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 \( \eta_i = x_i^T \beta \), so

\[
l(\beta) = \sum_i \left[ y_i \log \Phi(\eta_i) + (n_i - y_i) \log(1 - \Phi(\eta_i)) \right]
\]

\[
\frac{\partial l(\beta)}{\partial \beta_j} = \sum_i \left[ y_i \phi(\eta_i) x_{i,j} \Phi(\eta_i) + \frac{(n_i - y_i) \phi(\eta_i) x_{i,j}}{1 - \Phi(\eta_i)} \right]
\]

\[
\frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} = \sum_i x_{i,j} x_{i,k} \left[ \frac{-y_i \phi(\eta_i)^2}{\Phi(\eta_i)^2} + \frac{(n_i - y_i) \phi(\eta_i) \eta_i}{\Phi(\eta_i)} \right] + \frac{(n_i - y_i) \phi(\eta_i)^2}{(1 - \Phi(\eta_i))^2} + \frac{(n_i - y_i) \phi(\eta_i) \eta_i}{1 - \Phi(\eta_i)}
\]

\[
-\mathbb{E} \frac{\partial^2 l(\beta)}{\partial \beta_j \partial \beta_k} = \sum_i x_{i,j} x_{i,k} n_i \phi(\eta_i)^2 \left[ \frac{1}{\Phi(\eta_i)} + \frac{1}{1 - \Phi(\eta_i)} \right]
\]

1. Estimating \( \beta \)

```r
> library(faraway)
> data(orings)
> logL <- function(beta, orings) {
+  y <- orings$damage
+  X <- chind(1, orings$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)
```

```r
1
```
2. 95% CIs for $\beta_0$ and $\beta_1$

\[
X \leftarrow \text{cbind}(1, \text{orings$\text{temp}$})
\]
\[
\text{zetahat} \leftarrow X \%\% \beta
\]
\[
a \leftarrow \text{dnorm(zetahat)}^2 \times (1 / \text{pnorm(zetahat)} + 1 / (1 - \text{pnorm(zetahat))))
\]
\[
I_{11} \leftarrow \sum(6 \times X[,1]^2 \times a)
\]
\[
I_{12} \leftarrow \sum(6 \times X[,1] \times X[,2] \times a)
\]
\[
I_{22} \leftarrow \sum(6 \times X[,2]^2 \times a)
\]
\[
I_{\text{inv}} \leftarrow \text{solve(matrix}(c(I_{11}, I_{12}, I_{12}, I_{22}), 2, 2))
\]
\[
c(\text{betahat}[1] - 1.96 \times \sqrt{I_{\text{inv}}[1,1]}, \text{betahat}[1] + 1.96 \times \sqrt{I_{\text{inv}}[1,1]})
\]
\[1\] 2.239700 8.943748
\[
c(\text{betahat}[2] - 1.96 \times \sqrt{I_{\text{inv}}[2,2]}, \text{betahat}[2] + 1.96 \times \sqrt{I_{\text{inv}}[2,2]})
\]
\[1\] -0.15784765 -0.05375385

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

\[
\text{probitemod} \leftarrow \text{glm}(\text{cbind(damage, 6-damage)} \sim \text{temp, family=binomial(link=probit), orings})
\]
\[
\text{summary(probitemod)}
\]

Call:
\[
\text{glm(formula = cbind(damage, 6 - damage) \sim temp, family = binomial(link = probit), data = orings)}
\]

Deviance Residuals:
\[
\begin{array}{lllll}
\text{Min} & 1Q & \text{Median} & 3Q & \text{Max} \\
-1.0134 & -0.7761 & -0.4467 & -0.1581 & 1.9983
\end{array}
\]

Coefficients:
\[
\begin{array}{lccccc}
\text{Estimate} & \text{Std. Error} & z \text{ value} & \text{Pr(>|z|)} \\
(\text{Intercept}) & 5.59145 & 1.71055 & 3.269 & 0.00108 \star* \\
\text{temp} & -0.10580 & 0.02656 & -3.984 & 6.79\times10^{-5} \star\star\star
\end{array}
\]

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.

\[
y \leftarrow \text{orings$\text{damage}$}
\]
\[
n \leftarrow \text{rep}(6, \text{length}(y))
\]
\[
y\text{logxy} \leftarrow \text{function}(x, y) \text{ ifelse}(y == 0, 0, y \times \text{log}(x/\text{y}))
\]
\[
\text{phat} \leftarrow \text{pnorm(zetahat)}
\]
\[
(D \leftarrow -2 \times \text{sum}(y\text{logxy}(n*\text{phat}, y) + y\text{logxy}(n*(1-\text{phat}), n - y)))
\]
\[1\] 18.13058
Next we fit the null model and use a likelihood ratio test.

\[
\text{Next we fit the null model and use a likelihood ratio test.}
\]

\[
> (\text{df} <- \text{length}(y) - \text{length}(\text{betahat}))
\]

\[
[1] 21
\]

\[
\text{Next we fit the null model and use a likelihood ratio test.}
\]

\[
> (\text{phatN} <- \text{sum}(y)/\text{sum}(n))
\]

\[
[1] 0.07971014
\]

\[
> (\text{DN} <- -2*\text{sum}(y \log_2(x) + x \log_2(1-\text{phatN})) + 2 \cdot n)
\]

\[
[1] 38.89766
\]

\[
> (\text{dfN} <- \text{length}(y) - 1)
\]

\[
[1] 22
\]

\[
> \text{pchisq}(\text{DN} - D, \text{dfN} - \text{df}, \text{lower=FALSE}) \# p-value
\]

\[
[1] 5.186684e-06
\]

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.

\[
\text{Forecast for the probability of failure when the temperature is 31° Fahrenheit.}
\]

\[
> \text{si2 <- matrix(c(1, 31), 1, 2) \%*\% inv \%*\% matrix(c(1, 31), 2, 1)}
\]

\[
> \text{(p31 <- pnorm(betahat[1] + betahat[2]*31))}
\]

\[
[1] 0.9896084
\]

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

\[
[1] 0.7108118
\]

\[
> \text{pnorm(betahat[1] + betahat[2]*31 + 1.96*sqrt(si2))}[1]
\]

\[
[1] 0.9999763
\]

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.

\[
\text{Plot of the fitted probit (dashed line) and logit (solid line) models. They are very close, but the probit model puts a little more weight in the tails.}
\]

\[
> \text{plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim=c(0,1),}
\]

\[
+ xlab="Temperature", ylab="Prob of damage")
\]

\[
> x <- \text{seq}(25,85,1)
\]

\[
> \text{lines(x, pnorm(betahat[1] + betahat[2]*x), col="red", lty=2)}
\]

\[
> \text{betalogit <- glm(cbind(damage,6-damage) ~ temp, family=binomial, orings)$coefficients}
\]

\[
> \text{lines(x, ilogit(betalogit[1] + betalogit[2]*x), col="blue")}
\]