[1]` and $\hat{\beta}_1$ is `betah[2]`, due to the fact that R indexes the elements of a vector beginning with a 1.

To compute the 95% CI for x^* note that $\nabla h = (-1/\beta_1, \beta_0/\beta_1^2)^T$ and use the following

```
gradh <- c( -1/betah[2], betah[1]/(betah[2]^2) )
gradh
```

```
##        temp (Intercept)
##    4.624627  249.438380
```

```
se <- sqrt( t(gradh)%*% vcov(chal1)%*%gradh)
se
```

```
##          [,1]
## [1,] 2.515673
```

```
xh + c(-1,1)*qnorm(0.975)*c(se)
```

```
## [1] 49.00635 58.86760
```

This agrees with the results in the lecture – however this time we didn't have to do the calculations by hand.

Note that in the computation of the standard error we used the following commands:

- `t` gives the transpose of a vector/matrix
- `%*%` corresponds to matrix multiplication.

Exercises 1

1 For the Challenger example:

a) Use the output of `vcov` to compute the standard error of $\hat{\beta}_0$ and $\hat{\beta}_1$, and compare the answer to the output of `summary`.

b) Calculate 95% Wald and profile likelihood intervals for the intercept parameter.

c) Compare the width and symmetry of the Wald and profile likelihood intervals for both parameters. What do you notice? Which do you expect to be more accurate?

d) Conduct Wald and profile likelihood tests of the hypothesis $H_0 : \beta_{Temp} = -0.33$ vs $H_1 : \beta_{Temp} \neq -0.33$ at a 5% significance level.

2 Data have been collected in each of $n = 31$ areas. In the ith area, we have:

- x_i: the number of Eucalypt trees (named `Eucs` in R)
- y_i: the number of noisy miner birds (also called the *abundance*, named `Minerab` in R).

The data are given in the file `nminer2.csv`. Using R:

a) Fit a Poisson regression model with abundance as the response variable and the number of Eucalypt trees as the explanatory variable, and write down the fitted model.

b) Using *three different methods*, find a 95% confidence interval for ϕ, the expected abundance in an area with 6 Eucalypt trees. Comment on which method you think is likely to be best.

c) Using *three different methods*, find a 95% confidence interval for λ, the percentage increase in expected abundance associated with one additional Eucalypt tree. Comment on which method is likely to be best.

Fisher scoring

Recall from Lecture 6 that for a statistical model with parameter vector $\boldsymbol{\theta}$ the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ can be approximated numerically via an iterative procedure known as Fisher scoring, which proceeds as follows:

1. Set initial approximate values for the estimates, $\hat{\boldsymbol{\theta}}^0$
2. For $k = 0, 1, 2, \ldots$ until convergence set

$$\hat{\boldsymbol{\theta}}^{k+1} = \hat{\boldsymbol{\theta}}^k + \mathcal{I}(\hat{\boldsymbol{\theta}}^k)^{-1}\mathbf{u}(\hat{\boldsymbol{\theta}}^k)$$

Recall further from Lectures 5-6 that the score vector and Fisher information matrix for simple logistic regression are

$$\mathbf{u}(\boldsymbol{\beta}) = \begin{bmatrix} \sum_i m_i(y_i - \mu_i) \\ \sum_i x_i m_i(y_i - \mu_i) \end{bmatrix}, \qquad \mathcal{I}(\boldsymbol{\beta}) = \begin{bmatrix} \sum_i W_i & \sum_i W_i x_i \\ \sum_i W_i x_i & \sum_i W_i x_i^2 \end{bmatrix},$$

where $W_i = m_i\mu_i(1 - \mu_i)$, and that these can be calculated using the R functions `u` and `FIM` given in the file `Lab2-functions.R`.

Exercises 2

3 Using these functions with the Challenger data:

a) calculate the estimated asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$ and compare your result to the output of `vcov`.

b) without using the `glm` function, carry out 10 iterations of Fisher scoring starting with initial vector $\hat{\boldsymbol{\beta}}^0 = (0, 0)$

4 Recall from Example Sheet 3 that for simple Poisson regression the score vector and information matrix are

$$\mathbf{u}(\boldsymbol{\beta}) = \begin{bmatrix} \sum_i (y_i - \mu_i) \\ \sum_i x_i(y_i - \mu_i) \end{bmatrix}, \qquad \mathcal{I}(\boldsymbol{\beta}) = \begin{bmatrix} \sum_i \mu_i & \sum_i \mu_i x_i \\ \sum_i \mu_i x_i & \sum_i \mu_i x_i^2 \end{bmatrix}$$

where $\mu_i = e^{\beta_0 + \beta_1 x_i}$.

a) by adapting the given functions `u` and `FIM`, write functions to calculate the score vector and Fisher information matrix for simple Poisson regression

For the noisy miner model in Question 2:

b) use your function to calculate the estimated asymptotic variance matrix of $\hat{\boldsymbol{\beta}}$, and compare the result to `vcov`

c) use your functions to carry out 15 iterations of Fisher scoring, and verify that the results coincide with those from `glm`.