Models for count data

SISS - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-12-04

Load data

```
rm(list=ls())

library(VIM) # Useful to analyze missing data
```

```
## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(mice) # Useful to analyze missing data
```

```
##
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
##
##     filter

## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
library(foreign) # Package that allows you to read .dta data
library(psych)
library(MatchIt)
library(bestNormalize)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(GGally)
```

```
## Loading required package: ggplot2

##
## Attaching package: 'ggplot2'
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:bestNormalize':
##
##     boxcox
```

```
library(AER)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:psych':
##
##     logit
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
```

```
library(pscl)
```

```
## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis
```

Have a preliminary overview of data

look at the columns

```
data("DoctorVisits")
doc <- DoctorVisits
```

Data come from an Australian health survey, where visits is the number of doctor visits in past two weeks.

Overview

```
summary(doc)
```

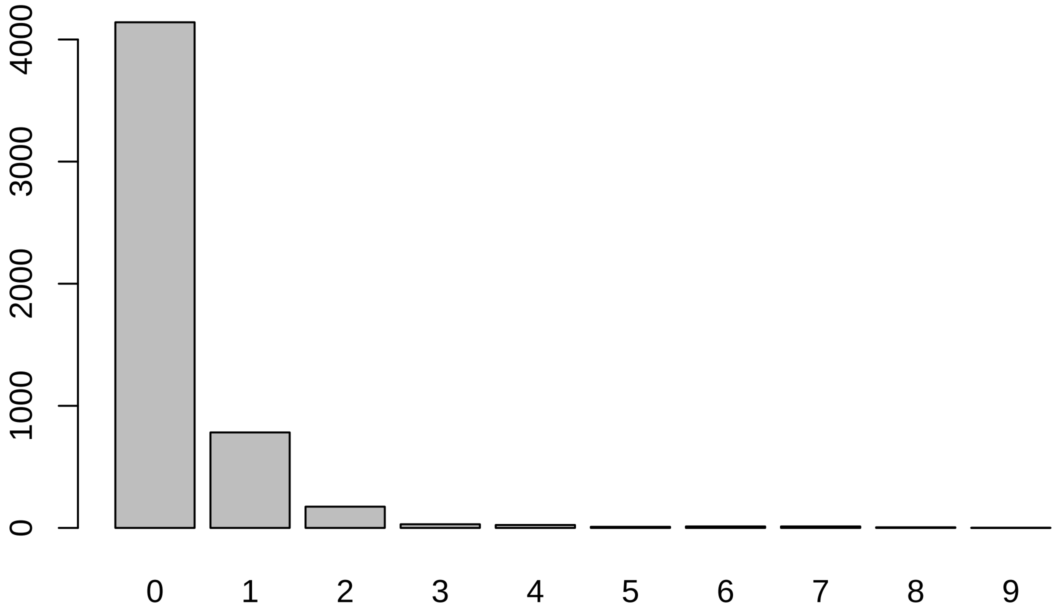
```
##      visits          gender          age             income      
##  Min.   :0.0000   male  :2488   Min.   :0.1900   Min.   :0.0000  
##  1st Qu.:0.0000   female:2702   1st Qu.:0.2200   1st Qu.:0.2500  
##  Median :0.0000                 Median :0.3200   Median :0.5500  
##  Mean   :0.3017                 Mean   :0.4064   Mean   :0.5832  
##  3rd Qu.:0.0000                 3rd Qu.:0.6200   3rd Qu.:0.9000  
##  Max.   :9.0000                 Max.   :0.7200   Max.   :1.5000  
##     illness         reduced            health       private    freepoor  
##  Min.   :0.000   Min.   : 0.0000   Min.   : 0.000   no :2892   no :4968  
##  1st Qu.:0.000   1st Qu.: 0.0000   1st Qu.: 0.000   yes:2298   yes: 222  
##  Median :1.000   Median : 0.0000   Median : 0.000                         
##  Mean   :1.432   Mean   : 0.8619   Mean   : 1.218                         
##  3rd Qu.:2.000   3rd Qu.: 0.0000   3rd Qu.: 2.000                         
##  Max.   :5.000   Max.   :14.0000   Max.   :12.000                         
##  freerepat  nchronic   lchronic  
##  no :4099   no :3098   no :4585  
##  yes:1091   yes:2092   yes: 605  
##                                  
##                                  
##                                  
##                                  
```

```
head(doc)
```

```
##   visits gender  age income illness reduced health private freepoor freerepat
## 1      1 female 0.19   0.55       1       4      1     yes       no        no
## 2      1 female 0.19   0.45       1       2      1     yes       no        no
## 3      1   male 0.19   0.90       3       0      0      no       no        no
## 4      1   male 0.19   0.15       1       0      0      no       no        no
## 5      1   male 0.19   0.45       2       5      1      no       no        no
## 6      1 female 0.19   0.35       5       1      9      no       no        no
##   nchronic lchronic
## 1       no       no
## 2       no       no
## 3       no       no
## 4       no       no
## 5      yes       no
## 6      yes       no
```

Preliminary inspection of the outcome variable

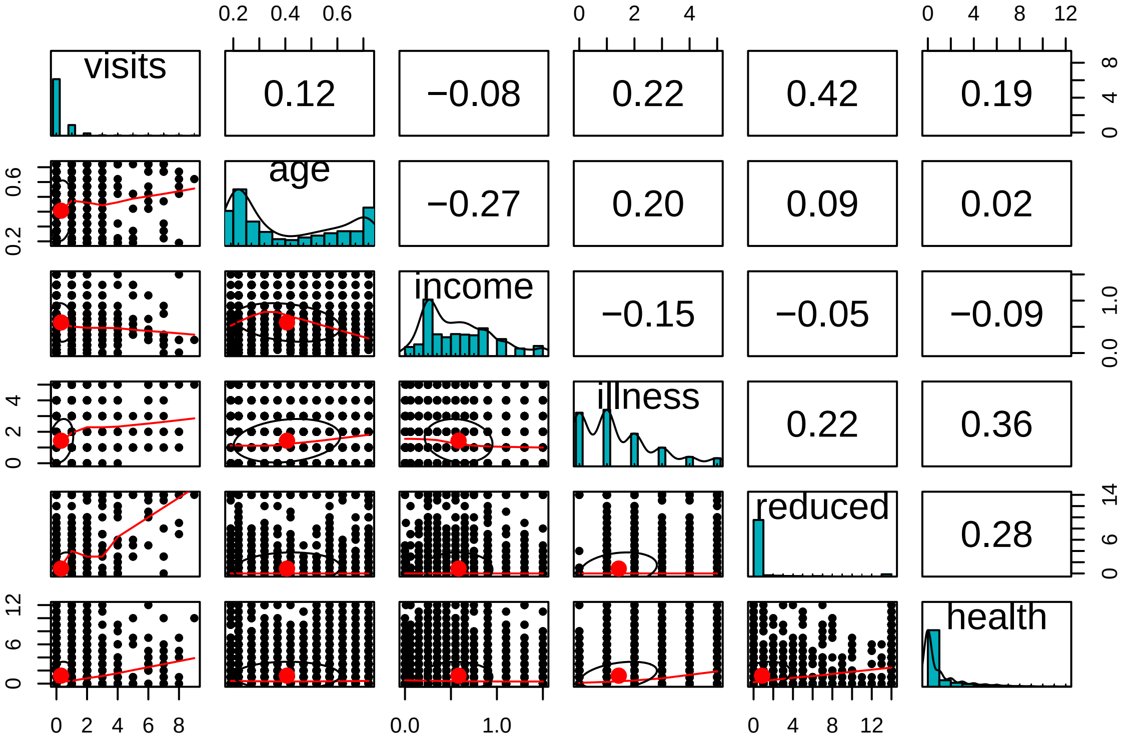
```
counts <- table(doc$visits)
barplot(counts)
```



Correlation in the data

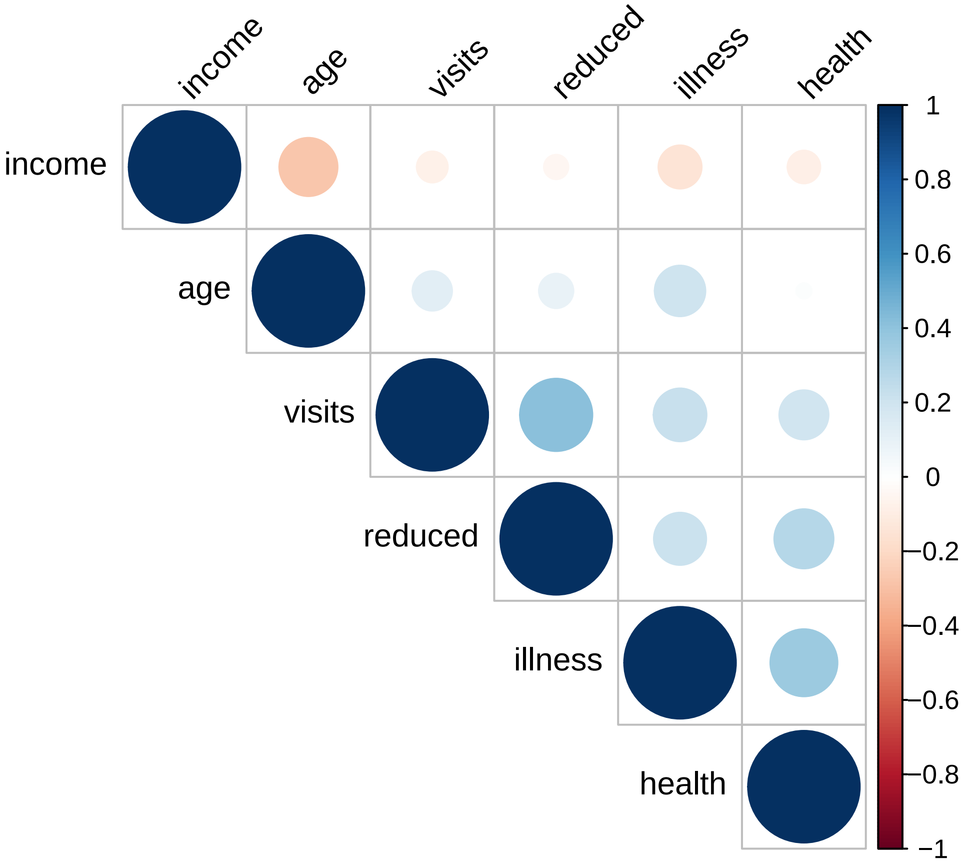
```
quantitative_variables <- c("visits","age","income",
                            "illness","reduced","health")

pairs.panels(doc[,quantitative_variables],
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE # show correlation ellipses
             )
```



```
correlation_matrix <- cor(doc[, quantitative_variables])

corrplot(correlation_matrix, type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Fit your models

Simple case

Naive linear model

```
simple_naive_lm <- lm(visits ~ age,
              data = doc)
summary(simple_naive_lm)
```

```
##
## Call:
## lm(formula = visits ~ age, data = doc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4540 -0.3569 -0.2113 -0.1967  8.5946
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.10448    0.02443   4.277 1.93e-05 ***
## age          0.48538    0.05369   9.040  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.792 on 5188 degrees of freedom
## Multiple R-squared:  0.01551,    Adjusted R-squared:  0.01532
## F-statistic: 81.73 on 1 and 5188 DF,  p-value: < 2.2e-16
```

```
simple_naive_lm_predicted <- predict(simple_naive_lm,
                                     type="response")
```

Poisson model

```
simple_poisson <- glm(visits ~ age, family="poisson",
              data = doc)
summary(simple_poisson)
```

```
##
## Call:
## glm(formula = visits ~ age, family = "poisson", data = doc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9650  -0.8266  -0.6552  -0.6402   6.5608
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.87944    0.06248  -30.08   <2e-16 ***
## age          1.54878    0.12060   12.84   <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: 5634.8  on 5189  degrees of freedom
## Residual deviance: 5470.0  on 5188  degrees of freedom
## AIC: 7805.6
##
## Number of Fisher Scoring iterations: 6
```

```
simple_poisson_predicted <- predict.glm(simple_poisson, type="response")
```

Multivariate model

Poisson model

```
full_poisson <- glm(visits ~ age + income + illness +
                   reduced + health + private + freepoor                  + freerepat + nchronic + lchr
                data = doc)

summary(full_poisson)
```

```
##
## Call:
## glm(formula = visits ~ age + income + illness + reduced + health +
##     private + freepoor + freerepat + nchronic + lchronic, family = "poisson",
##     data = doc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8966  -0.6837  -0.5783  -0.4935   5.8087
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.019864   0.097016 -20.820  < 2e-16 ***
## age           0.307919   0.165699   1.858  0.06313 .
## income       -0.236007   0.083480  -2.827  0.00470 **
## illness       0.186534   0.018300  10.193  < 2e-16 ***
## reduced       0.126523   0.005032  25.145  < 2e-16 ***
## health        0.031450   0.010097   3.115  0.00184 **
## privateyes    0.153021   0.070856   2.160  0.03080 *
## freepooryes  -0.454270   0.179667  -2.528  0.01146 *
## freerepatyes  0.111385   0.091495   1.217  0.22345
## nchronicyes   0.130685   0.066399   1.968  0.04905 *
## lchronicyes   0.151017   0.082302   1.835  0.06652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4388.0  on 5179  degrees of freedom
## AIC: 6741.5
##
## Number of Fisher Scoring iterations: 6
```

Have a look at the statistics of your model

```
glance(full_poisson)
```

```
## # A tibble: 1 x 8
##   null.deviance df.null logLik   AIC   BIC deviance df.residual  nobs
##           <dbl>   <int>  <dbl> <dbl> <dbl>    <dbl>       <int> <int>
## 1         5635.    5189 -3360. 6742. 6814.    4388.        5179  5190
```

Run an overdispersion test

```
dispersiontest(full_poisson, trafo = 1)
```

```
##
##  Overdispersion test
##
## data:  full_poisson
## z = 6.417, p-value = 6.949e-11
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##     alpha
## 0.4206546
```

```
dispersiontest(full_poisson, trafo = 2)
```

```
##
```

```
##  Overdispersion test
##
## data:  full_poisson
## z = 7.2137, p-value = 2.723e-13
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##     alpha
## 0.9533897
```

Quasi-poisson

```
full_quasipoisson <- glm(visits ~ age + income + illness +
                   reduced + health + private + freepoor                  + freerepat + nchronic + lchr
                data = doc)

summary(full_quasipoisson)
```

```
##
## Call:
## glm(formula = visits ~ age + income + illness + reduced + health +
##     private + freepoor + freerepat + nchronic + lchronic, family = "quasipoisson",
##     data = doc)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8966  -0.6837  -0.5783  -0.4935   5.8087
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.019864   0.112096 -18.019  < 2e-16 ***
## age           0.307919   0.191454   1.608  0.10783
## income       -0.236007   0.096456  -2.447  0.01445 *
## illness       0.186534   0.021145   8.822  < 2e-16 ***
## reduced       0.126523   0.005814  21.762  < 2e-16 ***
## health        0.031450   0.011667   2.696  0.00705 **
## privateyes    0.153021   0.081869   1.869  0.06167 .
## freepooryes  -0.454270   0.207593  -2.188  0.02869 *
## freerepatyes  0.111385   0.105716   1.054  0.29210
## nchronicyes   0.130685   0.076720   1.703  0.08855 .
## lchronicyes   0.151017   0.095094   1.588  0.11233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.335023)
##
##     Null deviance: 5634.8  on 5189  degrees of freedom
## Residual deviance: 4388.0  on 5179  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Negative-binomial

```
full_nb <- glm.nb(visits ~ age + income + illness +
                   reduced + health + private + freepoor                  + freerepat + nchronic + lchr
                data = doc)

summary(full_nb)
```

```
##
## Call:
## glm.nb(formula = visits ~ age + income + illness + reduced +
##     health + private + freepoor + freerepat + nchronic + lchronic,
##     data = doc, init.theta = 0.9263625049, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9404  -0.6340  -0.5317  -0.4526   4.1533
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.169191   0.117309 -18.491  < 2e-16 ***
## age           0.367328   0.207190   1.773  0.07624 .
## income       -0.222546   0.101391  -2.195  0.02817 *
## illness       0.216354   0.023525   9.197  < 2e-16 ***
## reduced       0.143227   0.007311  19.591  < 2e-16 ***
## health        0.037868   0.013632   2.778  0.00547 **
## privateyes    0.159512   0.084577   1.886  0.05930 .
## freepooryes  -0.508935   0.210260  -2.420  0.01550 *
## freerepatyes  0.187085   0.115156   1.625  0.10424
## nchronicyes   0.112568   0.078963   1.426  0.15399
## lchronicyes   0.188378   0.103101   1.827  0.06768 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.9264) family taken to be 1)
##
##     Null deviance: 3926.5  on 5189  degrees of freedom
## Residual deviance: 3036.5  on 5179  degrees of freedom
## AIC: 6431.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.9264
##           Std. Err.:  0.0866
##
##  2 x log-likelihood:  -6407.4470
```

Zero inflated poisson

```
full_zeroinfl_poisson <-zeroinfl(visits ~ age + income +                                         illnes
                         reduced + health + private + freepoor                                   + free
```

```
                dist="poisson", data = doc)
summary(full_zeroinfl_poisson )
```

```
## 
## Call:
## zeroinfl(formula = visits ~ age + income + illness + reduced + health + 
##     private + freepoor + freerepat + nchronic + lchronic, data = doc, 
##     dist = "poisson")
## 
## Pearson residuals:
##     Min      1Q  Median      3Q     Max 
## -1.6001 -0.4481 -0.2960 -0.1981 11.0898 
## 
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  -0.574052   0.136098  -4.218 2.47e-05 ***
## age           0.029502   0.217464   0.136   0.8921    
## income       -0.194272   0.107371  -1.809   0.0704 .  
## illness       0.041089   0.024596   1.671   0.0948 .  
## reduced       0.082462   0.005980  13.790  < 2e-16 ***
## health        0.023094   0.011268   2.049   0.0404 *  
## privateyes   -0.031874   0.095971  -0.332   0.7398    
## freepooryes  -0.397713   0.243856  -1.631   0.1029    
## freerepatyes -0.221033   0.117441  -1.882   0.0598 .  
## nchronicyes  -0.004719   0.092223  -0.051   0.9592    
## lchronicyes   0.005587   0.101978   0.055   0.9563    
## 
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   2.33996    0.27286   8.576  < 2e-16 ***
## age          -1.34358    0.50724  -2.649  0.00808 ** 
## income        0.11825    0.23002   0.514  0.60719    
## illness      -0.45239    0.08351  -5.417 6.06e-08 ***
## reduced      -1.26427    0.23692  -5.336 9.49e-08 ***
## health       -0.07919    0.03839  -2.063  0.03914 *  
## privateyes   -0.53198    0.19383  -2.745  0.00606 ** 
## freepooryes   0.28968    0.51774   0.560  0.57581    
## freerepatyes -1.31609    0.31447  -4.185 2.85e-05 ***
## nchronicyes  -0.11568    0.19637  -0.589  0.55579    
## lchronicyes  -0.43365    0.30287  -1.432  0.15220    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Number of iterations in BFGS optimization: 46 
## Log-likelihood: -3186 on 22 Df
```

```
logLik(full_poisson)
```

```
## 'log Lik.' -3359.774 (df=11)
```

```
logLik(full_zeroinfl_poisson)
```

```
## 'log Lik.' -3185.671 (df=22)
```

```
logLik(full_quasipoisson)
```

```
## 'log Lik.' NA (df=11)
```

```
logLik(full_nb)
```

```
## 'log Lik.' -3203.723 (df=12)
```