Poisson regression models for count data

So today I'm going to talk about Poisson regression models for count data.

I will first of all give a brief review of regression analysis. I will then introduce Poisson regression and looking at a simple model without a covariate first of all the so called a equiprobable model. I will be then assessing this particular model with the Pearson chi-squared test and the log likelihood ratio test statistics and also I will be looking at some residual analysis as well and then I will be introducing the Poisson regression model with a co-variate so basically a Poisson time trend model.

You may have come across different types of regression models already for example a linear regression model or a continuous dependent variable. You may have used logistic regression models already for a binary outcome variable. There are obviously other types of regression as well that are also part of the generalized linear models. So basically for example the multinomial logit model for a multi-category, un-ordered variable and also the sort of so called commutative logit for multi-category ordered variable so ordinal regression.

Here we are going to basically go step further. We are going to look at the outcome variable that is a count available using Poisson regression. Sometimes in the literature you may also find the expression of a log linear model.

Data for this particular session are assumed to be first of all a count variable Y so for example the number of accidents or the number of suicides in a particular geographical area or time period then we've got a categorical variable X for example which lets say a capital C possible categories such as days of the week or months. So basically Y here in this particular case has capital C possible outcomes so y 1 y 2 and so on until I yc . Obviously generally in Poisson regression modelling you may think of a number of categorical variables that you have or a number of even a continuous variables as explanatory variables in your models. Here we're going to start with something relatively simple.

Just to sort of introduce the basic principles of Poisson regression so basically it's a form of regression analysis here to model count data and in particular case if all the explanatory variables are categorical then we basically model a contingency table so basically cell counts. And the model basically models expected frequencies. The model specifies also how the count variable obviously relates to any of these explanatory variables or for example the level of the categorical variables.

Poisson models are a form of generalized linear modeling. It uses the logarithm the log as the canonical link function in this particular case. We basically assume that the outcome variable Y the dependent variable the variable that we are particularly interested in has a Poisson distribution and the logarithm basically is its expected value that can be modelled by a linear combination of any of these unknown parameters so basically of this unknown beta coefficients the regression coefficients in your model. Sometimes it's referred to as a Log linear model in particular when used to model contingency tables.

Let's have a look at a brief example for example the number of suicides by weekday in France So we've got a number of week days in the first column and the second column just simply the frequencies the occurrences the events and then let's say the percentages how it is distributed according to two days of the week.

So that's the type of a model or the type of data that we would like to model. Let's first look at a very simple case the equiprobable model. The equiprobable model means that basically all outcomes are equally probable so they are equally likely that is for our particular example we assume a uniform distribution for the outcome across days of week so Y does not vary with the days of week X basically.

So the equiprobable model is basically given by this formula here: So the probability of a particular event across these categories basically of the days of the week is equally distributed so it's 1/C so 1 over the capital C so we basically expect an equal distribution across days of week. And given this particular data we can test then the assumption of our interest , basically the assumption of the equiprobable model so H0 that this assumption holds. So looking at our example again let's say suicides by week days in France. Basically H0 the assumption that we would like to test means that each day is equally likely for the suicides to happen that means the expected proportion of suicides is about 100 / 7 so 7 days of the week. So basically just over 14% per day and if you look at the third column of the table we see the actual observed distribution and obviously that depends a little bit on each day of the week possibly and diverges a little bit from 14% per day. But may be the divergence is not very much and we are satisfied with actually our assumption and to do that properly we obviously would need to do a formal test and I'll come to that in in the next session and i will explain the extra formal test in in further detail.

Looking at another example 2 looking at traffic accidents per week again the amount to make the H0 assumption of the equiprobable model that means that each day is equally likely for an accident that means the expected proportion is again at the number of accidents is 100/7 so basically just over 14% per day we would expect. And there may be in this particular example we see a greater distribution in particular for Sunday that seems to be a greater percentage than in just fourteen percent. So you may want to continue testing if the observed distribution that we have is may be different from the expected distribution or if it's still ok to assume that they are actually equal.

Looking at hypothesis testing we may say in this particular case H0 that each day is equally likely for an accident to happen but we can also think of other alternative null hypothesis for example that each working day is equally likely for an accident or that maybe Saturday Sunday the weekends are equally likely for an accident. You could also think of course of other extra or additional variables for example the distance driven each day of the week and you may want to take into account those types of an explanatory variable as well. Just thinking about this a bit further, basically we cannot express the equiprobable model more formally as actual Poisson regression model without a covariate and that models the expected frequencies. So basically we assume a Poisson distribution with parameter move for a random component that means the response variable Y follows a Poisson distribution that means basically that that Y follows this notation here or this formula here using the exponential function and move the parameter of interest and also the Y the outcome variable of interest where Y is just simply the count variable 1 2 and 3 and so on. So basically Y is a random variable that takes on only positive integer values and also this Poisson distribution has only one single parameter μ which actually is the mean and the variance of this distribution. And we assume that our outcome follows i.e. this Poisson distribution follows the integer count distribution. Looking at basically that the simple model to start with, we aim to model the expected value of y and it can be shown that this is the parameter μ, hence we aim to model the parameter μ effectively in our Poisson model. So defining the equiprobable

model that I had on an earlier slide and sort of the intuitive notation, I'm now formalising this writing it down as the expected value of y, the parameter μ and that is 1/ C because we are making the assumption of the equal probability across weekdays. Or using the link function the log of μ would then be a coefficient alpha so that is basically the coefficient that i would like to estimate as part of my model and alpha is then basically the log of 1/C in this particular case.