MATH302/4082: Poisson GLMs in R

Work on this by using the lung cancer example for Poisson GLMs in `R` covered in the lecture. Many explanations follow along identical lines; I have hence not given detailed explanations of the output in this exercise and have focussed more on the deviance and residuals.

The dataset analysed here was used in a study conducted by Sir Richard Doll and colleagues which explored the effects of smoking on coronary heart disease. In 1951 all British doctors were sent a questionaire about whether they smoked tobacco. The data may be read into R and printed on the screen as follows:

```
age=rep(c(40,50,60,70,80),2)
smoke=c(rep(1,5),rep(2,5))
deaths=c(32,104,206,186,102,2,2,28,28,31)
years=c(52407, 43248,28612,12663,5317,18790,10673,5710,2585,1462)
smokedata=cbind(age,smoke,deaths,years)
smokedata
```

The following points should be noted:

- the ages 40, 50,...,80 correspond to age ranges 35–44, 45–54,...,75–84, respectively;
- the variable `smoke` is a factor, where `smoke`=1 refers to smokers and `smoke`=2 refers to non-smokers;
- the variable `deaths` refers to the number of deaths in the relevant age range and smoking group (so, for example, 32 refers to the number of doctors who died in the age range 35-44 and were smokers);
- the variable `years` is the number of person-years observed in each age range and each smoking category.

  Person-years takes into account both the number of people in the study and the amount of time each person spends in the study. This is done since length of exposure (in this case, number of years of smoking) is different for different subjects, and this needs to be accounted for when modelling rates. For example, roughly, a study that followed 1000 people for 1 year would contain 1000 person years of data. A study that followed 100 people for 10 years would also contain 1000 person years of data

The main *questions of interest* in this study are the following.

Q1 Is the death **rate** higher for smokers than non-smokers?

Q2 If so, by how much?

Q3 Does the differential effect due to smoking change with age?

We can think of the response variable as being `deaths`. Since this is a count variable, Poisson regression seems a natural starting point here. Unless there is some specific reason to do otherwise, it usually makes sense to use a log link with the Poisson model. We shall do so here. The log link is the default link for the Poisson model so we do not need to declare it explicitly in R. The variable `years` needs to be used as an offset in the model. Recall how we did this in the lecture notes.

1. Under the log link function with `years` as the offset, the model supposes that log of the mean number of deaths is a linear function of log of `years` (the offset) and `age`. Let us convert `age` also to the log scale—this has nothing to do with the link function but is just a transformation of an existing predictor. We need to remember to declare `smoke` as a factor; we shall call this factor `smokef`.

   ```
   smokef=factor(smoke)
   Lage=log(age)
   Lyears=log(years)
   ```

Plot `log(deaths)` against `log(age)` and see if the plot appears more linear and than with just `age`. This is graphical check to see if the Poisson GLM seems reasonable.

2. We first consider the model

   M1: deaths ~smokef + Lage

   with `Lyears` declared as an offset, and fit it using the following commands:

   ```
   out=glm(deaths ~ smokef + Lage, family=poisson, offset=Lyears)
   summary(out)
   ```

   Match the mathematical form on the model in the notes to what is fitted here interpret the output in `summary(out)`. What responses can you offer to questions Q1-Q3?

3. Now examine the fit of this model. To do this, extract the fitted values and the residuals. R distinguishes between Pearson and Deviance (see lecture notes). Obtain each separately.

   ```
   fv=out$fitted.values
   p.res=resid(out,'pear')
   d.res=resid(out,'dev')
   ```

   The type of residuals in `resid` is in single quotes. Copying and pasting into R may cause errors.

   Plot the fitted values and the two types of residuals against the variable `deaths`. Recall that in the linear model a plot of fitted values versus the residuals was used to check for violations from assumption of nonlinearity of the mean (expected value of response) as a function of the predictors, and assumption of equal error variances - a random scatter is good, and indicates that the assumptions are not violated. The same is being done here (the model is linear once the link function is used on the expected response) with appropriately defined residuals.

4. Recall that $a * b$ in the formula for `lm` or `glm` includes the individual predictors `a` and `b` along with the interaction `a*b`. Now consider the following models, all of which should include `Lyears` as an offset:

   M2: deaths ~smokef * Lage

   M3: deaths ~smokef + Lage + Lage2
   where `Lage2` is a new variable defined by `Lage2 = Lage*Lage`; that is, we have created a new predictor that is the square of the predictor `Lage` to check if that explains variability in the response better (lower residual deviance). Note that `Lage2` is NOT linearly related to `Lage`, and hence may have complementary information to `Lage` (since taking the square is nonlinear).

   M4: deaths ~smokef * Lage + Lage2

   (i) Note the number of regression coefficients in each model, and interpret them.

   (ii) Note the nesting structure amongst the models: $M1 \subset M2$, $M1 \subset M3$, $M1 \subset M4$; $M2 \subset M4$; $M3 \subset M4$. Here $A \subset B$ denotes that model A is nested within model B.

   (iii) Fit models M2, M3 and M4, performing residual and fitted value plots in each case.

   (iv) For each model M1-M4, note down the questions amongst questions Q1-Q3 in the *questions of interest* that can be answered.

   Which of the models considered do you think is best and why? It should be clear by now that when modelling data, there is no one correct answer. Recall that smaller the deviance, better the model. But also bear in mind that deviance is always lower for a more complex model, i.e., a model with more predictors and interaction terms. So there is a need to balance smaller deviance against complexity of the model, especially since the results need to be interpretable (imagine explaining to the Oncologist about the effect of squaring the log of age!).