tle: “LNL_HW_week4”
thor: “Patrick Kelly”
te: “Saturday, February 07, 2015”
tput: pdf_document

Week 4 Homework Assignment

This assignment helps understanding binomial regression

The assignment is due on the day of the next class at 11:59 pm.

This assignment is individual

1. Select an area of application, for example, biomedical studies, engineering, marketing, etc.

I am interested in exploring different marketing data set for application of a binomial regression.

2. For the selected area of application postulate a problem that could be solved using binomial regression. Think of the two versions of the experiment design: with prospective and retrospective sampling. It is recommended to postulate the problem with two predictors.

I wanted to use data from within the marketing industry. I found a competition on kaggle that had click-through data, however through further investication I found that this data was a bit too complicated for this assignment and that I could not effectively condense the 22 independent variables down to only two.

As an alternative, I decided to take the same approach but simulate my own data. Therefore I created a sceario in which we are measuring whether or not a user clicked on a link and are capturing two independent varaibles: the user’s sex (M or F) and the hour of the day that the user was exposed to the link (1pm through 5pm).

Prospective experimental design:

Designate a random group of people, half male and half female, and split them into five groups randomly. Then have each

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

group browse the internet (or the target website) during a different times of the day, splitting the groups across each of the five hours (1pm - 5pm). Then record whether or not each person clicks on the link that you are intending to measure.

Retrospective experimental design:

Access the click stream database to the target website. Gather the the following historical data between 1pm and 5pm: the website visitor's sex, the time the visited the site, and whether or not they clicked on the link that you want to measure.

3. Find or simulate data for the model.

Below is the setup of the simulated data. Please note that this simulated data is over engineered and therefore produces a very strong model, but the learnings from the exercise still apply.

```
library(faraway)
library(plyr)
```

```
##
## Attaching package: 'plyr'
##
## The following object is masked from 'package:faraway':
##
##     ozone
```

```
#initialize vectors
click <- numeric()
hour <- numeric()
count1 <- 0
count2 <- 0
#creating the data
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
for (i in c(5,7,18,45,60)){
  count1 <- count1+1
  p <- i*.01
  d <- rbinom(100,1,p)
  click <- c(click,d)
  hour <- c(hour,rep(count1,100))
}
for (i in c(10,11,23,50,65)){
  count2 <- count2+1
  p <- i*.01
  d <- rbinom(100,1,p)
  click <- c(click,d)
  hour <- c(hour,rep(count2,100))
}
sex <- c(rep("M",500),rep("F",500))

in.data <- data.frame(click,hour,sex)

table.data <- ddply(in.data, .(hour, sex), summarise, clicks=sum(click==1), noclicks=sum(click==0))
table.data
```

```
##    hour sex clicks noclicks
## 1     1   F     10       90
## 2     1   M      7       93
## 3     2   F      9       91
## 4     2   M      5       95
## 5     3   F     19       81
## 6     3   M     19       81
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
## 7     4   F     52        48
## 8     4   M     45        55
## 9     5   F     58        42
## 10    5   M     59        41
```

4. Analize the use of the three different link functions and compare them to each other.

There are three commonly used links for binomial output data: - Logit: n = ln(p/(1-p)); p2 = e^n / (1+e^n) - Probit: n = (phi)^-1(p); p2 = phi(n) - Complementary log log: n = ln(-ln(1-p)); p2 = 1 - $e^{(-e}$n)

As we know from lecture, the three links appear very similar for probabilities around 0.5. They show signiÖcant relative di§erences in cases of extreme probabilities (p t 0, 1). But distinguishing the link functions in the tail areas is not easy: need very long samples in which very rare events may be observed.

The logit is the most common transformation. Since probabilities fall between 0 and 1, a transformation is necessary so that we do not have to put considerable constraints on our coeffificents and independent variables. Hence the initial odds transformation which changes the range to 0 to infinity. And then finally the log odds transforms the range to negative infinity to infinity, allowing for the model to takes its most flexible and fitting form.

The probit function is necessary to transform the data using the same logic as above, but probit models are more tractable in some situations than logit models (e.g. in a Bayesian setting in which normally distributed prior distributions are placed on the parameters)

Lastly, the complementary log-log function log(???log(1???p)) may also be used. This link function is asymmetric and will often produce different results from the probit and logit link functions but still accomplishes the necessary transformation / range adjustment. The complementary log log can be used over the other link functions above certain circumstances, such as where an underlying random variable is reduced to a dichotomous form.

5. Conduct the fit of binomial regression model using glm() with family=binomial.

```
#creating the 3 models that differ by link fn
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```r
#logit
modl<-glm(formula = cbind(clicks,noclicks) ~ hour + sex,family=binomial,data=table.data)
#probit
modp<-glm(formula = cbind(clicks,noclicks) ~ hour + sex,family=binomial(link=probit),data=table.data)
#complementary log log
modc<-glm(formula = cbind(clicks,noclicks) ~ hour + sex,family=binomial(link=cloglog),data=table.data)

modl
```

```
##
## Call:  glm(formula = cbind(clicks, noclicks) ~ hour + sex, family = binomial,
##     data = table.data)
##
## Coefficients:
## (Intercept)         hour         sexM
##     -3.6555       0.8397      -0.1616
##
## Degrees of Freedom: 9 Total (i.e. Null);  7 Residual
## Null Deviance:       233.6
## Residual Deviance: 17.12     AIC: 67.64
```

```r
modp
```

```
##
## Call:  glm(formula = cbind(clicks, noclicks) ~ hour + sex, family = binomial(link = probit),
##     data = table.data)
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
##
## Coefficients:
## (Intercept)         hour         sexM
##     -2.0984       0.4805      -0.1008
##
## Degrees of Freedom: 9 Total (i.e. Null);  7 Residual
## Null Deviance:       233.6
## Residual Deviance: 19.26     AIC: 69.77
```

```
modc
```

```
##
## Call:  glm(formula = cbind(clicks, noclicks) ~ hour + sex, family = binomial(link = cloglog),
##     data = table.data)
##
## Coefficients:
## (Intercept)         hour         sexM
##     -3.3954       0.6853      -0.1078
##
## Degrees of Freedom: 9 Total (i.e. Null);  7 Residual
## Null Deviance:       233.6
## Residual Deviance: 18.77     AIC: 69.28
```

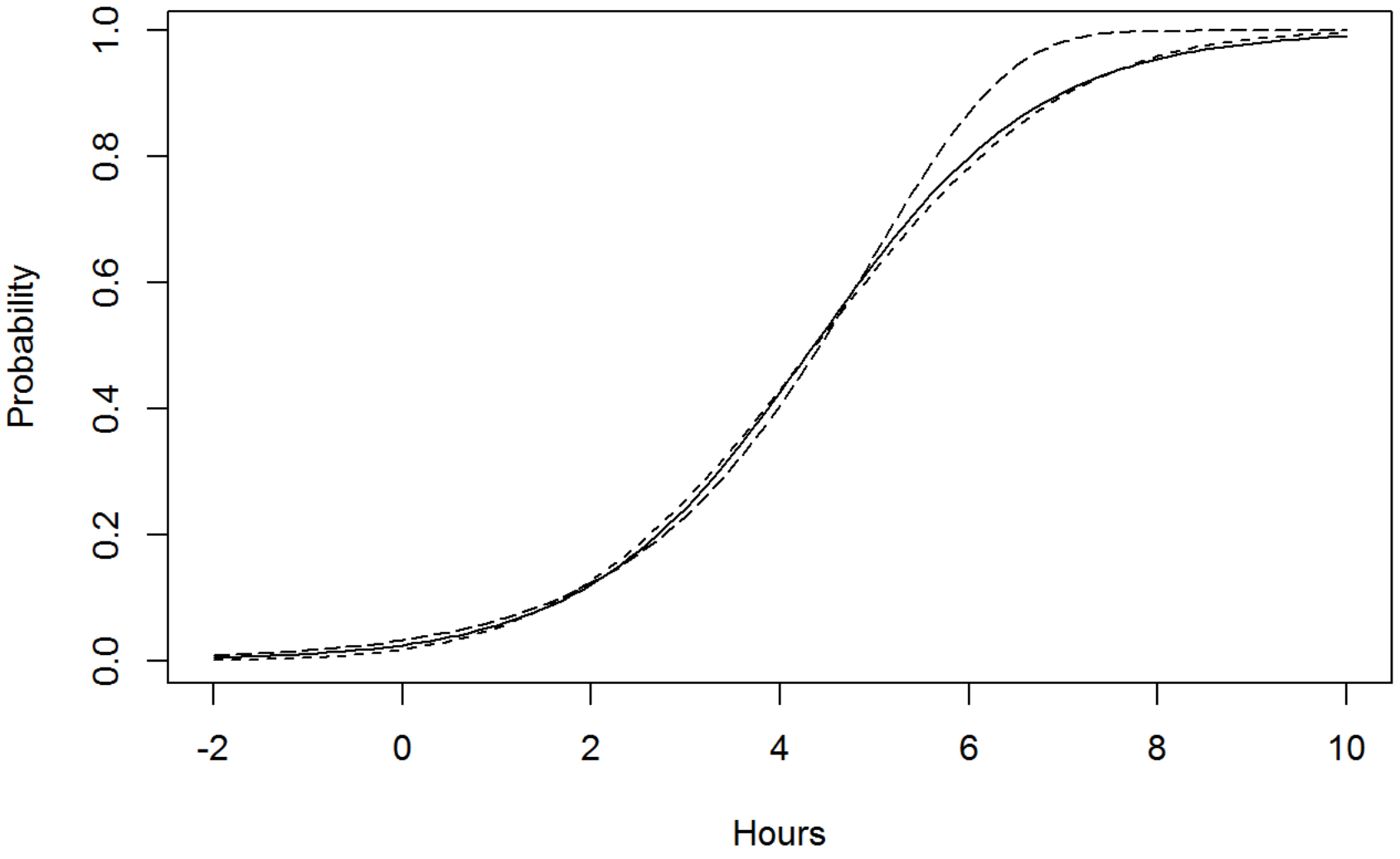
```
fit.check <- cbind(fitted(modl),fitted(modp),fitted(modc))
summary(fit.check)
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

```
##        V1                V2                V3
##  Min.   :0.04846   Min.   :0.04285   Min.   :0.05798
##  1st Qu.:0.10954   1st Qu.:0.11282   1st Qu.:0.11474
##  Median :0.22873   Median :0.24002   Median :0.22002
##  Mean   :0.28300   Mean   :0.28325   Mean   :0.28236
##  3rd Qu.:0.41660   3rd Qu.:0.42028   3rd Qu.:0.39727
##  Max.   :0.63249   Max.   :0.61957   Max.   :0.64355
```

```
x<-seq(-2,10,.1)
pl<-ilogit(modl$coef[1]+modl$coef[2]*x)
pp<-pnorm(modp$coef[1]+modp$coef[2]*x)
pc<-1-exp(-exp((modc$coef[1]+modc$coef[2]*x)))
plot(x,pl,type="l",ylab="Probability",xlab="Hours")
lines(x,pp,lty=2)
lines(x,pc,lty=5)
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com



6. Describe the problem, all steps of obtaining solution and conclusions about the results.

The problem was that we want to understand the impact on sex and time of day on whether or not a person visiting our website will click on a certain link (ex: a link that leads to the “products for sale” page).

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com

In order to obtain our solution, I first collected the date from the server that tracks click stream activity for my website and limited my data to my two independent variables (viewer's sex and the time of the day that the view visited the page) and my dependent variable (binary variable indicating whether or not the user clicked on the target link). I ensured that I collected a sample of 100 visitors for each combination of sex and hour. (THIS DATA WAS SIMULATED ABOVE) This is a retrospective experimental design. I then created a table of counts of clicks for each of my combinations of sex and hour, 10 total rows, and created a binomial regression with the following link funtions: logit, probit, and complementary log log. I determined that the logit link function was the most appropriate model due to its fit to the data, however the learning from each of the models was the same (as the results were very simliar). In conclusion, using the coeffients from the logit model, we can conclude that an additional hour increases the odds of a click by ~250% and if the view is male then this reduces the odds of the clink by ~64% (see below for exp() caluclations). Therefore we should try to drive traffic on our website later in the day and focus on a female audience for a higher conversion rate on this link.

```
exp(0.9252)
```

```
## [1] 2.522373
```

```
exp(-0.4508)
```

```
## [1] 0.6371183
```

PRO version Are you a developer? Try out the HTML to PDF API pdfcrowd.com