Diagnostics in Poisson regression models-residual analysis

Let's look at some diagnostics as well in Poisson regression models. So I 'll introduce some residual analysis briefly just to sort of introduce some of these ideas as well although we've already discovered that based on the log likelihood test statistic and the Pearson chi-squared statistic the model may not hold but let's look at what it does to the residuals and introduce what type of residuals they might be for Poisson regression models and then we can later on extend our model to a more sophisticated model in particular using a categorical variable in the model or using another explanatory variable in the model.

So let's look at diagnostics in Poisson regression models residual analysis and let's continue with our example from our previous session about the recall of stressful event.

Residuals represent variation in the data that cannot be otherwise explained by the model. So that's quite a nice feature generally about the residuals and they can help us residual plots can help us to understand our model better and to diagnose any particular problems. So the residual plots can be used to discover certain patterns certain outliers, misspecification of the models. So basically ideally we would like to see a sort of random pattern in our residual plots and if they are awesome sort of more systematic patterns then we can identify certain particular problems and it can help us to reformulate the actual model. So if the residual exhibits no pattern then in a way that's a good indication because that would imply that the model is probably appropriate for the particular data attained.

I would like to introduce three different types of residuals for Poisson regression model so the raw residuals will just be the difference between the observed and expected values. The Pearson residuals or the standardized residuals are basically the raw residuals divided by the square root of the expected values. And then also the adjusted residuals and they can be rather helpful for actually the diagnostics and the actual plots that we are going to look at. So that is defined as the observed minus expected value is divided by the standard deviation of these observed minus expected values.

So basically we've got adjusted residuals that you would like to look at and to use for our residual plots. So if H0 is true, the adjusted residual have a standard normal distribution with the zero mean and the standard variants of 1. So basically at least for large sample that should be the case and looking at that for the recall of stressful events example that basically means that we look at the adjusted residuals and we would like to see how they compare. So we compare the observed and the expected values divided by the standard deviation of that so we obtain the adjusted residuals and for each category basically for each month we look at the adjusted residuals that are greater or smaller than 1.96 comparing it with the normal distribution that would hold if H0 actually holds. And if you see this bigger discrepancy or adjusted residuals larger than 1.96 or smaller than -1.96 that would give us an indication that there's a divergence from the H0 hypothesis. So if the adjusted residuals follow indeed the normal distribution which is true on the H0 we would expect a roughly one adjusted residual being larger than 1.96 or smaller than - 1.96 so we would only be finding one large or very small adjusted residual we would expect. Now looking at the actual data we saw that in months 1 3 & 4 we had actually positive adjusted residuals and in months 16 and 17 we had negative adjusted residuals that are a larger or smaller than 1.96. And basically we see that it's actually more likely to report more recent events so the positive residuals mean that observed data is larger than the expected data and it's more likely to report a stressful event in a month immediately prior to interview.

So we do see some sort of time trend probably in our data set that obviously isn't captured with the equiprobable model so we want to define our model in a better way or improve our model and we see in the next session how we're going to do that. Looking at the plot of adjusted residuals for months we also see a downward trend so the adjusted residuals are on the y axis and the months are given on the x-axis and we can see just by plotting those types of adjusted residuals that there is a downward trend so again it's not a random pattern there's a downward trend and we might see basically the sort of time trend.

Another way of looking at the residuals is the normal QQ plots. So basically these are probability plots that plot the quantile of one distribution with the quantiles of another distribution and here we would like to compare the distribution of the observed adjusted residuals with the expected residuals i.e. the normal residuals from a normal distribution. So basically here yeah and Q stands for quantiles i.e. for a quantile against quantiles plots effectively. So basically we are plotting observes quantiles against the expected quantiles and hence we have plotted quantities of adjusted residuals against the quantiles of the standard normal. That means that the points should actually lie just on the straight line of the y equal x line at least if the adjusted residuals indeed follow up the normal distribution which is true under the H0 hypothesis. So again we can compare divergence from the H0 hypothesis. And here looking at the QQ plots from this particular data said we can see that in the tail so in the upper end and in the lower end that is divergence from the straight line , relationship so again we would conclude that there is some kind of time trend in our data set.

So conclusions are clearly that there is divergence in the tales from these straight lines. There is overall strong evidence that the equiprobable model doesn't really hold doesn't fit the data that well which may be isn't too surprising and we would like or will see sort of more likely that the data is more likely to report recent events. So basically such a tendency would result if respondents were more likely to remember recent events than distant events.

So basically again there is strong evidence that we should be using some kind of other model and which one to use and we are going to now explore the Poisson time trend model a Poisson model with a co-variate.