DEPARTMENT OF STATISTICS

Course STATS 330: Advanced Statistical Modelling

Tutorial Sheet 9: October 9, 2008

This tutorial is designed to give you practice in fitting a Poisson regression model.

In this tutorial you will investigate a famous problem of disputed authorship. Many scholars have questioned whether St Paul actually wrote the fourteen epistles of the New Testament. The table below gives the number of occurrences of the Greek word "kai" (meaning "and") in sentences from 5 of the Epistles, namely Romans, 1st Corinthians, 2nd Corinthians, Galatians, and Philippians. Many scholars argue that the frequency of common words such as conjunctions that are independent of the subject matter are good indicators of authorship: writing by the same author should have similar frequencies.

	Work				
# of sentences with	**Romans**	**1st Corinthians**	**2nd Corinthians**	**Galatians**	Philippians
0"kais"	**386**	**424**	**192**	**128**	42
1 "kais"	**141**	**152**	**86**	**48**	29
2 "kais"	**34**	**35**	**28**	**5**	19
3 "kais"	**15**	**14**	**11**	**5**	10
4 "kais"	2	2	2	1	2

The data has been put into a file called **`St.Paul.txt`** which can accessed from the webpage. This data set contains 3 variables: **`work, number,`** and **`weight.`** These are

`work` having the values I, II, III, IV and V indicating the epistle.
`number` the number of times "kai" occurs in the sentence.
`weight` the number of sentences having the indicated value of **number**.

We will use Poisson regression, using **`number`** as the response and **`work`** as the explanatory to investigate the authorship of the first 5 epistles. You need to 1) fit a model that explains the connection between the mean number of "kai"s per sentence and the epistle, and 2) interpret the results.

Task 1: Read in the data and create a suitable data frame for the analysis

This should be routine by now. However, if the file had not been provided (as in Assignment 5), you could create a data frame using the function **`expand.grid`** to make the factors. You type

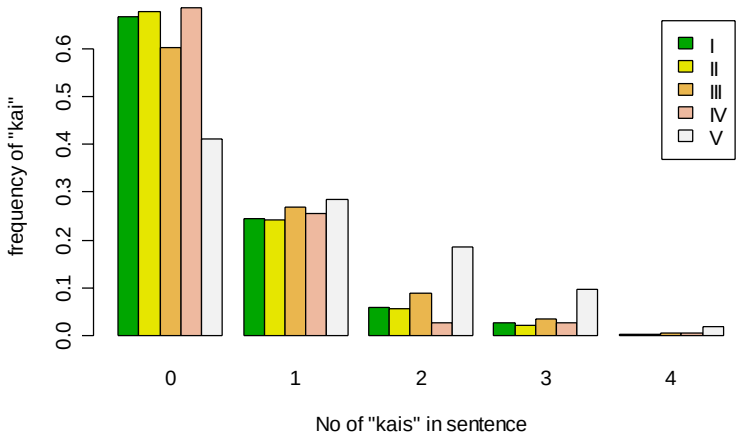
```
counts<- c(386,141,34,15,2,424,152,35,14,2,192,86,28,11,2,128,48,5,5,1,
42,29,19,10,2)

St.Paul.df<-data.frame(expand.grid(number= 0:4,
work=c("I","II","III","IV","V")),
weight=counts)
```

Task 2: Plot the data

A good plot is to plot the relative frequencies of the sentences having 0 ,1,2, … occurrences of "kai" and compare the frequencies across the works. Try the function **`barplot`**. Look up the explanation of this function by typing **`?barplot`** in R. Then try the following code (you will need to consult the R documentation to make sense of it):

```
# barplot needs a matrix of heights

freqs<-matrix(0,5,5)
# columns correspond to the number of "kais", 0,1,2,3,4
# rows correspond the works
# elements are frequencies

# use tapply to get work totals

totals<-tapply(St.Paul.df$weight, St.Paul.df$work, sum)

index<-1:5
for(i in 1:5){
freqs[i,]<-St.Paul.df[index,3]/totals[i]
index<-index + 5
}

# label rows and columns of matrix, for labelling barplot
dimnames(freqs)<-list(c("I","II","III","IV","V"),0:4)

xlabels<-0:4    # label for number of kais

bar.colours<-terrain.colors(5)  # colours for bars (5 in all, one for each
work)

barplot(height=freqs,
names.arg=xlabels,
legend.text=TRUE,
beside=TRUE,
col=bar.colours,
xlab="No of \"kais\" in sentence",
ylab="frequency of \"kai\" ")
```

This produces



Task 3: Interpret the plot

Do all the 5 works seem similar in terms of the distribution of the number of “kais” in a sentence? If not, which ones are different?

Task 4: Fit the model

Next we examine the differences, if any, between the means of these 5 distributions (works). We fit a Poisson regression model using **`number`** as response and **`work`** as a (categorical) explanatory.)

The data are in grouped form in the data set supplied. We get round this by using the variable **`weight`** to indicate a repeat count. See the onion example in Lecture 26.To fit a mean to each work, we type

```
St.Paul.glm<-glm(number~work, family=poisson, weight=weight, data=St.Paul.df)
summary(St.Paul.glm)
Call:
glm(formula = number ~ work, family = poisson, data = St.Paul.df,
    weights = weight)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.79123    0.06177 -12.809  < 2e-16 ***
workII      -0.04392    0.08655  -0.507   0.6119
workIII      0.23552    0.09633   2.445   0.0145 *
workIV      -0.09607    0.12961  -0.741   0.4585
workV        0.82022    0.11550   7.102 1.23e-12 ***

Null deviance: 2059.1  on 24  degrees of freedom
Residual deviance: 2000.7  on 20  degrees of freedom
AIC: 3432.9

Number of Fisher Scoring iterations: 5
```

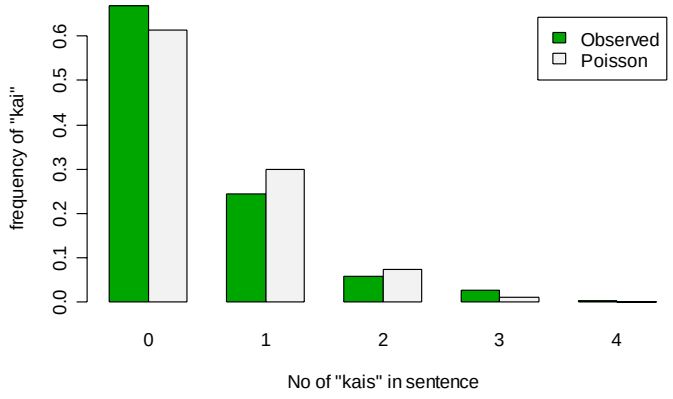
Task 5. Interpret the coefficients

Looking at the parameters, we see that work V is clearly different from the baseline (work I). Also, it seems that the mean of work III is also significantly different from the baseline, so that the authorship of works III and V seems to be different from those of works I, II, and IV.

Task 6: check the model

The main assumption here is that the number of "kais" per sentence has a Poisson distribution. We can check this by comparing the observed frequencies to the corresponding probabilities given by the Poisson distribution with mean equal to the observed mean. Let's plot the observed frequencies against the corresponding Poisson probabilities, separately for each work. Use the code (for e.g. work I)

```
heights<-rbind(freqs[1,],dpois(0:4,282/578)) # work I mean is 282/578

# label rows and columns of matrix, for labelling barplot
dimnames(heights)<-list(c("Observed", "Poisson"),0:4)

xlabels<-0:4    # label for number of kais

bar.colours<-terrain.colors(2)  # colours for bars (2 in all, one observed
and one for Poisson)

barplot(height=heights,
names.arg=xlabels,
legend.text=TRUE,
beside=TRUE,
col=bar.colours,
xlab="No of \"kais\" in sentence",
ylab="frequency of \"kai\" ")
```



Agreement seems reasonable, could check with a chi-squared test. What about the other works?