ST2304 Exercises Week 12: Log-linear Models

Bob O'Hara

14 March 2018

Log-linear Models in R

We can first walk through the analysis of the Hastings rarities data. First we need to read in the data. Note that I have used `stringsAsFactors = FALSE`, so strings aren't converted to factors. I want to do this explicitly later, so I can set the contrasts in a nice way.

```
Hastings <- read.csv("../Data/HastingsData.csv")

# Make sure baselines are correct
Hastings$Era <- relevel(Hastings$Era, ref = "B")
Hastings$Area <- relevel(Hastings$Area, ref = "Sussex")
# for plots
Hastings$Colour <- c("red2", "blue", "red4")[as.numeric(Hastings$Area)]

str(Hastings)
```

```
'data.frame':   540 obs. of  7 variables:
 $ X     : int  1 2 3 4 5 6 7 8 9 10 ...
 $ Year  : int  1895 1896 1897 1898 1899 1900 1901 1902 1903 1904 ...
 $ Era   : Factor w/ 2 levels "B","A": 2 2 2 2 2 2 2 2 2 2 ...
 $ Class : Factor w/ 3 levels "I","II","III": 1 1 1 1 1 1 1 1 1 1 ...
 $ Area  : Factor w/ 3 levels "Sussex","Hastings",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Count : int  2 4 2 0 1 4 7 7 10 7 ...
 $ Colour: chr  "blue" "blue" "blue" "blue" ...
```

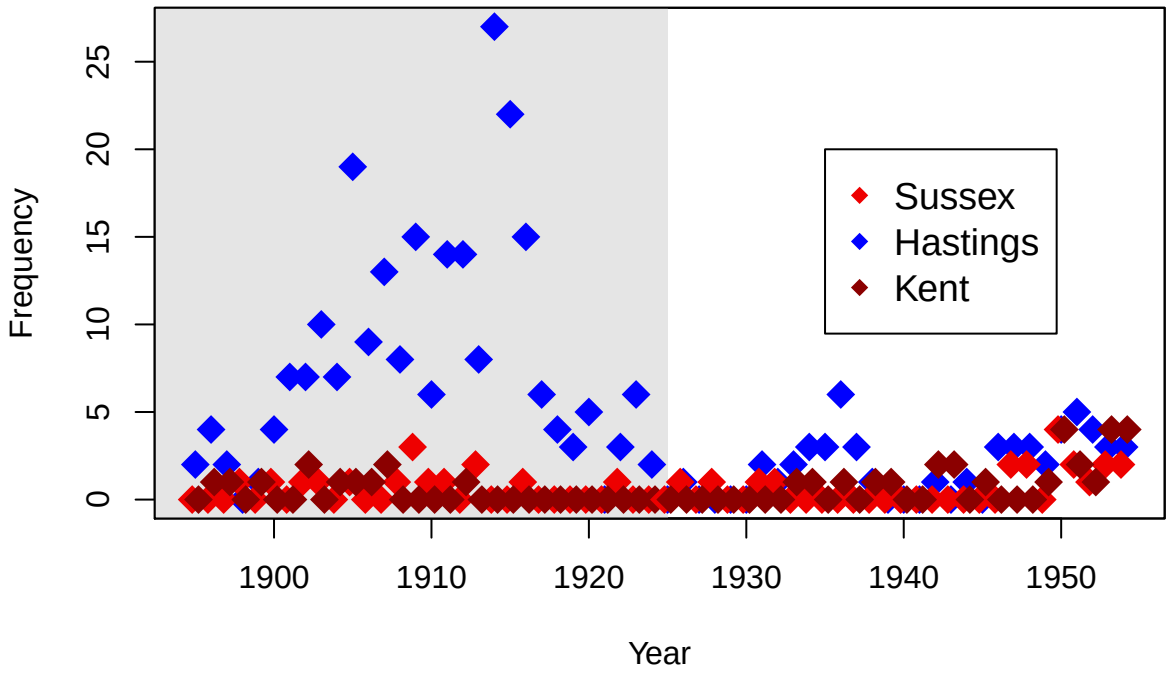
- Year is the year
- Era is whether it is the years under suspicion (A, i.e. before 1925) or afterwards (B)
- Class is the class of rarity, derived from an independent source. I is the rarest, III is the least rare
- Area is the geographic area: Hastings (and surrounding area), the rest of Sussex, and Kent
- Count is the number of observations of rare species in that class.

I set the baselines of the factors to being Era B (i.e. when there wasn't any fraud suspected), and the baseline area to Sussex (where no fraud was suspected). I also create a variable for colour, which is nice when plotting.

We only want to look at Class I, so here we extract it and then plot the data:

```
# Get class 1
HastingsClassI <- Hastings[Hastings$Class=="I",]

plot(HastingsClassI$Year + 0.2*(as.numeric(HastingsClassI$Area)-2),
     HastingsClassI$Count, type="n", xlab="Year", ylab="Frequency")
rect(1850, -5, 1925, 59, col="grey90", border=NA)
points(HastingsClassI$Year + 0.2*(as.numeric(HastingsClassI$Area)-2),
     HastingsClassI$Count, col = HastingsClassI$Colour, pch=18, cex=2)
legend(1935, 20, levels(HastingsClassI$Area), col=c("red2", "blue", "red4"), pch=18, cex=1.2)
box(bty="o")
```



Now we can fit the model. The null model is that the number of rarities will depend on the era (if hte number of birders is different before and afterwards) and the region: the regions have different sizes, and rare birds may appear on different parts of the coast. The alternative model is that there in an interaction: for some reason there might be extra variation beyond this. In particular, one area (Hastings) might have more rarities recorded in Era A than would be expected if only time and space had an effect. This, then is the interaction.

So, we can fit the model:

```
Hast.mod <- glm(Count ~ Area*Era, family=poisson,
                data=HastingsClassI)
anova(Hast.mod, test = "Chisq")
```

```
Analysis of Deviance Table

Model: poisson, link: log

Response: Count

Terms added sequentially (first to last)

         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                       179     848.02
Area      2   349.31       177     498.71 < 2.2e-16 ***
Era       1    81.60       176     417.11 < 2.2e-16 ***
Area:Era  2    55.23       174     361.89 1.018e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The R code is almost the same as for a linear model, the only difference is that we have the extra argument, `family="poisson"`, which says that we should assume a Poisson distribution (if we forget this, R will assume a normal distribution, and the model will be fitted just like lm()). If we want to specify the link function, we can write `family=poisson(log)`.

`anova()` gives an Analysis of Deviance table. The Area:Era effect is the one we are most intersted in: the deviance is 55.23 with 2 degrees of freedom, so the p-value for the test is 10^{-12} (we should correct for

over-dispersion, but this will not make much difference).

So that suggests that something is going on. But is it related to Hastings? We can look at the parameter estimates:

Parameter Estimates

```
round(summary(Hast.mod)$coefficients, 2)
```

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.46       0.23   -1.99     0.05
AreaHastings          1.04       0.27    3.92     0.00
AreaKent              0.31       0.30    1.04     0.30
EraA                 -0.24       0.35   -0.68     0.49
AreaHastings:EraA     1.74       0.38    4.62     0.00
AreaKent:EraA        -0.62       0.50   -1.25     0.21
```

The main result is that before 1925 there were about $\exp(1.74) = 5.7$ times more rarities recorded than would have been expected (the 95% Confidence interval is 2.73, 12.1, so we can be sure that there is a large effect). Note that this is the only big effect: there is little difference in the interaction between Kent and Sussex (although the standard errors are fairly large, so we won't see any small effects).

We can test for over-dispersion:

```
ResidDev <- deviance(Hast.mod)
ResidDF <- df.residual(Hast.mod)

# Need to test the upper tail, i.e. a big test statistic is significant
pchisq(ResidDev, df=ResidDF, lower.tail = FALSE)
```

```
[1] 2.868446e-15
```

And the estimate of the overdispersion is `ResidDev/ResidDF` $= 2.08$, which is fairly small, but still worth correcting for: the effect is to increast the standrd errors by 1.44 times.

```
round(summary(Hast.mod, dispersion = ResidDev/ResidDF)$coefficients, 2)
```

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.46       0.33   -1.38     0.17
AreaHastings          1.04       0.38    2.72     0.01
AreaKent              0.31       0.44    0.72     0.47
EraA                 -0.24       0.50   -0.47     0.64
AreaHastings:EraA     1.74       0.54    3.20     0.00
AreaKent:EraA        -0.62       0.72   -0.87     0.39
```

Residuals

it is straightforward to plot the deviance residuals. The plot of the residuals agaisnt the fitted values isn't terribly informative, because ther are only 6 levels. But the plot against time is helpful, as it suggests an increase at the end of the period, suggesting we could add Year as an effect if we want to model the data well. But this would use a lot of parameters, and for this problem would not change the conclusions

```
plot(HastingsClassI$Year, resid(Hast.mod), col=HastingsClassI$Colour)
legend(1925, 4.6, levels(HastingsClassI$Area), col=c("red2", "red4", "blue"), pch=1, cex=1.2)
```

We could also fit a negative binomial distribution. This can be done a few ways, here we will use the `glm.nb()` function in the MASS package:

```
library(MASS) # this should come with R as standard
Hast.NB <- glm.nb(Count ~ Area*Era, data=HastingsClassI)

round(summary(Hast.NB)$coefficients, 2)
```

```
                Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.46       0.27   -1.70     0.09
AreaHastings        1.04       0.33    3.15     0.00
AreaKent            0.31       0.36    0.87     0.38
EraA               -0.24       0.40   -0.59     0.55
AreaHastings:EraA   1.74       0.47    3.72     0.00
AreaKent:EraA      -0.62       0.57   -1.09     0.27
```

We see the same result as before: the point estimates are the sme, but the standard errors are a bit different. The differences are because the assumptions about how the mean and variance are related are different.

Question 1

The short version: repeat the analysis above for rarity classes II and III. This should mainly be copying and pasting the code above. To extract the data do this (download the data from Blackboard: you will probably have to change the file path):

```
Hastings <- read.csv("../Data/HastingsData.csv")

# Make sure baselines are correct
Hastings$Era <- relevel(Hastings$Era, ref = "B")
Hastings$Area <- relevel(Hastings$Area, ref = "Sussex")
# for plots
Hastings$Colour <- c("red2", "blue", "red4")[as.numeric(Hastings$Area)]
```

```
# Make data frames for classes II and III
# Get class 1I
HastingsClassII <- Hastings[Hastings$Class=="II",]
HastingsClassIII <- Hastings[Hastings$Class=="III",]
```

1. For Class II (the intermediate rarities), fit the same model as I did for Class I. Is Hastings any different between eras A and B? If so, what is the difference?
2. Is there any evidence for overdispersion? If there is, what effect does correcting for it have on the model?
3. Plot the residuals against year. Does there seem to be any extra effect of year?

Now for Class III:

4. For Class III (the more common rarities), fit the same model as I did for Class I. Is Hastings any different between eras A and B? If so, what is the difference?
5. Is there any evidence for overdispersion? If there is, what effect does correcting for it have on the model?
6. Plot the residuals against year. Does there seem to be any extra effect of year?
7. Is there any difference between the estimated coefficients for the AreaHastings:EraA interactions between the three different classes?

Exam-type Questions

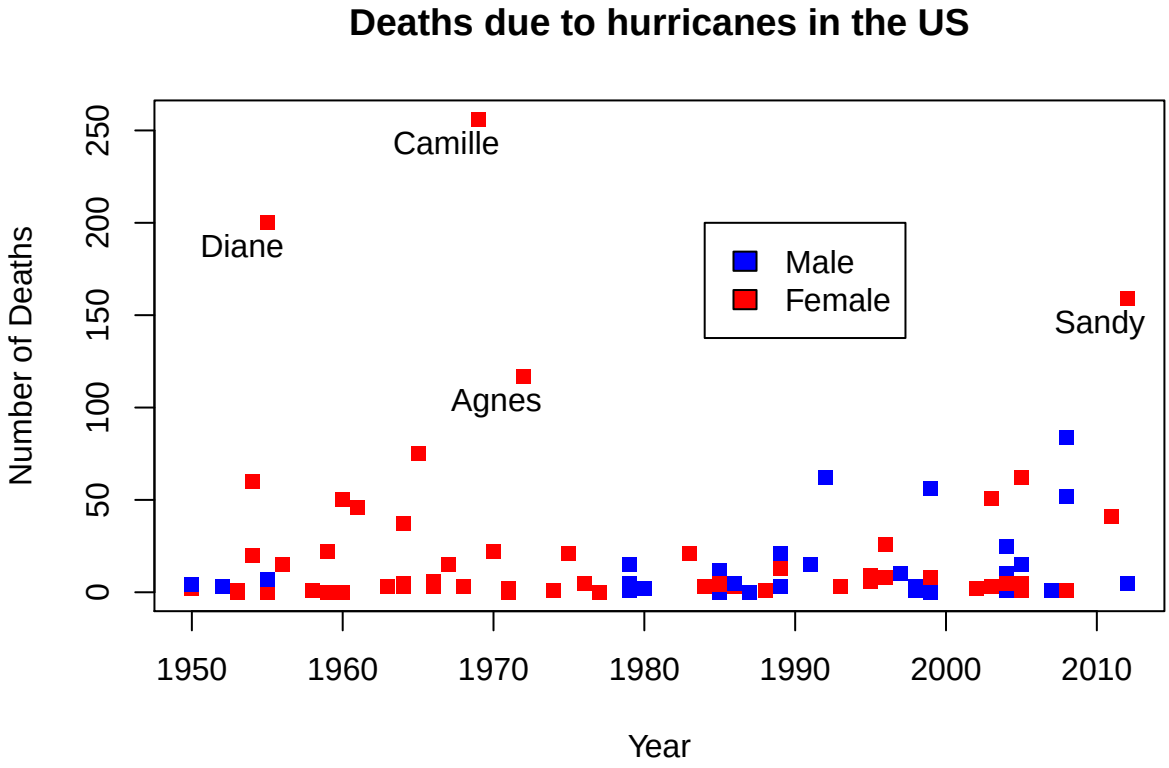
1. From the output of the analysis of Class II, does the Analysis of Deviance table suggest there is variation between areas and eras?
2. What statistics would you use to investigate if there is over-dispersion, and how strong it is? What additional information would you need?
3. From the summary, for Class II, what are the main differences between eras and areas?
4. The data were collected to see if there was evidence of fraud, which we would see if there were more rarities in Era A in the Hastings area. Does the analysis from Class II suggest that there could be fraud?

Question 2: Himmicanes

A few years a go a strange paper appeared in PNAS that suggested that hurricanes in the USA with female names caused more deaths than those with male names. This was viewed with some scepticism. Here we can find out why.

```
Data <- read.csv("../Data/Himmicanes.csv", stringsAsFactors = FALSE)
BigH <- which(Data$alldeaths>100) # Select hurricanes with > 100 deaths

plot(Data$Year, Data$alldeaths, col=Data$ColourMF, type="p", pch=15,
     xlab="Year", ylab="Number of Deaths", main="Deaths due to hurricanes in the US")
text(Data$Year[BigH], Data$alldeaths[BigH], Data$Name[BigH], adj=c(0.8,1.5))
legend(1984, 200, c("Male", "Female"), fill=c("blue", "red"))
```



The relevant variables are:

- Year: Year
- Name: Hurricane's name
- MasFem: A scoring of how feminine the name sounds (we won't use this here)
- Minpressure: minimum air pressure in the hurricane (a measure of stength)
- Gender: Gender (0: Male, 1: Female)
- Category: Category of hurricane (larger is more severe)
- alldeaths: Number of deaths
- NDAM: Normalised damage (i.e. how much the hurricane cost, corrected for inflation etc.)

The aim is to predict the number of deaths. This is a count, so a Poisson distribution makes sense (but it is very overdispersed). After some poking at the data, the authors arrived at a model with Minpressure and NDAM, and the interaction of each of these with Gender.

1. How would you write the model? (if it's easier, write the model as the interaction of Gender with Minpressure, plus the interaction of Gender and NDAM).
2. Fit the model assuming a Poisson distribution. Does it suggest an effect of gender?
3. Is there any evidence for over-dispersion?
4. Correct the overdispersion using a negative binomial distribution (this is what was done in the paper), using the glm.nb() function in the MASS package
5. Now look at the residuals: plot them against the predicted values and the covariates. Does the model look OK, or can it be improved?
6. Based on this, try to find a model that fits better. Does this model show an effect of gender?