MAST30027: Modern Applied Statistics

Assignment 1

Due: 1:00pm Friday 14 August (week 3)

This assignment is worth 3 1/3% of your total mark. Fill in a plagiarism declaration form and hand it in together with this assignment.

Fit a binomial regression model to the O-rings data from the Challenger disaster, using a *probit* link. You must use R but may not use the `glm` function; I want you to work from first principles.

Your solution should include the following:

1. parameter estimates;
2. 95% CIs for the parameter estimates;
3. a likelihood ratio test for the significance of the temperature coefficient;
4. 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);
5. a plot comparing the fitted probit model to the fitted logit model.

Show your working, that is, the R code you use.

Solution

For a binomial regression with a probit link we have $y_i \sim \text{bin}(n_i, \Phi(\eta_i))$, where $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$, so

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_i [y_i \log \Phi(\eta_i) + (n_i - y_i)\log(1 - \Phi(\eta_i))] \\
\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} &= \sum_i \left[\frac{y_i \phi(\eta_i) x_{i,j}}{\Phi(\eta_i)} + \frac{(n_i - y_i)\phi(\eta_i) x_{i,j}}{1 - \Phi(\eta_i)} \right] \\
\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} &= \sum_i x_{i,j} x_{i,k} \left[\frac{-y_i \phi(\eta_i)^2}{\Phi(\eta_i)^2} + \frac{-y_i \phi(\eta_i)\eta_i}{\Phi(\eta_i)} \right. \\
&\qquad \left. + \frac{-(n_i - y_i)\phi(\eta_i)^2}{(1 - \Phi(\eta_i))^2} + \frac{(n_i - y_i)\phi(\eta_i)\eta_i}{1 - \Phi(\eta_i)} \right] \\
-\mathbb{E}\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k} &= \sum_i x_{i,j} x_{i,k} n_i \phi(\eta_i)^2 \left[\frac{1}{\Phi(\eta_i)} + \frac{1}{1 - \Phi(\eta_i)} \right]
\end{aligned}
$$

1. Estimating $\boldsymbol{\beta}$

```
> 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
```

2. 95% CIs for β_0 and β_1

```
> 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
```

3. Testing $H_0 : \beta_1 = 0$. 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`.

4. Forecast for the probability of failure when the temperature is 31° Fahrenheit.

```
> 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]
```

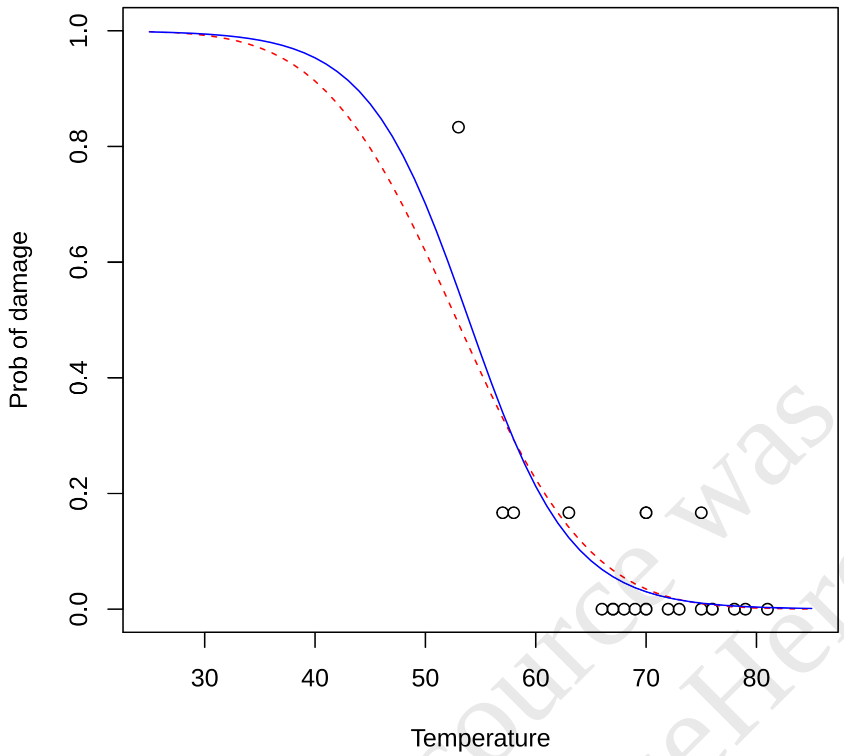
```
[1] 0.7108118
```

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

```
[1] 0.9999763
```

5. Plot of the fitted probit (dashed line) and logit (solid line) models. 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")
```


Prob of damage
0.0
0.2
0.4
0.6
0.8
1.0
30
40
50
60
70
80
Temperature