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)
```

```
##  [1] "idauniq"  "iadl"     "iintdtm"  "iintdty"  "sex"      "age"
##  [7] "angina"   "diabete"  "arthriti" "limitill" "currsmk"  "physact"
## [13] "marstat2" "nssec3"   "adl"      "time"
```

These are described below.

[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")`.

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
`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)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   2.000   1.785   3.000   6.000
```

```
length(pd$adl)
```

```
## [1] 11213
```

```
median(pd$adl)
```

```
## [1] 2
```

```r
range(pd$adl)
```

```
## [1] 0 6
```

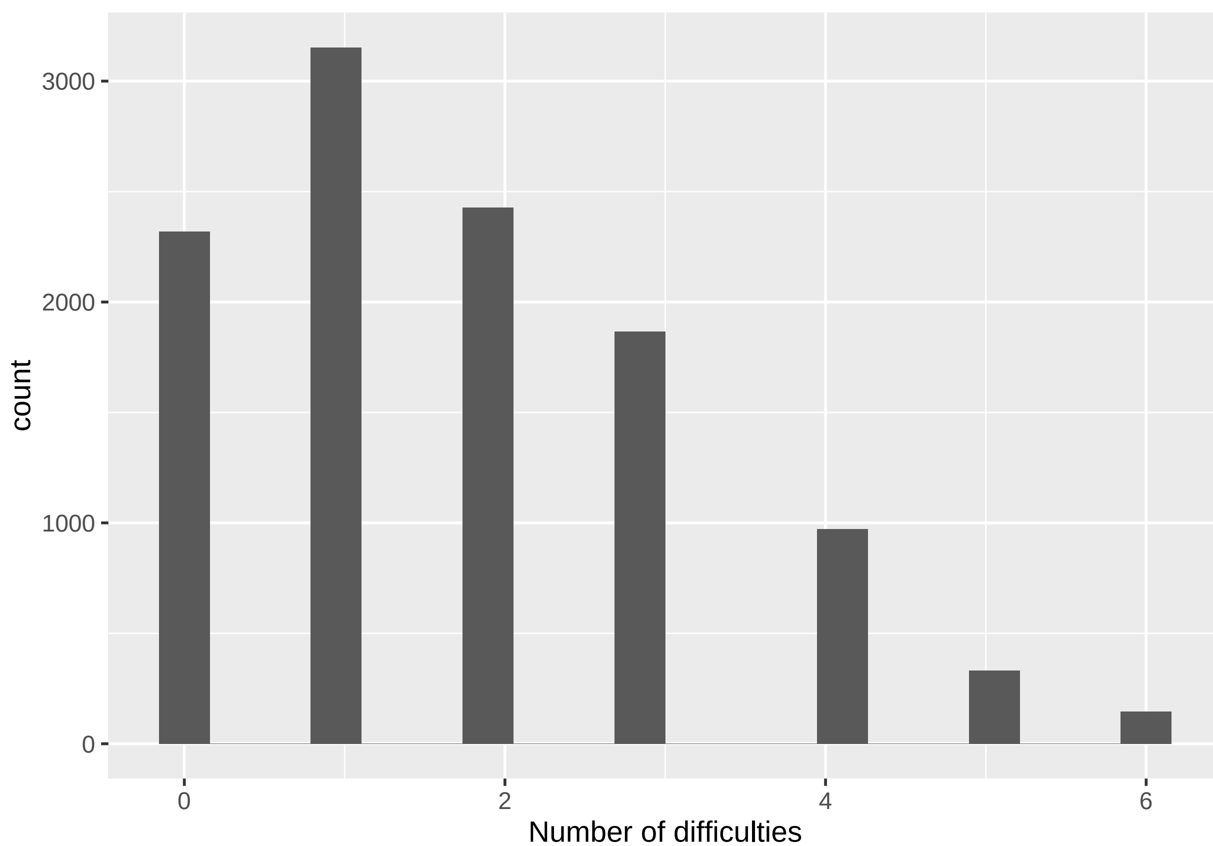
```r
print(adl_tab <- table(pd$adl))
```

```
##
##    0    1    2    3    4    5    6
## 2319 3152 2427 1867  972  331  145
```

```r
round(prop.table(adl_tab), 3)
```

```
##
##     0     1     2     3     4     5     6
## 0.207 0.281 0.216 0.167 0.087 0.030 0.013
```

```r
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))
```

```
##
## Call:
## glm(formula = adl ~ 1, family = poisson, data = pd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8897  -0.6415   0.1575   0.8274   2.4731
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.579658   0.007068   82.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15051  on 11212  degrees of freedom
## Residual deviance: 15051  on 11212  degrees of freedom
## AIC: 38143
##
## Number of Fisher Scoring iterations: 5
```

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.

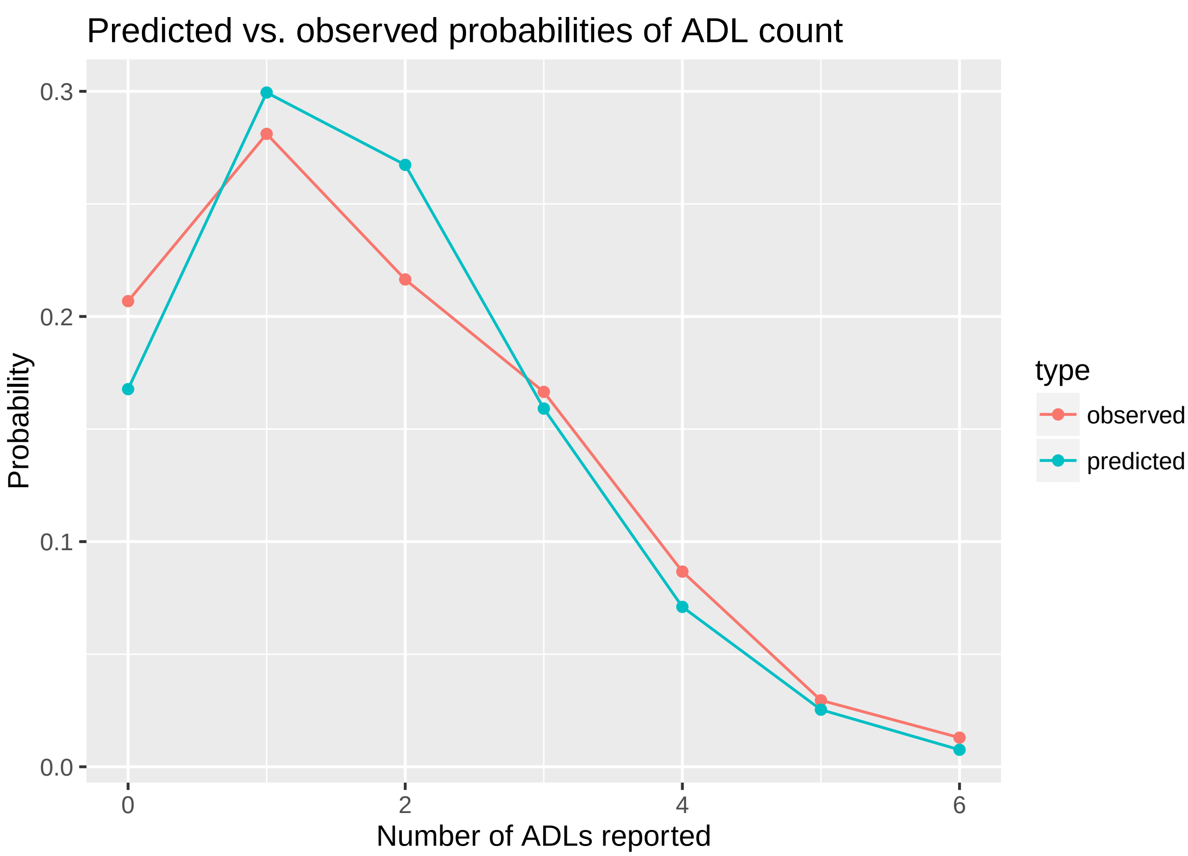
We can compare the observed distribution of `adl` with the probabilities predicted by our model. First, calculate the predicted probability of each count (ranging from 0 to 6) from the above model (`m1`):

```
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:

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

```
##
## Call:
## glm(formula = adl ~ factor(sex) + age, family = poisson, data = pd)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0925  -1.2514  -0.1335   0.4703   3.7112
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.3485063  0.0449914  -29.97   <2e-16 ***
## factor(sex)female  0.7277099  0.0156852   46.40   <2e-16 ***
## age                0.0220767  0.0006402   34.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15051  on 11212  degrees of freedom
## Residual deviance: 11430  on 11210  degrees of freedom
## AIC: 34526
##
## Number of Fisher Scoring iterations: 5
```

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

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

```
##       (Intercept) factor(sex)female               age
##             0.260             2.070             1.022
```

[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

- 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**.

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

```
##      res.deviance    df          p
## [1,]     11430.13 11210 0.07144796
```

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`:

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

```
##
##    0    1    2    3    4    5    6
## 8872 1153  547  289  193  104   41
```

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

```
##
##     0     1     2     3     4     5     6
## 0.792 0.103 0.049 0.026 0.017 0.009 0.004
```

- 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.

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

```
##
## Call:
## zeroinfl(formula = adl ~ indager + indsex + smok2 + ed2 + limitill +
##     livpart, data = zip)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.1968 -0.3351 -0.2254 -0.1643 11.3836
##
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.5368370  0.1515994  -3.541 0.000398 ***
## indager     -0.0001166  0.0018707  -0.062 0.950302
## indsex       0.0200817  0.0403045   0.498 0.618308
## smok2        0.0269320  0.0493584   0.546 0.585313
## ed2          0.0693986  0.0420627   1.650 0.098967 .
## limitill     1.0921985  0.0728836  14.986  < 2e-16 ***
## livpart     -0.0085028  0.0418821  -0.203 0.839121
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.953758   0.268037  18.482  < 2e-16 ***
## indager     -0.045110   0.003809 -11.843  < 2e-16 ***
## indsex       0.103889   0.075067   1.384   0.1664
## smok2       -0.162843   0.092279  -1.765   0.0776 .
## ed2         -0.359084   0.075592  -4.750 2.03e-06 ***
## limitill    -1.817614   0.092467 -19.657  < 2e-16 ***
## livpart     -0.367961   0.080681  -4.561 5.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 42
## Log-likelihood: -7504 on 14 Df
```

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

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

round(irr, 3)
```

```
## count_(Intercept)     count_indager      count_indsex       count_smok2
##             0.585             1.000             1.020             1.027
##         count_ed2    count_limitill     count_livpart
##             1.072             2.981             0.992
```

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).

```r
# 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)
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishible)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   -16.97164 model2 > model1 < 2.22e-16
## AIC-corrected         -16.82572 model2 > model1 < 2.22e-16
## BIC-corrected         -16.29140 model2 > model1 < 2.22e-16
```

- 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?