Department of Statistics
STATS 330: Advanced Statistical Modelling
Assignment 1
Semester 2, 2019

Total: 60 marks Due: 12:00 (noon), 16 August, 2019

Notes:

(i) Write your assignment using R Markdown. Knit your report to either a Word or PDF document.

(ii) Create a section for each question. Include all relevant code and output in the final document.

(iii) Marks may be deducted for poor style. Please keep your code and plots neat.

(iv) Please remember to hand in your hard copy, with signed cover sheet, by the due date.

Introduction

The magazine Mother Jones has recorded the number of mass-killings in the USA since 1982. Click [here] to visit this website:

This, from their website:

"Our research focused on indiscriminate rampages in public places resulting in four or more victims killed by the attacker. We exclude shootings stemming from more conventionally motivated crimes such as armed robbery or gang violence. Other news outlets and researchers have since published larger tallies that include a wide range of gun crimes in which four or more people have been either wounded or killed. While those larger data sets of multiple-victim shootings are useful for studying the broader problem of gun violence, our investigation provides an in-depth look at a distinct phenomenon-from the firearms used and mental health factors to the growing copycat problem. Tracking mass shootings is complex; we believe ours is the most useful approach."

In this assignment we will examine modelling from a number of perspectives — all of which should be considered when deciding how to model data:

- Inspecting the data and making allowances for other background events pertinent to these data,
- looking at how these models fit your data,
- see how well the model predicts future events,
- using all the above ideas to see which model you are most 'comfortable' with?

The data

The data set `masskill.csv` contains the following variables:

- `year`: the year the incidents occurred,
- `popn`: the population of the USA for a given year (millions),
- `masskill`: the number incidents where a mass killing occurred.

(a) Plot the number of mass killing incidents per year over this period of time. Comment briefly.

[5 marks]

(b) Plot the population of the USA over this period of time. Comment briefly on how the population is changing over this period of time.

[5 marks]

(c) Make a plot that shows the number of mass killing per year, per 100 million people. Comment briefly.

[5 marks]

Some background information

The variable population (per 100 million) is an example of an exposure variable. As it would be natural to assume that if the population increased then it would there may be an increase in the total number of mass killings (all other things being equal). That is, we have increased the exposure to this event occuring.

Mathematically, we adjust for this greater exposure as follows :

$\log(E(\text{count})/\text{popn}) = log(E(\text{count})) - \log(\text{popn}) = \mathbf{x}^T\boldsymbol{\beta}$

where $\mathbf{x}^T\boldsymbol{\beta}$ is short-hand for $\beta_0 + \beta_1 x_{1i} + \cdots \beta_p x_{pi}$ i.e. the variables you are using to explain these counts.

Alternatively, this this cane be expressed as: $\log(E(\text{count})) = \mathbf{x}^T\boldsymbol{\beta} + \log(\text{popn})$

Here the term log(popn) is called an offset term as it offsets the effect of this exposure variable.

(d) Fit a 'linear model' in this model for the mass killing count, starting from 1982 as year 0, with the offset population exposure variable as follows:

```
glm(formula = masskill
~ I(year - 1982), family = "poisson",
offset = log(popn/100)),data = MK.df }
```

Comment, briefly, on what you conclude from this output.

[5 marks]

(e) State, mathematically, the model for the count of mass killings that you are fitting here.

[5 marks]

(f) Include an additional quadratic term in a new model model in this model for the mass killing count, starting from 1982 as year 0, with the offset population exposure variable as follows:

```
glm(formula = masskill
~ I(year - 1982)+ I((year - 1982)^2),
family = "poisson",
offset = log(popn/100)),data = MK.df }
```

Comment, briefly, on what you conclude from this output.

[5 marks]

(g) State, mathematically, the model for the count of mass killings that you are fitting here.

[5 marks]

(h) Plot the data again with the linear model's expected counts superimposed along with the 95% confidence interval band for these expected values. Comment, briefly, on how well your model fits these data.

[5 marks]

Hint: the `R` function `predict` default setting will give you predicted values of the "link" type which in this case will give you the log of the expected/fitted count. Specifying `se.fit=TRUE` will give its associated standard error. You can add fitted lines to plots using the `R` function `lines`.

(i) Plot the data again with the quadratic model's expected counts superimposed along with the 95% confidence interval band for these expected values. Comment, briefly, on how well your model fits these data.

[5 marks]

(j) Compute a confidence interval for the mean number of mass killing in 2019 using the linear and quadratic models. Assume that the population of the USA is 327,170,000 people. Comment briefly.

[5 marks]

(k) One of the features of the Poisson distribution is that if the interval of interest changes so to does the rate value change (all other things being equal). So a rate value per year can be halved to obtain a half yearly rate value.

In the first half year of 2019 (January to June) there have been four mass killing incidents.

Use the linear and quadratic models predicted expected value and the code (changed by you), below, for the number of mass killings (adjusted for this half year scale) to see which of your models seems most appropriate for these data.

Comment, briefly.

[5 marks]

```
# distribution of Poisson(lambda=mean) distribution
pred.mean=5.5 # change this
barplot(dpois(0:15,pred.mean),ylab="Probability",xlab="x",
space=1, names.arg=0:15,
main=paste("Distribution of Poisson with mean = ",
           round(pred.mean,2)," mass killings per year")
)
```

(l) As a consequence of the above analyses which of these two models do you believe is the best description for these data. Comment, briefly, on what model you prefer and why?

[10 marks]

Hint: at this stage there is no correct answer (only time will, literally, tell) — we are interested in what your opinion is and why?