453 HW4

Max Tjen

2023-07-05

Packages

```
library(countreg)
library(pscl)
library(lmtest)
library(dplyr)
library(tidyverse)
```

Question 16

```
data16 <- read_csv("/Users/mtjen/Desktop/453/hw4/dt.csv")
```

Part A

```
mod16a <- glm(ofp ~ hosp + numchron + gender + school +
                privins + health_excellent + health_poor,
              family = poisson(link = "log"),
              data = data16)

mod16a
```

```
##
## Call:  glm(formula = ofp ~ hosp + numchron + gender + school + privins +
##     health_excellent + health_poor, family = poisson(link = "log"),
##     data = data16)
##
## Coefficients:
##      (Intercept)             hosp          numchron            gender
##          1.02887           0.16480           0.14664          -0.11232
##           school           privins  health_excellent       health_poor
##          0.02614           0.20169          -0.36199           0.24831
##
## Degrees of Freedom: 4405 Total (i.e. Null);  4398 Residual
## Null Deviance:       26940
## Residual Deviance: 23170     AIC: 35960
```

Part B

```
# parameter effects
effects <- round(100 * (exp(mod16a$coefficients) - 1)[2:8], 3)

hospInt <- round(100 * (exp(confint(mod16a, parm = "hosp")) - 1), 3)
numchronInt <- round(100 * (exp(confint(mod16a, parm = "numchron")) - 1), 3)
genderInt <- round(100 * (exp(confint(mod16a, parm = "gender")) - 1), 3)
schoolInt <- round(100 * (exp(confint(mod16a, parm = "school")) - 1), 3)
privinsInt <- round(100 * (exp(confint(mod16a, parm = "privins")) - 1), 3)
health_excellentInt <- round(100 * (exp(confint(mod16a,
                                                parm = "health_excellent")) - 1), 3)
health_poorInt <- round(100 * (exp(confint(mod16a,
                                           parm = "health_poor")) - 1), 3)

# parameter CI lower bound
confLow <- c(hospInt[1], numchronInt[1], genderInt[1],
             schoolInt[1], privinsInt[1],
             health_excellentInt[1], health_poorInt[1])

# parameter CI lower bound
confHigh <- c(hospInt[2], numchronInt[2], genderInt[2],
             schoolInt[2], privinsInt[2],
             health_excellentInt[2], health_poorInt[2])

results <- data.frame(matrix(nrow = 7))[,-1]
results$parameter <- colnames(data16)[2:8]
results$effect <- effects
results$low <- confLow
results$high <- confHigh
results
```

```
##          parameter  effect     low    high
## 1             hosp  17.915  16.529  19.301
## 2         numchron  15.794  14.757  16.836
## 3           gender -10.624 -12.866  -8.330
## 4           school   2.649   2.279   3.021
## 5          privins  22.346  18.380  26.469
## 6 health_excellent -30.371 -34.420 -26.147
## 7      health_poor  28.185  23.769  32.737
```

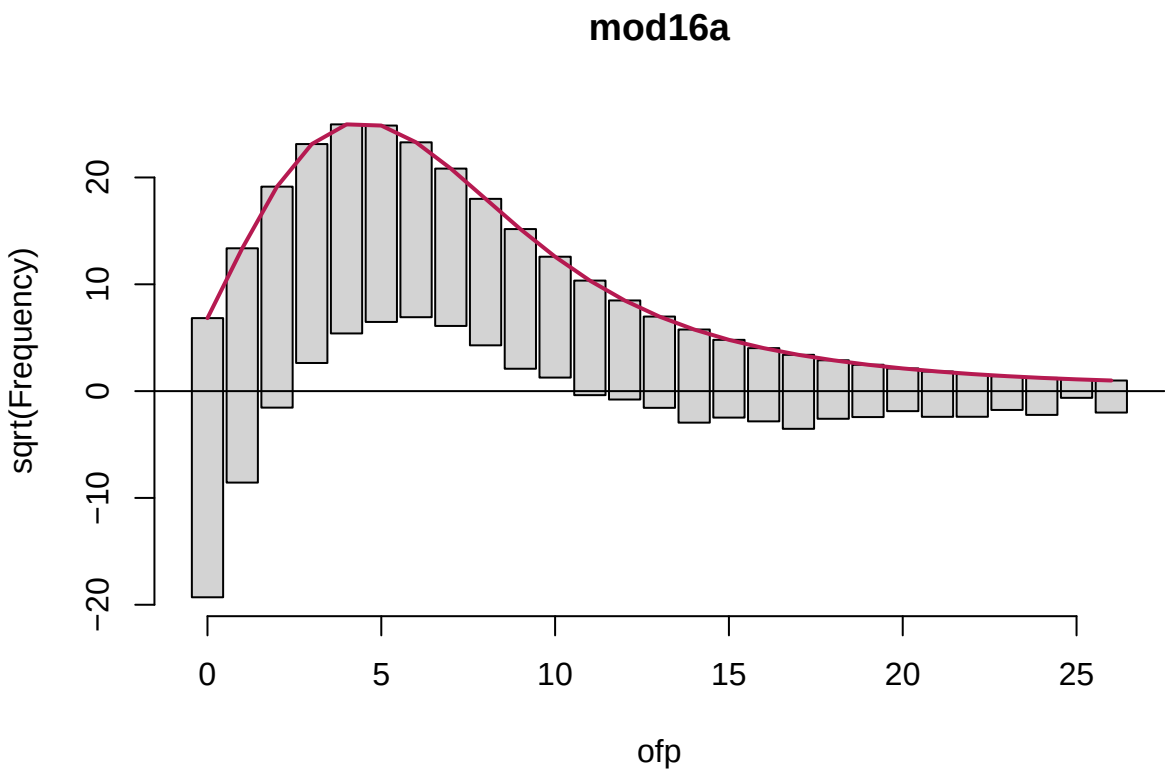
We can see that there is a positive association between the number of physician office visits someone has with the number of hospital stays, number of chronic conditions, number of years of education, having private insurance, and being labeled as having poor health. There is a negative association between gender (being male) and being labeled as having excellent health. All of these variables' confidence intervals don't include 0, confirming their effect direction and relationship with one's number of physician office visits.

Part C

```
# actual visits
table(data16$ofp)[1]
```

```
##   0
## 683
```

```
rootogram(mod16a)
```



We can see via the rootogram that the model heavily underfits 0 values for the number of physician office visits one has. This means that the model doesn't predict many 0 values despite the data having a decent amount. There may be many zero values in the data because people typically don't go for small issues and only go when it's really needed or required. As such, most people don't often go to the physician's office even if they probably should.

Part D

```
zipMod16 <- zeroinfl(ofp ~ hosp + numchron + gender + school +
                       privins + health_excellent + health_poor | 1,
                     dist = "poisson",
                     data = data16)

# parameter effects
round(100 * (exp(zipMod16$coefficients$count) - 1)[2:8], 3)
```

```
##             hosp         numchron           gender           school
##           17.249           10.903           -6.317            1.991
##          privins health_excellent      health_poor
##            8.966          -27.326           28.961
```

```
round(100 * (exp(confint(zipMod16)) - 1)[2:8,], 3)
```

```
##                           2.5 %  97.5 %
## count_hosp               15.865  18.650
## count_numchron            9.878  11.938
## count_gender             -8.696  -3.876
## count_school              1.614   2.369
## count_privins             5.328  12.731
## count_health_excellent -31.724 -22.644
## count_health_poor        24.559  33.519
```

We can see that similar to before, there is a positive association between the number of physician office visits someone has with the number of hospital stays, number of chronic conditions, number of years of education, having private insurance, and being labeled as having poor health. There is a negative association between gender (being male) and being labeled as having excellent health. The effect values for number of hospital stays, number of years of education, being labeled as having excellent health, and being labeled as having poor health are relatively similar to the poisson regression model. However, the effect values of number of chronic conditions, gender (being male), and having private insurance are noticeably different than before, although all still have the same effect direction. As before, the confidence intervals for all variables don't include 0, which means that we can be 95% confident about the true direction of effects.

Question 32

```
data32 <- read_csv("/Users/mtjen/Desktop/453/hw4/pregnancy.csv")
```

Part A

```
data32 <- data32 |>
  mutate(smokef = factor(Smoke),
         socialf = factor(Social))

cTab <- xtabs(Count ~ HT + PU + smokef + socialf,
              data = data32)

cTab
```

```
## , , smokef = 1, socialf = 1
##
##    PU
## HT     n    y
##   n  286   21
##   y   82   28
##
```

```
## , , smokef = 2, socialf = 1
##
##    PU
## HT     n    y
##   n   71    5
##   y   24    5
##
## , , smokef = 3, socialf = 1
##
##    PU
## HT     n    y
##   n   13    0
##   y    3    1
##
## , , smokef = 1, socialf = 2
##
##    PU
## HT     n    y
##   n  785   34
##   y  266   50
##
## , , smokef = 2, socialf = 2
##
##    PU
## HT     n    y
##   n  284   17
##   y   92   13
##
## , , smokef = 3, socialf = 2
##
##    PU
## HT     n    y
##   n   34    3
##   y   15    0
##
## , , smokef = 1, socialf = 3
##
##    PU
## HT     n    y
##   n 3160  164
##   y 1101  278
##
## , , smokef = 2, socialf = 3
##
##    PU
## HT     n    y
##   n 2300  142
##   y  492  120
##
## , , smokef = 3, socialf = 3
##
##    PU
## HT     n    y
##   n  383   32
```

```
##   y   92   16
##
## , , smokef = 1, socialf = 4
##
##    PU
## HT     n    y
##   n  656   52
##   y  213   63
##
## , , smokef = 2, socialf = 4
##
##    PU
## HT     n    y
##   n  649   46
##   y  129   35
##
## , , smokef = 3, socialf = 4
##
##    PU
## HT     n    y
##   n  163   12
##   y   40    7
##
## , , smokef = 1, socialf = 5
##
##    PU
## HT     n    y
##   n  245   23
##   y   78   20
##
## , , smokef = 2, socialf = 5
##
##    PU
## HT     n    y
##   n  321   34
##   y   74   22
##
## , , smokef = 3, socialf = 5
##
##    PU
## HT     n    y
##   n   65    4
##   y   14    7
```

```
fTab <- ftable(cTab,
               row.vars = c("smokef", "socialf"),
               col.vars = c("HT", "PU"))

fTab
```

```
##                HT    n         y
##                PU    n    y    n    y
## smokef socialf
## 1      1           286   21   82   28
```

```
##        2           785   34  266   50
##        3          3160  164 1101  278
##        4           656   52  213   63
##        5           245   23   78   20
## 2      1            71    5   24    5
##        2           284   17   92   13
##        3          2300  142  492  120
##        4           649   46  129   35
##        5           321   34   74   22
## 3      1            13    0    3    1
##        2            34    3   15    0
##        3           383   32   92   16
##        4           163   12   40    7
##        5            65    4   14    7
```

```
round(prop.table(fTab, margin = 1), 3)
```

```
##                HT     n           y
##                PU     n     y     n     y
## smokef socialf
## 1      1          0.686 0.050 0.197 0.067
##        2          0.692 0.030 0.234 0.044
##        3          0.672 0.035 0.234 0.059
##        4          0.667 0.053 0.216 0.064
##        5          0.669 0.063 0.213 0.055
## 2      1          0.676 0.048 0.229 0.048
##        2          0.700 0.042 0.227 0.032
##        3          0.753 0.046 0.161 0.039
##        4          0.756 0.054 0.150 0.041
##        5          0.712 0.075 0.164 0.049
## 3      1          0.765 0.000 0.176 0.059
##        2          0.654 0.058 0.288 0.000
##        3          0.732 0.061 0.176 0.031
##        4          0.734 0.054 0.180 0.032
##        5          0.722 0.044 0.156 0.078
```

The results of the first part is very difficult to interpret because of the output. From the f-table and proportion version, it looks like the values among sub-groups are very similar, which may mean that the explanatory variables may not have much of an effect on the two symptoms.

Part C

```
mod32 <- glm(Count ~ (HT + PU + smokef + socialf)^3,
             family = poisson(link = "log"),
             data = data32)

mod32
```

```
##
## Call:  glm(formula = Count ~ (HT + PU + smokef + socialf)^3, family = poisson(link = "log"),
##     data = data32)
```

```
##
## Coefficients:
##          (Intercept)                   HTy                   PUy
##             5.656045             -1.249510             -2.612241
##              smokef2               smokef3               socialf2
##            -1.385567             -3.136007              1.005807
##              socialf3              socialf4              socialf5
##              2.401709              0.833597             -0.144586
##               HTy:PUy           HTy:smokef2           HTy:smokef3
##              1.538450              0.133653              0.002299
##          HTy:socialf2          HTy:socialf3          HTy:socialf4
##              0.182377              0.197375              0.110361
##          HTy:socialf5           PUy:smokef2           PUy:smokef3
##              0.062027             -0.166645             -0.468278
##          PUy:socialf2          PUy:socialf3          PUy:socialf4
##             -0.438634             -0.334655              0.028853
##          PUy:socialf5      smokef2:socialf2      smokef3:socialf2
##              0.120632              0.378541              0.039033
##     smokef2:socialf3      smokef3:socialf3      smokef2:socialf4
##              1.068214              1.032585              1.368012
##     smokef3:socialf4      smokef2:socialf5      smokef3:socialf5
##              1.739376              1.647265              1.750667
##      HTy:PUy:smokef2       HTy:PUy:smokef3      HTy:PUy:socialf2
##             -0.199337             -0.576903             -0.232193
##     HTy:PUy:socialf3      HTy:PUy:socialf4      HTy:PUy:socialf5
##              0.023908             -0.126752             -0.256776
## HTy:smokef2:socialf2  HTy:smokef3:socialf2  HTy:smokef2:socialf3
##             -0.217845              0.114783             -0.622208
## HTy:smokef3:socialf3  HTy:smokef2:socialf4  HTy:smokef3:socialf4
##             -0.409354             -0.590119             -0.264277
## HTy:smokef2:socialf5  HTy:smokef3:socialf5  PUy:smokef2:socialf2
##             -0.422739             -0.104025              0.292699
## PUy:smokef3:socialf2  PUy:smokef2:socialf3  PUy:smokef3:socialf3
##              0.464012              0.333231              0.847969
## PUy:smokef2:socialf4  PUy:smokef3:socialf4  PUy:smokef2:socialf5
##              0.152625              0.453546              0.395115
## PUy:smokef3:socialf5
##              0.788634
##
## Degrees of Freedom: 59 Total (i.e. Null);  8 Residual
## Null Deviance:       33140
## Residual Deviance: 12.68     AIC: 458.1
```

```
# get p-value
1 - pchisq(q = 12.68, df = 8)
```

```
## [1] 0.1233444
```

```
car::Anova(mod32)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Count
```

```
##                    LR Chisq Df Pr(>Chisq)
## HT                   3427.3  1  < 2.2e-16 ***
## PU                  10229.3  1  < 2.2e-16 ***
## smokef               6090.7  2  < 2.2e-16 ***
## socialf             12283.8  4  < 2.2e-16 ***
## HT:PU                 497.1  1  < 2.2e-16 ***
## HT:smokef             100.0  2  < 2.2e-16 ***
## HT:socialf              5.6  4    0.22860
## PU:smokef               0.2  2    0.91136
## PU:socialf             26.1  4  3.037e-05 ***
## smokef:socialf        431.0  8  < 2.2e-16 ***
## HT:PU:smokef            6.2  2    0.04569 *
## HT:PU:socialf           3.0  4    0.56332
## HT:smokef:socialf      15.0  8    0.05873 .
## PU:smokef:socialf       3.7  8    0.88377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model's residual deviance is 12.68 with 8 residual degrees of freedom. To see whether the four variable interaction term is needed, we have to calculate the p-value. The p-value is 0.123, telling that there isn't statistically significance in the interaction term, so the interaction isn't needed in the model.

By running an Anova test on the model, we can see that each of the main effects are significant and should be included in the model. There are also several interaction terms that should be included, which are:

- HT:PU
- HT:smokef
- PU:smokef
- smokef:socialf
- HT:PU:smokef

These signficanct interaction terms help us to make conclusions about correlations. We can find that three of the predictors - hypertension, proteinurea, and smoking status - are highly correlated with each other. We can also see that social class is only highly correlated with smoking status.

Question 16

Part E

```
zipMod16e <- zeroinfl(ofp ~ hosp + numchron + gender + school +
                        privins + health_excellent + health_poor |
                        hosp + numchron + gender + school +
                        privins + health_excellent + health_poor,
                      dist = "poisson",
                      data = data16)

lrtest(zipMod16, zipMod16e)
```

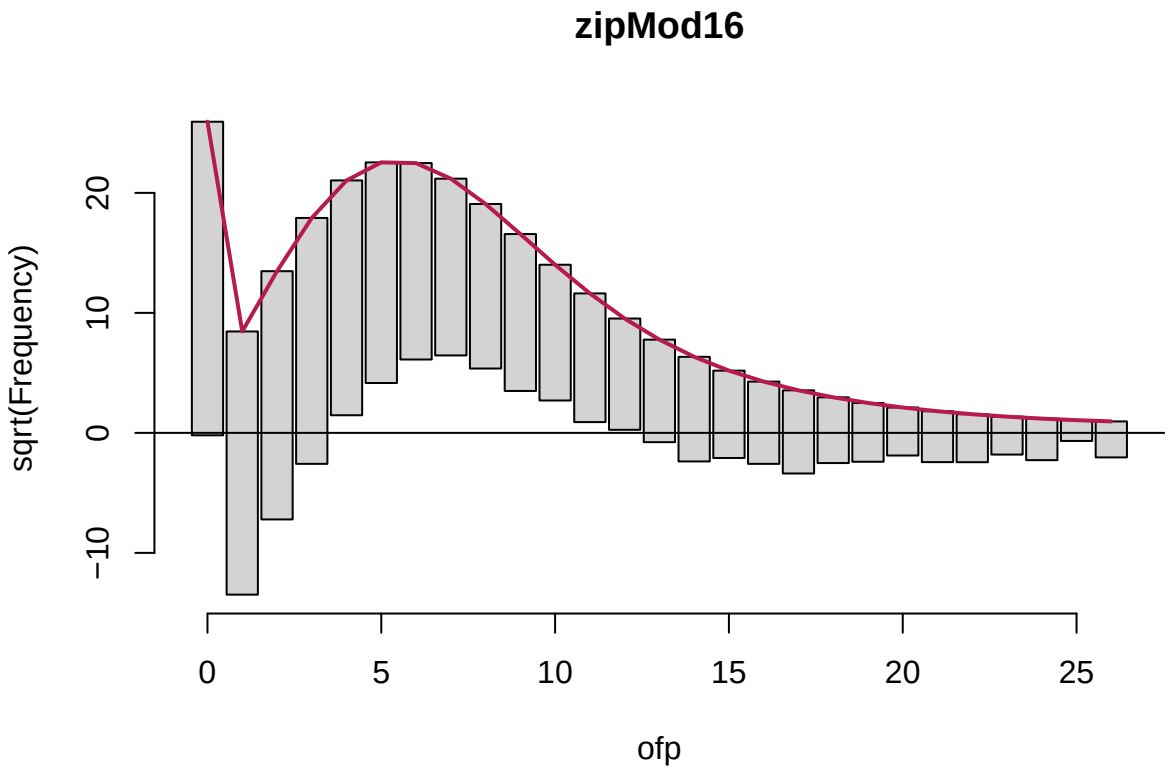
```
## Likelihood ratio test
##
```

```
## Model 1: ofp ~ hosp + numchron + gender + school + privins + health_excellent +
##     health_poor | 1
## Model 2: ofp ~ hosp + numchron + gender + school + privins + health_excellent +
##     health_poor | hosp + numchron + gender + school + privins +
##     health_excellent + health_poor
##   #Df LogLik Df  Chisq Pr(>Chisq)
## 1   9 -16302
## 2  16 -16134  7 335.77  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

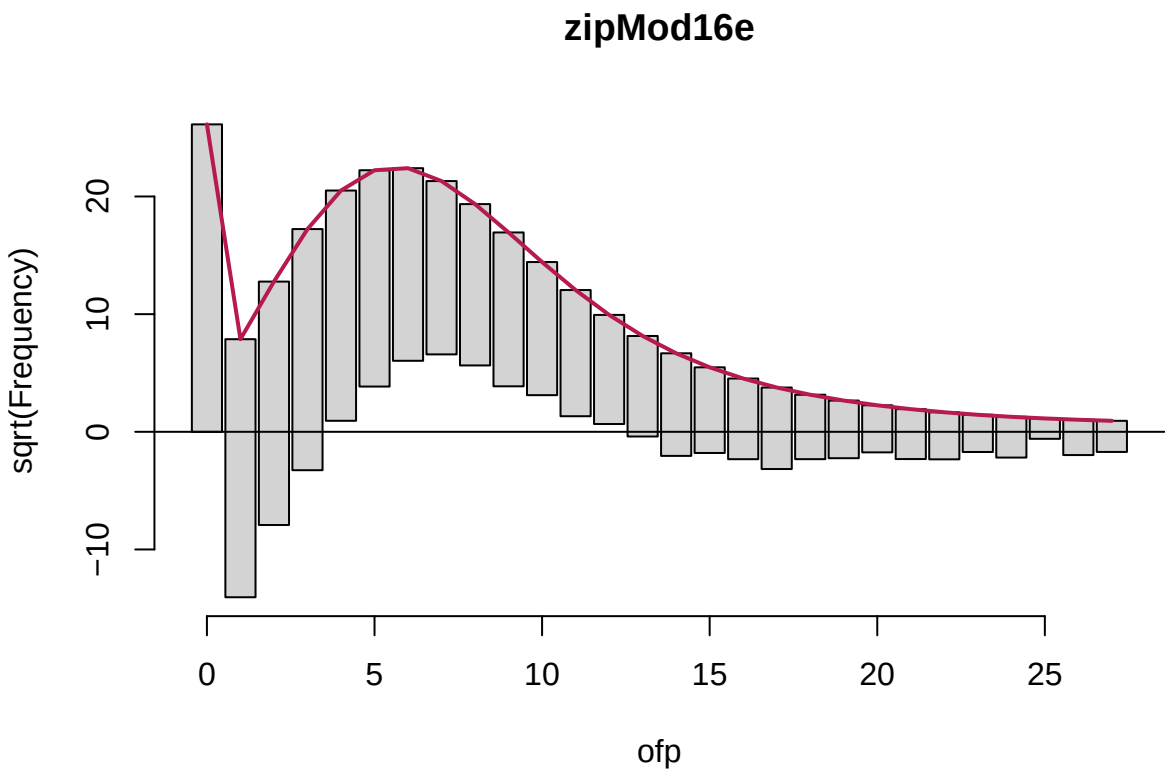
From the likelihood ratio test returning a statistically significant p-value, we can conclude that there's sufficient evidence that the explanatory variables do help to predict the number of physician office visits someone has.

Part F

```
rootogram(zipMod16)
```




```
rootogram(zipMod16e)
```



These rootograms show that both models predict relatively similarly, with the model using explanatory variables to estimate π_i performing slighly better than the base zip model.