8004 Homework 9

Nooreen Dabbish
April 9, 2015

1 Data is generated from the exponetial distribution with density

$$f(y) = \lambda \exp(-\lambda y), \text{ where } \lambda, y > 0.$$

(a) Show that it belongs to the exponential family distributions be indentifyting θ, b(θ), ϕ, a(ϕ) and c(y;ϕ).

An exponential family distribution can be written in the form

$$\exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right\}.$$

We write:

$$f(y|\lambda) = \exp(-\lambda y + \log\lambda)$$

and equate θ = -λ, b(θ) = -log λ =-log(-θ), (note that λ >0, so θ < 0), ϕ = 1, a(ϕ) = 1, c(y;ϕ) = 1.

(b) What is the canonical link and variance functions for a GLM with the response following the exponential distribution?

The link function connects the linear predictor μ to the parameter θ in the exponential family distribution definition above. To find the canonical link, we want μ = E(Y) = b'(θ). we find the first moment of y:

$$EY = \int_0^\infty \lambda y e^{-y} dy = \frac{1}{\lambda}$$

$$\text{Note } b'(\theta) = \frac{-1}{\theta} = \frac{1}{\lambda} = \mu \text{ and } b''(\theta) = \frac{1}{\theta^2} = \frac{1}{\lambda^2} = \mu^2$$

We write θ as a funcion of μ

$$\theta(\mu) = -\frac{1}{\mu}$$

$$b'^{-1}(\cdot) = \text{negative inverse function.}$$

Since var(Y) = b"(θ)a(ϕ), and a(ϕ) = 1, var(Y) = b"(θ) = μ^2.

(c) Is there any practical difficulty for using the canonical link in practice?

Especially in small samples, canoncial links have desirable properties. However, they may not be the best fit for a model (McCullagh and Nelder pg 32).

Note in this case that the exponential mean is restricted to positive values. However our μ is a linear combination of predictors. This does not guarantee a positive restriction on our estimates of the mean.

(d) Express the deviance as a function of y$_i$ and fitted mean μ_i (i = 1, . . . , n).

We have scaled deviance given by

$$\frac{D(y;\hat{\mu})}{\phi} = 2\sum \frac{w_i}{\phi}\{y_i(\widetilde{\theta}_i - \hat{\theta}_i) - b(\widetilde{\theta}_i) + b(\hat{\theta}_i)\}$$

with a(ϕ) = ϕ/w, $\tilde{\theta} = \theta(y)$ denoting the full model (n parameter) estimate of θ, and $\hat{\theta} = \theta(\hat{\mu})$ denoting the null model (one parameter) estimate of θ.

Evaluating for b(θ) = - log (- θ) and ϕ = 1, w$_i$ = 1 gives

$$D(y;\hat{\mu}) = 2\sum\left(y_i(\tilde{\theta}_i - \hat{\theta}_i) + \log\left(\frac{\tilde{\theta}_i}{\hat{\theta}_i}\right)\right)$$

From above, we have that $\theta(\mu) = -\frac{1}{\mu}$. Evaluating for $\hat{\theta}_i = 1/\hat{\mu}_i$) and $\tilde{\theta}_i = y_i$ gives

$$D(y;\hat{\mu}) = 2\sum\left(y_i\left(-\frac{1}{y_i} + \frac{1}{\hat{\mu}_i}\right) + \log\left(\frac{\hat{\mu}_i}{y_i}\right)\right) = 2\sum\left(\left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right) + \log\left(\frac{\hat{\mu}_i}{y_i}\right)\right)$$

2 Consider the Orings data of Faraway(2006) where the number of damaged ones out of six orings and corresponding temperatures of space shuttle launches are recorded.

```
library(faraway)
data(orings)
```

(a) Construct the appropriate test statistic for testing the effect of the temperature. State the approriate null distribution and give the p-value.

(Reference Chapter 2 of Faraway 2006)

We will use the deviance as a log likelihood test to test for the effect of temp. The null model, with only an intercept, is nested in the model including temperature. The difference of deviance measures D$_S$ - D$_L$ follows a χ^2 distribution with degrees of freedom equal to the number of parameters excluded from the nested model, in this case df=1 under the null hypothesis that the additional parameters do not contribute significantly to the model.

We write:

```
logitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial, orings)
devtest <- pchisq(logitmod$null-logitmod$deviance,
                  logitmod$df.null-logitmod$df.residual,lower=FALSE)
```

And obtain a highly significant p-value of 2.747e-06.

(b) Does it affect the conclusion by changing the link function to probit and other link functions.

The deviance function derivation in part 1d) illustrated that the form of the link function will impact the form of the deviance function. It therefore seems possible that changing the link function could alter the conclusion. With such a highly significant p-value for the effect of temperature, this seems unlikely. Let's see.

(b).1 Probit link function

```
probitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=probit), orings)
probitdevtest <- pchisq(probitmod$null-probitmod$deviance,
                  probitmod$df.null-probitmod$df.residual,lower=FALSE)
```

And agian obtain a highly significant p-value of 5.187e-06.

(b).2 Complementary Log-Log link function

```
clogmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=cloglog), orings)
clogdevtest <- pchisq(clogmod$null-clogmod$deviance,
                  clogmod$df.null-clogmod$df.residual,lower=FALSE)
```

And again obtain a highly significant p-value of 1.734e-06.

(b).3 Cauchit link function

```
cauchitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=cauchit), orings)
cauchitdevtest <- pchisq(cauchitmod$null-cauchitmod$deviance,
                  cauchitmod$df.null-cauchitmod$df.residual,lower=FALSE)
```

And again obtain a highly significant p-value of `3.174e-07`.

(b).4 Log link function

```
logmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=log), orings)
logdevtest <- pchisq(logmod$null-logmod$deviance,
                  logmod$df.null-logmod$df.residual,lower=FALSE)
```

And again obtain a highly significant p-value of `9.24e-07`.

While all of these p-values were significant, it is clear there is some amount of variation. Moreover, this may not be the case for a larger p-value or a different test run with different link functions. Quite helpfully, R restricts the use of link functions with binomial data to the five tested here. I infer this is in order to guarantee a value for p between 0 and 1.

(c) Creating a new column of response as the indicator on whether or not there is some oring damage in that launch. Refit this binary reponse to a GLM with the logit link.

With the new model, the significance of the launch temperature is much lower (but still significant) at `0.004804`.

(d) Which model do you prefer? The one in part (a) or part (c)? Why?

I prefer the model in part a, because the model in part (c) is throwing away additional data. However, one advantage of the indicator/binary model is that there is one data point at lower temperature with a high proportion of damage which may be acting as an outlier or influential point.

3 Appendix: Tangled R Code

```
library(MASS); library(xtable);library(nlme)
  lvector <- function(x, dig = 2, dsply=rep("f",ncol(x)+1)) {
   x <- xtable(x, align=rep("",ncol(x)+1),display=dsply,digits=dig) # We repeat empty string 6 time
   print(x, floating=FALSE, tabular.environment="pmatrix",
     hline.after=NULL, include.rownames=FALSE, include.colnames=FALSE)
   }

library(faraway)
data(orings)

plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim =c(0,1), xlab ="Temperature", ylab = "Prob of dam

##fit a line to the data
lmod <- lm(damage/6 ~ temp, orings)
abline(lmod)

###logit is default link choice
```

```
logitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial, orings)
summary(logitmod)


plot(damage/6 ~ temp, orings, xlim=c(25,85), ylim =c(0,1), xlab ="Temperature", ylab = "Prob of dam
x <- seq(25,85,1)
lines(x, ilogit(11.6630-0.2162*x)) #ilogit is inverse logit function

probitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=probit), orings)
summary(probitmod)
lines(x, pnorm(5.59145-0.10580*x),lty=2)

logitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial, orings)
devtest <- pchisq(logitmod$null-logitmod$deviance,
                    logitmod$df.null-logitmod$df.residual,lower=FALSE)

probitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=probit), orings)
probitdevtest <- pchisq(probitmod$null-probitmod$deviance,
                    probitmod$df.null-probitmod$df.residual,lower=FALSE)

clogmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=cloglog), orings)
clogdevtest <- pchisq(clogmod$null-clogmod$deviance,
                    clogmod$df.null-clogmod$df.residual,lower=FALSE)

cauchitmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=cauchit), orings)
cauchitdevtest <- pchisq(cauchitmod$null-cauchitmod$deviance,
                    cauchitmod$df.null-cauchitmod$df.residual,lower=FALSE)

logmod <- glm(cbind(damage, 6-damage) ~temp, family=binomial(link=log), orings)
logdevtest <- pchisq(logmod$null-logmod$deviance,
                    logmod$df.null-logmod$df.residual,lower=FALSE)

myorings <- orings
myorings[,3] <- ifelse(myorings[,2]>0,1,0)
colnames(myorings)[3] <- "Ind"
binmod <- glm(Ind ~ temp, family=binomial(link=logit), data=myorings)
summary(binmod)

bindevtest <- pchisq(binmod$null-binmod$deviance,
                    binmod$df.null-binmod$df.residual,lower=FALSE)
```