Poisson Regression

Practical in R

5th May 2016

1 Getting started

- Log into Moodle at *www.ucl.ac.uk/moodle*
- Download the file `poisson.Rdata` to your space in the workstation.
- Start *RStudio*

2 Examine data

Import the dataset into R and load the required packages[1]:

```
library(ggplot2)
library(pscl)
load('poisson.Rdata')
```

This creates two new data frames in the current `R` environment, labelled `pd` and `zip`. This section will use the `pd` data frame. You can obtain a list of the variables in this dataset using the `names` command:

```
names(pd)
```

These are described below.

Table 1: Variables in the data frame `pd`

Variable	Label
`idauniq`	Unique individual serial number
`iadl`	How many difficulties with iadl
`iintdtm`	Month of individual interview
`iintdty`	Year of individual interview
`sex`	Sex

[1] If you haven't used these packages before, you'll need to install them on your machine. Do this by typing, e.g. `install.packages("ggplot2")`.

Variable	Label
age	Age (collapsed at 90)
angina	
diabete	
arthriti	Doctor diagnosed arthritis
limitill	Limiting longstanding illness
currsmk	Current smoker
physact	
marstat2	Whether living with partner
nssec3	Socio-economic classification (NS-SEC3)
adl	Number of ADL difficulties
time	

- Explore the variable `adl`. This variable measures the number of difficulties with six *Activities of Daily Living (ADL)* such as dressing, walk across a room, bathing, eating, getting in and out of bed and using the toilet. This is an important measure of physical functioning in old age.
- Obtain summary statistics for this variable, and explore its distribution.

```
summary(pd$adl)

length(pd$adl)

median(pd$adl)

range(pd$adl)

print(adl_tab <- table(pd$adl))
round(prop.table(adl_tab), 3)

qplot(pd$adl, geom = 'histogram',
      xlab ='Number of difficulties',
      bins = 20)
```

3 Poisson regression

A useful place to begin when analyzing a count outcome is to compare the observed distribution of the variable (`adl`) with a Poisson distribution that has the same mean.

First we run a Poisson regression without any independent variables in order to fit a univariate Poisson distribution with the mean equal to that of our variable `adl`.

To do so type:

```
summary(m1 <- glm(adl ~ 1, data = pd, family = poisson))
```

Because the intercept from this model is equal to 0.580, $\mu = exp(0.580) = 1.786$, which is the same as the estimated mean obtained with `summary` earlier.

We can compare the observed distribution of `adl` with the probabilities predicted by our model. First, calculate the predicted probability of each count (ranging from 0 to 6) from the above model (`m1`):

```
pred <- predict(m1, type="response")[1]
pred_probs <- apply(array(0:6), 1,
                    function(count) dpois(lambda = pred, x = count))
```

We can now compare graphically the observed probabilities for each value of `adl` with the predicted probabilities from fitting a Poisson distribution with no independent variables (the null or empty Poisson regression) – we do so to try to understand whether Poisson could be the adequate regression for our outcome.

```
# Create a data frame containing the predicted and observed
# probabilities for each count of ADL:
observed_probs <- prop.table(table(pd$adl))
plot_data <- data.frame(count = rep(0:6, 2),
                        type = c(rep('observed', 7),
                                 rep('predicted', 7)),
                        y = c(observed_probs, pred_probs))

# Then plot this.
ggplot(plot_data, aes(y = y, x = count, group = type, color = type)) +
    geom_line() +
    geom_point() +
    ggtitle('Predicted vs. observed probabilities of ADL count') +
    ylab('Probability') +
    xlab('Number of ADLs reported')
```

On the x-axis we have the reported 'Number of difficulties with ADL' while on the y-axis we have the observed and predicted probabilities with which each count occurs.

- Question: **What can you say about the fitted Poisson distribution?**

3.0.1 Explore relationships between ADLs, sex and age.

We want to explore the relationship between number of ADLs reported (`adl`), sex and age. Run the following model:

```r
summary(m2 <- glm(adl ~  factor(sex) + age, data = pd, family = poisson))
```

We can calculate the incidence rate ratios (IRRs) by exponentiating the Poisson regression coefficients[2]:

```r
round(exp(coef(m2)), 3)
```

- Interpret the results from this model.
- Perform a goodness-of-fit test of the model and write down the null hypothesis. Based on the p-value say whether the Poisson distribution is appropriate or not for our outcome variable **adl**.

```r
with(m2, cbind(res.deviance = deviance,
               df = df.residual,
               p = pchisq(deviance,
                          df.residual,
                          lower.tail=FALSE)))
```

4 Zero inflated poisson regression (ZIP)

This section will use the `zip` data frame.. These data have a reclassified version of the Activities of Daily Living (ADL) measure examined above. Run a tabulation of the variable `adl`:

```r
adl_freq <- table(zip$adl)
adl_freq

round(prop.table(adl_freq), 3)
```

- Question: **What is the percentage of zeros?**

We'll fitting ZIP models using the `zeroinfl` function from the `pscl` library. Details about this library can be found here: https://cran.r-project.org/web/packages/pscl/vignettes/countreg.pdf

Run the ZIP model show below, and interpret the results.

[2]For a more efficient way of doing this in R, take a look at the `stargazer` package: https://cran.r-project.org/web/packages/stargazer/vignettes/stargazer.pdf

```
summary(m3 <- zeroinfl(adl ~ indager + indsex + smok2 + ed2 +
                           limitill + livpart, data = zip))
```

As before, to extract the incidence rate ratios (IRRs) for the Poisson part, we need to exponentiate the Poisson coefficients:

```
irr <- exp(coef(m3)[1:7])

round(irr, 3)
```

To carry out a Vuong test, comparing the ZIP model to the Poisson Regression model, we use the `vuong` function (also from the `pscl` package).

```
# Estimate the equivalent Poisson regression model
m3_poisson <- glm(adl ~ indager + indsex + smok2 + ed2 +
                        limitill + livpart,
                        data = zip,
                        family = poisson)
# Compare this to the ZIP equivalent
vuong(m3_poisson, m3)
```

- Question: **Does the Vuong test suggest that the ZIP fits the data better than the Poisson?**

5 Exercises

5.1 Exercise 1

- Using the data frame `pd`, run a Poisson regression with `adl` as the dependent variable and `sex`, `age`, `limitill`, `arthriti` and `physact` as independent variables. Interpret the results.
- Tabulate `arthriti` and `physact` first.
- After your initial regression, include time as an exposure variable.
- Question: **What effect does including time have on your results? Can you explain the reasons why your results do/do not change?**

5.2 Exercise 2

- Repeat a Poisson regression with `iadl` as dependent variable and `sex`, `age` as independent variables.
  - `iadl` is a count variable which indicates the number of difficulties with Instrumental Activity of Daily Living (complex skills needed to live independently)
- Run a goodness of fit test and explain the results. Is there a better model to use?