高念慈

B082040005

```
require(pscl)      # zero+poisson
```

```
## 載入需要的套件：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
```

```
require(ggplot2)
```

```
## 載入需要的套件：ggplot2
```

```
require(foreign)
```

```
## 載入需要的套件：foreign
```

```
# 讀取由“Minitab”、“S”、“SAS”、“SPSS”、“Stata”、“Systat”、“Weka”、“dBase”...存儲的數據
require(MASS)      # NB
```

```
## 載入需要的套件：MASS
```

資料

- https://drive.google.com/drive/folders/1jss5EZ9lL1_81R4YrKYculaQ0I4BROms (https://drive.google.com/drive/folders/1jss5EZ9lL1_81R4YrKYculaQ0I4BROms)

第三次作業的Dataset還有之前學長的PPT已經上傳在上面的連結，
請各位同學在治療前中後，分別做4個模型poisson, zero inflated, NB, zero+NB
第三次作業要做covariates包含age, gender, income

```
df = read.csv("C:/Users/user/Desktop/regression_note/teeth去連結.csv")
# View(df)

df1 = df[c("age","gender","income","teeth1","teeth2","teeth3","year1","day2","year3")]
summary(df1)
```

```
##       age            gender          income          teeth1       
##  Min.   :20.17   Min.   :0.0000   Min.   :0.0000   Min.   : 0.0000  
##  1st Qu.:41.67   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 0.0000  
##  Median :49.34   Median :1.0000   Median :1.0000   Median : 0.0000  
##  Mean   :49.43   Mean   :0.5194   Mean   :0.9694   Mean   : 0.5543  
##  3rd Qu.:56.07   3rd Qu.:1.0000   3rd Qu.:2.0000   3rd Qu.: 1.0000  
##  Max.   :87.25   Max.   :1.0000   Max.   :2.0000   Max.   :11.0000  
##      teeth2            teeth3           year1             day2      
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 4.800   Min.   :  0.0  
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 6.840   1st Qu.: 14.0  
##  Median : 0.0000   Median : 0.000   Median : 8.420   Median : 32.0  
##  Mean   : 0.4851   Mean   : 1.334   Mean   : 8.475   Mean   : 51.2  
##  3rd Qu.: 0.0000   3rd Qu.: 2.000   3rd Qu.:10.230   3rd Qu.: 77.0  
##  Max.   :14.0000   Max.   :21.000   Max.   :12.000   Max.   :184.0  
##      year3       
##  Min.   : 5.000  
##  1st Qu.: 6.590  
##  Median : 8.335  
##  Mean   : 8.329  
##  3rd Qu.: 9.930  
##  Max.   :11.990
```

治療前資料(刪year1=0)

```
befoedata = subset(df1, select = -c(teeth2,teeth3,day2,year3))
befoedata = befoedata[befoedata["year1"] != 0,]
```

治療中資料(刪day2=0)

```
middata = subset(df1, select = -c(teeth1,teeth3,year1,year3))
middata = middata[middata["day2"] != 0,]
```

治療後資料(刪year3=0)

```
afterdata = subset(df1, select = -c(teeth2,teeth1,day2,year1))
afterdata = afterdata[afterdata["year3"] != 0,]
```

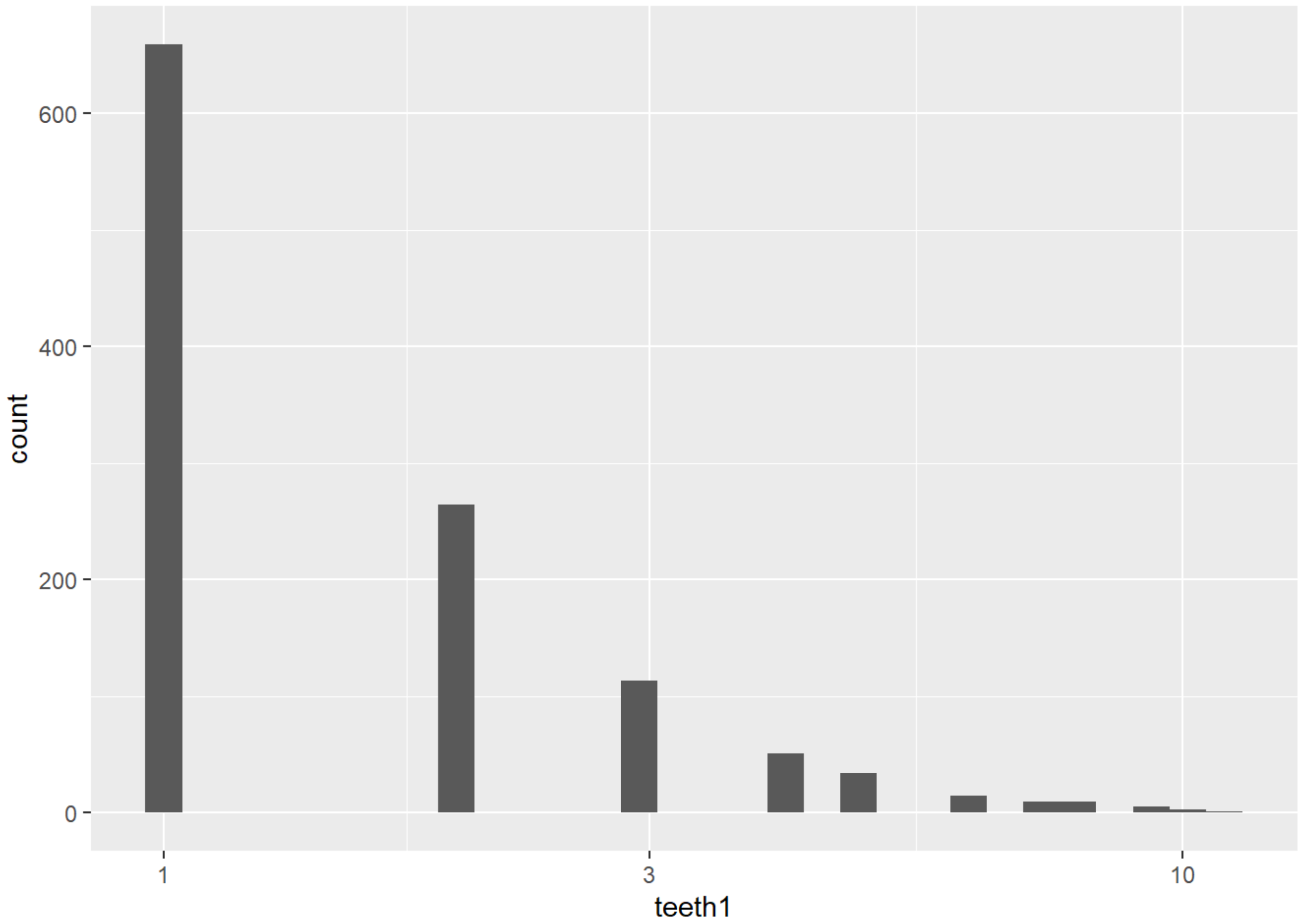
治療前Model (with offset)

```
ggplot(befoedata, aes(teeth1)) +
  geom_histogram() +
  scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2851 rows containing non-finite values (stat_bin).
```



1.poisson

- Source: http://rfunction.com/archives/223 (http://rfunction.com/archives/223)

```
head(befoedata)
```

```
##     age gender income teeth1 year1
## 1 35.72      0      0      0  5.33
## 2 44.11      1      0      1  4.87
## 3 21.94      0      0      0  9.82
## 4 22.66      0      0      0  7.99
## 5 24.46      1      0      0  9.40
## 6 25.19      1      0      0  6.95
```

```
model1 = glm(teeth1 ~ offset(log(year1)) + (age + gender + income), family = poisson, data =
befoedata)

summary(model1)
```

```
##
## Call:
## glm(formula = teeth1 ~ offset(log(year1)) + (age + gender + income),
##     family = poisson, data = befoedata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9273  -1.0598  -0.8720   0.2867   6.7713
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.061237   0.106044 -38.298  < 2e-16 ***
## age          0.023389   0.001781  13.134  < 2e-16 ***
## gender       0.387891   0.043785   8.859  < 2e-16 ***
## income      -0.091468   0.026154  -3.497  0.00047 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 6561.4  on 4015  degrees of freedom
## Residual deviance: 6260.0  on 4012  degrees of freedom
## AIC: 9068.3
##
## Number of Fisher Scoring iterations: 6
```

- 三個變數(age、gender、income)都有顯著影響
- age 每增加一單位(歲)，會導致每年平均拔牙顆數的log 增加 0.023389 個單位(顆)
- gender 每增加一單位，從 0 變 1 (從女生變男生/男生變女生)，會導致每年平均拔牙顆數的log 增加 0.387891 個單位(顆)
- income 每增加一單位(一個等級)，會導致每年平均拔牙顆數的log 增加 -0.091468 個單位(顆)
- AIC: 9068.3

泊松回歸的擬合優度偏差檢驗

- https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/ (https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/)

```
# overall goodness of fit test for Poisson model
# pchisq(model1$deviance, df=model1$df.residual, lower.tail=FALSE)

with(model1, cbind(res.deviance = deviance, df = df.residual,
  p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

```
##      res.deviance   df             p
## [1,]     6259.992 4012 4.364616e-103
```

結果

零假設是我們的模型被正確指定，我們有強有力的證據拒絕該假設。
所以我們有強有力的證據表明我們的模型擬合不佳。
也許是因為此模型沒考慮 zero inflation。

- 但也許我們只是運氣不好——即使原假設為真，檢驗也有 5% 的機率會被拒絕。

2.zero+poisson

- https://stats.oarc.ucla.edu/r/dae/zip/ (https://stats.oarc.ucla.edu/r/dae/zip/)

```
model2 <- zeroinfl(teeth1 ~ offset(log(year1)) + (age + gender + income) | offset(log(year1))
+ (age + gender + income), data = befoedata)

summary(model2)
```

```
##
## Call:
## zeroinfl(formula = teeth1 ~ offset(log(year1)) + (age + gender + income) |
##     offset(log(year1)) + (age + gender + income), data = befoedata)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.9309 -0.5601 -0.4695  0.2531 10.9416
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.204805   0.155911 -14.141  < 2e-16 ***
## age          0.007957   0.002575   3.090  0.00200 **
## gender       0.147111   0.058882   2.498  0.01248 *
## income      -0.109138   0.036500  -2.990  0.00279 **
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.054536   0.252535   0.216    0.829
## age         -0.029908   0.004447  -6.725 1.76e-11 ***
## gender      -0.430645   0.097942  -4.397 1.10e-05 ***
## income      -0.052541   0.060824  -0.864    0.388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 14
## Log-likelihood: -4030 on 8 Df
```

```
AIC(model2)
```

```
## [1] 8075.904
```

- 三個變數(age、gender、income)在 poisson model 都有顯著影響
- age 每增加一單位(歲)，會導致每年平均拔牙顆數的log 增加 0.007957 個單位(顆)
- gender 每增加一單位，從 0 變 1 (從女生變男生/男生變女生)，會導致每年平均拔牙顆數的log 增加 0.147111 個單位(顆)
- income 每增加一單位(一個等級)，會導致每年平均拔牙顆數的log 減少 -0.109138 個單位(顆)

- 兩個變數(age、gender)顯著影響 Zero-inflation model
- age 每增加一單位(歲)，會導致 logit link with 拔牙顆數為 0 的機率增加 -0.029908 個單位(顆)

- gender 每增加一單位，從 0 變 1 (從女生變男生/男生變女生)，會導致 logit link with 拔牙顆數為 0 的機率增加 -0.430645 個單位(顆)
- income 每增加一單位(一個等級)，會導致 logit link with 拔牙顆數為 0 的機率增加 -0.052541 個單位(顆)
- AIC:8075.904
- Log-likelihood: -4030 on 8 Df

Over dispersion:

One of the important assumptions of the Poisson model is equi-dispersion.

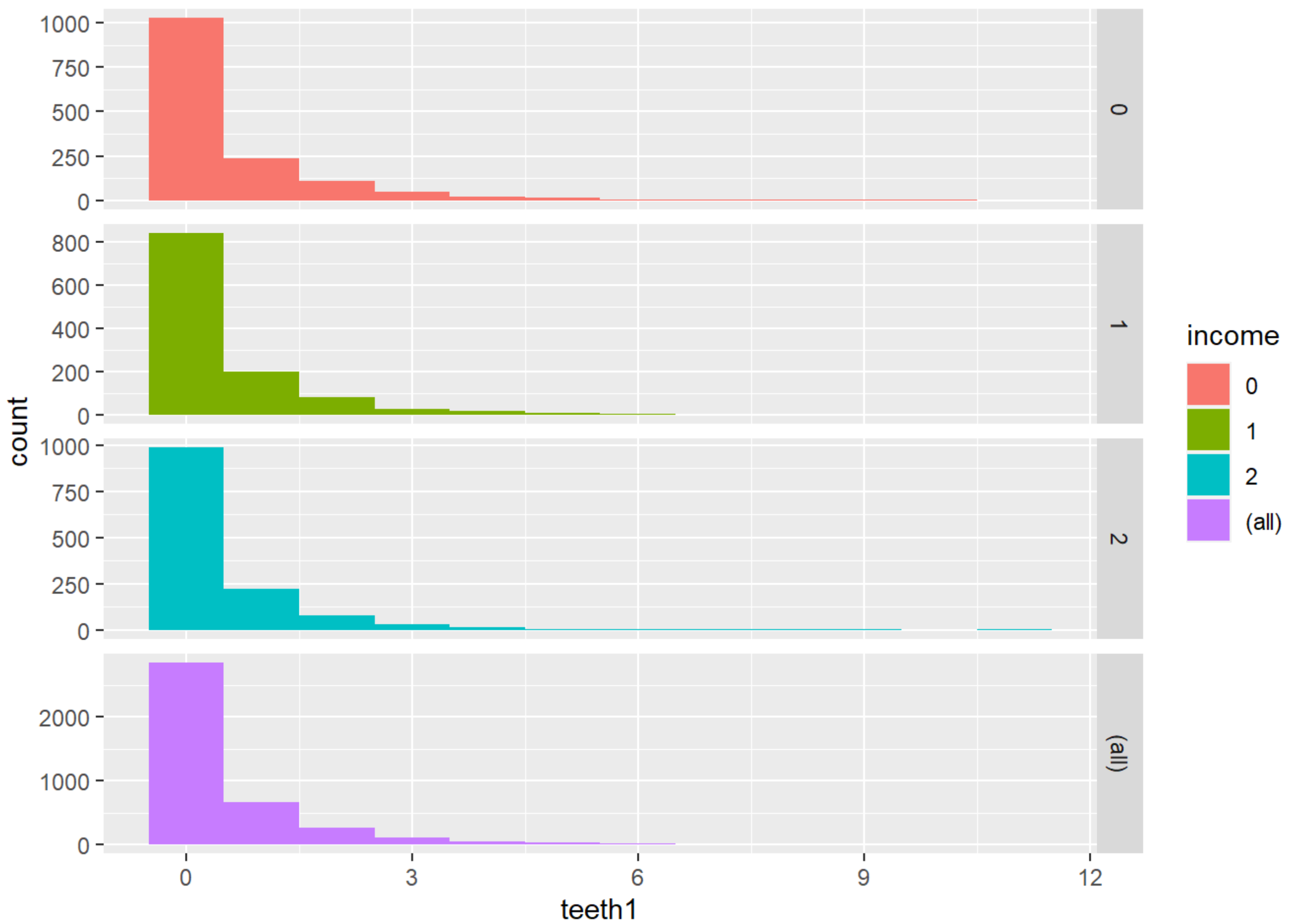
That is,
the mean and variance are equal:

One way to solve the over dispersion problem is to use an alternative distribution for count data.
Negative Binomial regression model:

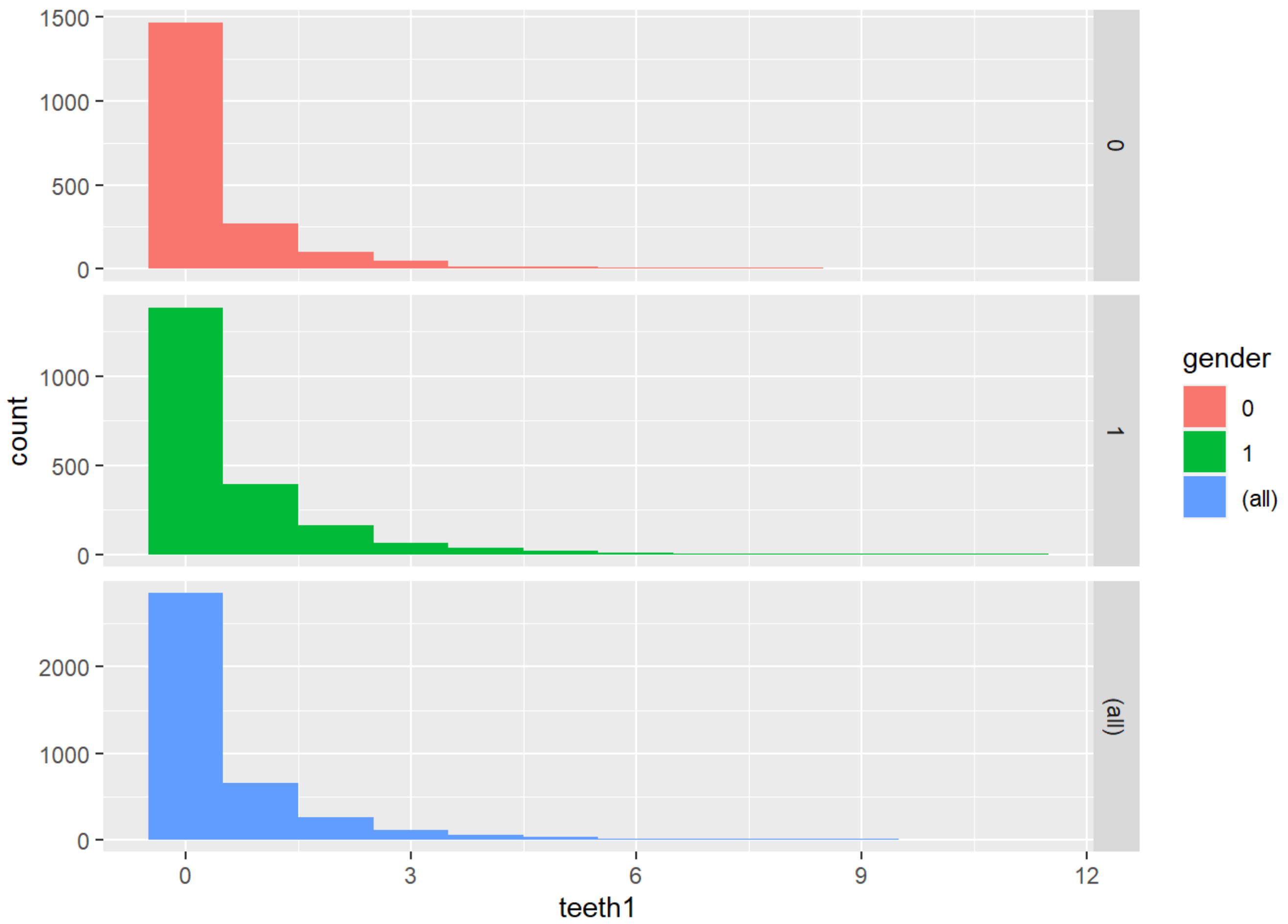
3.NB

- https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/ (https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/)

```
ggplot(befoedata, aes(teeth1, fill = income)) +
  geom_histogram(binwidth=1) +
  facet_grid(income ~ ., margins=TRUE, scales="free")
```



```
ggplot(befoedata, aes(teeth1, fill = gender)) +
  geom_histogram(binwidth=1) +
  facet_grid(gender ~ ., margins=TRUE, scales="free")
```




```
with(befoedata, tapply(teeth1, income, function(x) {
  sprintf("M (SD^2) = %1.2f (%1.2f)", mean(x), (sd(x))^2)
}))
```

```
##                        0                        1                        2
## "M (SD^2) = 0.61 (1.65)" "M (SD^2) = 0.55 (1.31)" "M (SD^2) = 0.49 (1.23)"
```

```
with(befoedata, tapply(teeth1, gender, function(x) {
  sprintf("M (SD^2) = %1.2f (%1.2f)", mean(x), (sd(x))^2)
}))
```

```
##                        0                        1
## "M (SD^2) = 0.44 (1.10)" "M (SD^2) = 0.66 (1.67)"
```

- 可看到上面變數的變異數都大於平均，使用 poisson 可能會有較大的誤差

```
model3 <- glm.nb(teeth1 ~ offset(log(year1)) + (age + gender + income), data = befoedata)

summary(model3)
```

```
##
## Call:
## glm.nb(formula = teeth1 ~ offset(log(year1)) + (age + gender +
##     income), data = befoedata, init.theta = 0.4210759226, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2370  -0.8490  -0.7311   0.1339   3.4547
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.253931   0.162175 -26.231  < 2e-16 ***
## age          0.027064   0.002839   9.533  < 2e-16 ***
## gender       0.436381   0.066790   6.534 6.42e-11 ***
## income      -0.095396   0.040153  -2.376   0.0175 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.4211) family taken to be 1)
##
##     Null deviance: 2966.4  on 4015  degrees of freedom
## Residual deviance: 2826.2  on 4012  degrees of freedom
## AIC: 7782.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.4211
##           Std. Err.:  0.0247
##
##  2 x log-likelihood:  -7772.7250
```

- 三個變數(age、gender、income)都有顯著影響
- age 每增加一單位(歲)，會導致每年平均拔牙顆數的log 增加 0.027064 個單位(顆)
- gender 每增加一單位，從 0 變 1 (從女生變男生/男生變女生)，會導致每年平均拔牙顆數的log 增加 0.436381 個單位(顆)
- income 每增加一單位(一個等級)，會導致每年平均拔牙顆數的log 增加 -0.095396 個單位(顆)
- AIC: 7782.7

Checking goodness of fit for Poisson regression model

```
X2 <- 2 * (logLik(model1) - logLik(model3))
X2
```

```
## 'log Lik.' -1287.566 (df=4)
```

```
pchisq(X2, df = 1, lower.tail=FALSE)
```

```
## 'log Lik.' 1 (df=4)
```

4.zero+NB

- https://www.rdocumentation.org/packages/pscl/versions/1.5.5/topics/zeroinfl (https://www.rdocumentation.org/packages/pscl/versions/1.5.5/topics/zeroinfl)

```
model4 <- zeroinfl(teeth1 ~ offset(log(year1)) + (age + gender + income) | offset(log(year1))
+ (age + gender + income), data = befoedata, dist = "negbin")

summary(model4)
```

```
##
## Call:
## zeroinfl(formula = teeth1 ~ offset(log(year1)) + (age + gender + income) |
##     offset(log(year1)) + (age + gender + income), data = befoedata, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.6241 -0.5072 -0.4365  0.1592 10.3723
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.226068   0.282895 -11.404  < 2e-16 ***
## age          0.011834   0.004072   2.906 0.003660 **
## gender       0.322695   0.096157   3.356 0.000791 ***
## income      -0.120912   0.048992  -2.468 0.013587 *
## Log(theta)  -0.652599   0.114126  -5.718 1.08e-08 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.19832    1.25346   1.754  0.07946 .
## age         -0.12302    0.03887  -3.165  0.00155 **
## gender      -0.84159    0.37530  -2.242  0.02493 *
## income      -0.06978    0.25076  -0.278  0.78081
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.5207
## Number of iterations in BFGS optimization: 35
## Log-likelihood: -3868 on 9 Df
```

```
AIC(model4)
```

```
## [1] 7754.268
```

- 主要的三個變數(age、gender、income)在 poisson model 都有顯著影響
- age 每增加一單位(歲)，會導致每年平均拔牙顆數的log 增加 0.011834 個單位(顆)
- gender 每增加一單位，從 0 變 1 (從女生變男生/男生變女生)，會導致每年平均拔牙顆數的log 增加 0.322695 個單位(顆)
- income 每增加一單位(一個等級)，會導致每年平均拔牙顆數的log 減少 -0.120912 個單位(顆)

- 兩個變數(age、gender)顯著影響 Zero-inflation model
- age 每增加一單位(歲)，會導致 logit link with 拔牙顆數為 0 的機率增加 -0.12302 個單位(顆)
- gender 每增加一單位，從 0 變 1 (從女生變男生/男生變女生)，會導致 logit link with 拔牙顆數為 0 的機率增加 -0.84159 個單位(顆)
- income 每增加一單位(一個等級)，會導致 logit link with 拔牙顆數為 0 的機率增加 -0.06978 個單位(顆)
- AIC: 7754.268
- Log-likelihood: -3868 on 9 Df

- https://wangcc.me/LSHTMlearningnote/count-outcomes.html (https://wangcc.me/LSHTMlearningnote/count-outcomes.html)

負二項式分佈迴歸的結果最底下出現的 Theta 部分，它的倒數是個體的隨機效應部分a
它是關鍵的離散程度參數 (dispersion parameter)

Vuong non-nested hypothesis testing to compare different models

```
# 基於對兩個不嵌套模型的預測概率的比較

vuong(model1, model2) # 普通泊松 vs 零膨脹泊鬆         AIC: 9068.3/ AIC: 8075.904
```

```
## 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                   -10.56709 model2 > model1 < 2.22e-16
## AIC-corrected         -10.48258 model2 > model1 < 2.22e-16
## BIC-corrected         -10.21648 model2 > model1 < 2.22e-16
```

```
vuong(model3, model4) # 普通負二項式與零膨脹負二項式   AIC: 7782.7/ AIC: 7754.268
```

```
## 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                  -3.0832857 model2 > model1 0.0010236
## AIC-corrected        -2.4067110 model2 > model1 0.0080485
## BIC-corrected        -0.2761629 model2 > model1 0.3912115
```

```
# model2: Log-likelihood: -4030 on 8 Df
# model4: Log-likelihood: -3868 on 9 Df
```

結果

- 第一個結果顯示 普通泊松 < 零膨脹泊鬆(好)

- 第二個結果顯示 普通負二項式 < 零膨脹負二項式(好)
- 負二項式分佈迴歸的模型更加擬合數據
- 由 AIC: 9068.3/ AIC: 8075.904/ AIC: 7782.7/ AIC: 7754.268 也能得出 零膨脹負二項式較好

以下變數解釋方式皆相同，直接最後模型比較

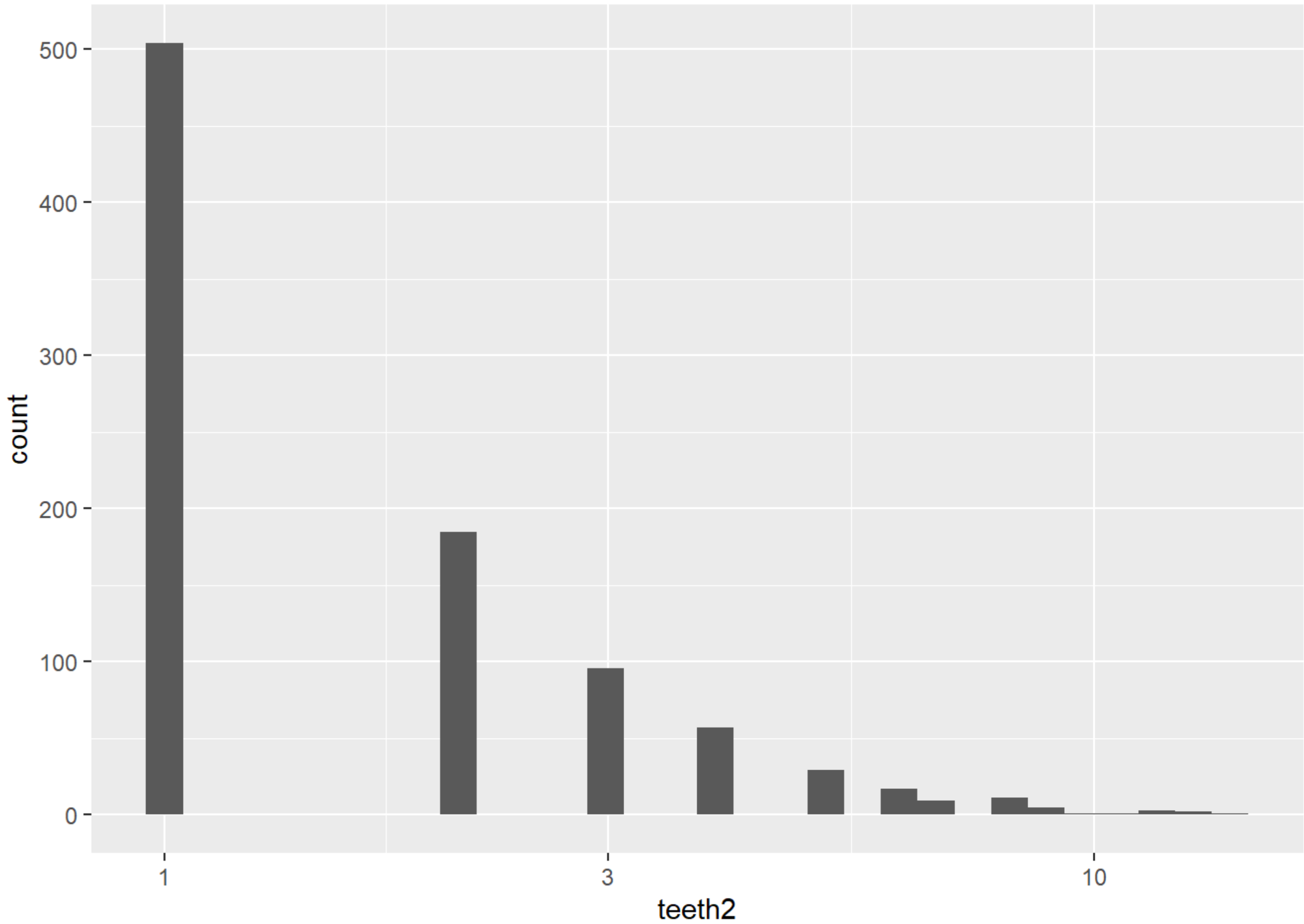
治療中Model(with offset)

```
ggplot(middata, aes(teeth2)) +
  geom_histogram() +
  scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2701 rows containing non-finite values (stat_bin).
```



1.poisson

```
head(middata)
```

```
##     age gender income teeth2 day2
## 1 35.72      0      0      0   32
## 2 44.11      1      0      0   71
## 3 21.94      0      0      0   56
## 4 22.66      0      0      0   22
## 5 24.46      1      0      2   59
## 6 25.19      1      0      0   42
```

```
model1 = glm(teeth2 ~ offset(log(day2)) + (age + gender + income), family = poisson, data = m
iddata)

summary(model1)
```

```
##
## Call:
## glm(formula = teeth2 ~ offset(log(day2)) + (age + gender + income),
##     family = poisson, data = middata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3473  -0.9783  -0.6014  -0.2716   6.6100
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.400162   0.114833 -47.026  < 2e-16 ***
## age          0.012295   0.002072   5.935 2.93e-09 ***
## gender       0.347394   0.046985   7.394 1.43e-13 ***
## income      -0.065893   0.027613  -2.386    0.017 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 5866.2  on 3621  degrees of freedom
## Residual deviance: 5761.7  on 3618  degrees of freedom
## AIC: 8030.3
##
## Number of Fisher Scoring iterations: 6
```

泊松回歸的擬合優度偏差檢驗

- https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/ (https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/)

```
# overall goodness of fit test for Poisson model
# pchisq(model1$deviance, df=model1$df.residual, lower.tail=FALSE)

with(model1, cbind(res.deviance = deviance, df = df.residual,
  p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

```
##      res.deviance   df             p
## [1,]     5761.699 3618 1.846925e-102
```

結果

零假設是我們的模型被正確指定，我們有強有力的證據拒絕該假設。
所以我們有強有力的證據表明我們的模型擬合不佳。

- 但也許我們只是運氣不好——即使原假設為真，檢驗也有 5% 的機率會被拒絕。

2.zero+poisson

- https://stats.oarc.ucla.edu/r/dae/zip/ (https://stats.oarc.ucla.edu/r/dae/zip/)

```
model2 <- zeroinfl(teeth2 ~ offset(log(day2)) + (age + gender + income) | offset(log(day2)) +
(age + gender + income), data = middata)

summary(model2)
```

```
##
## Call:
## zeroinfl(formula = teeth2 ~ offset(log(day2)) + (age + gender + income) |
##     offset(log(day2)) + (age + gender + income), data = middata)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -0.69836 -0.60146 -0.49895 -0.05068 18.60809
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.849574   0.164897 -29.410  < 2e-16 ***
## age          0.015031   0.002912   5.161 2.46e-07 ***
## gender       0.337290   0.060004   5.621 1.90e-08 ***
## income      -0.124895   0.035840  -3.485 0.000492 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.960224   0.416355 -11.913   <2e-16 ***
## age          0.015025   0.007396   2.031   0.0422 *
## gender       0.037104   0.148165   0.250   0.8023
## income      -0.211653   0.090931  -2.328   0.0199 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 14
## Log-likelihood: -3620 on 8 Df
```

```
AIC(model2)
```

```
## [1] 7256.85
```

Over dispersion:

One of the important assumptions of the Poisson model is equi-dispersion.

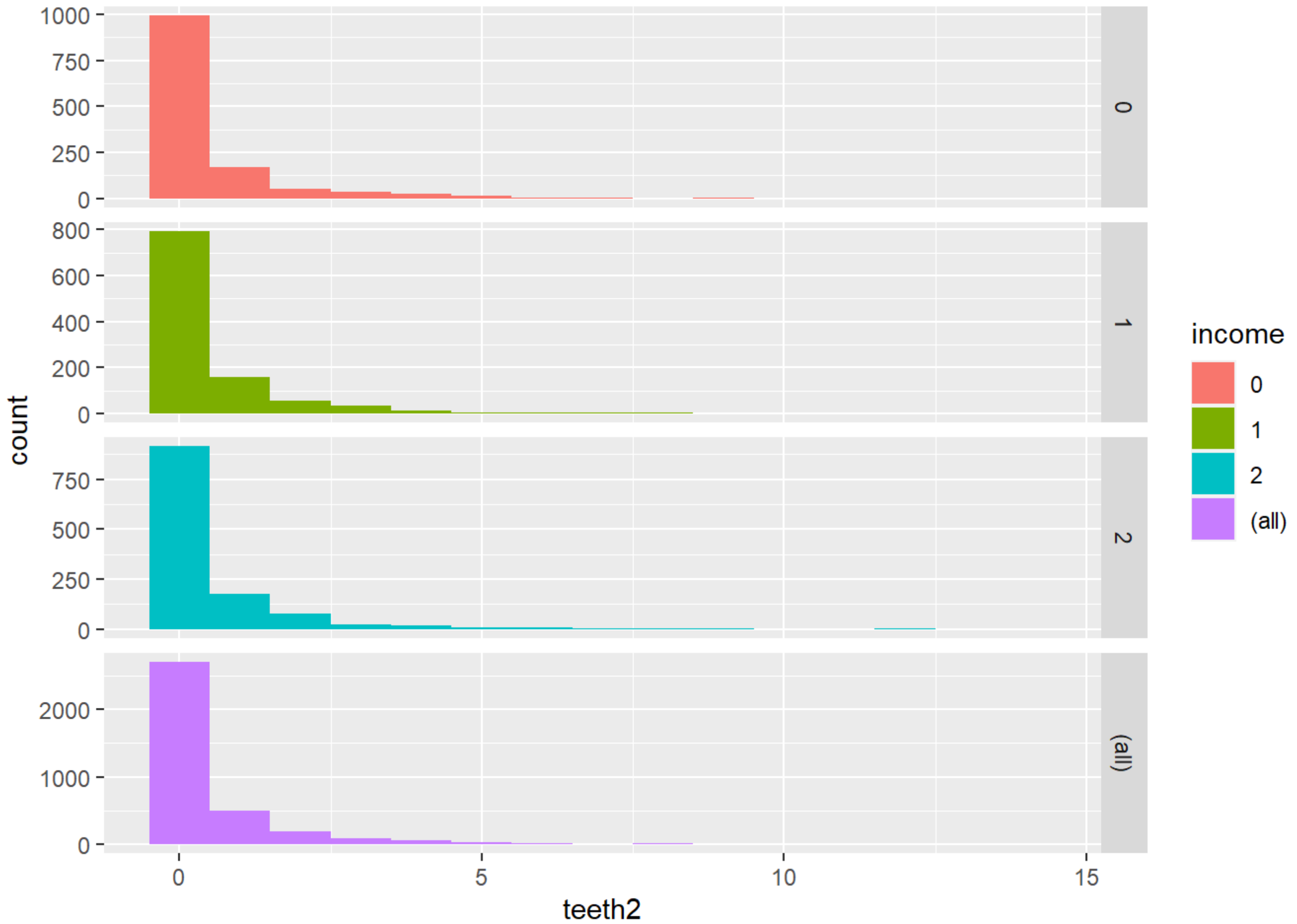
That is,
the mean and variance are equal:

One way to solve the over dispersion problem is to use an alternative distribution for count data.
Negative Binomial regression model:

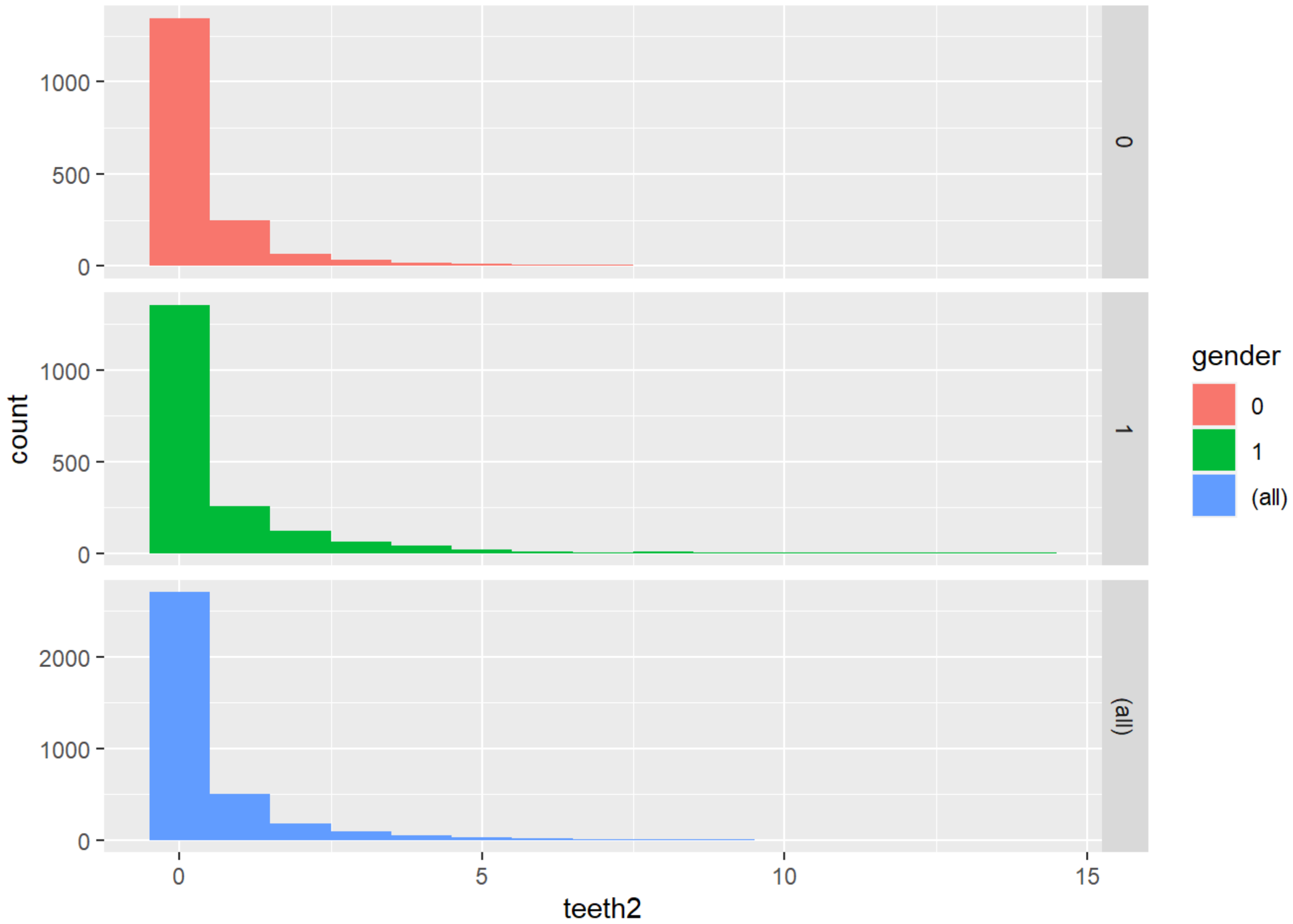
3.NB

- https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/ (https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/)

```
ggplot(middata, aes(teeth2, fill = income)) +
  geom_histogram(binwidth=1) +
  facet_grid(income ~ ., margins=TRUE, scales="free")
```




```
ggplot(middata, aes(teeth2, fill = gender)) +
  geom_histogram(binwidth=1) +
  facet_grid(gender ~ ., margins=TRUE, scales="free")
```




```
with(middata, tapply(teeth2, income, function(x) {
  sprintf("M (SD^2) = %1.2f (%1.2f)", mean(x), (sd(x))^2)
}))
```

```
##                         0                        1                        2
## "M (SD^2) = 0.52 (1.69)" "M (SD^2) = 0.58 (1.90)" "M (SD^2) = 0.51 (1.47)"
```

```
with(middata, tapply(teeth2, gender, function(x) {
  sprintf("M (SD^2) = %1.2f (%1.2f)", mean(x), (sd(x))^2)
}))
```

```
##                         0                        1
## "M (SD^2) = 0.43 (1.23)" "M (SD^2) = 0.62 (2.07)"
```

```
model3 <- glm.nb(teeth2 ~ offset(log(day2)) + (age + gender + income), data = middata)

summary(model3)
```

```
##
## Call:
## glm.nb(formula = teeth2 ~ offset(log(day2)) + (age + gender +
##     income), data = middata, init.theta = 0.3372043089, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2899  -0.8081  -0.5668  -0.2405   4.0759
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.490339   0.192474 -28.525  < 2e-16 ***
## age          0.016622   0.003483   4.772 1.82e-06 ***
## gender       0.367023   0.079461   4.619 3.86e-06 ***
## income      -0.058444   0.047580  -1.228    0.219
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.3372) family taken to be 1)
##
##     Null deviance: 2334.8  on 3621  degrees of freedom
## Residual deviance: 2291.9  on 3618  degrees of freedom
## AIC: 6530.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.3372
##           Std. Err.:  0.0205
##
##  2 x log-likelihood:  -6520.6640
```

Checking goodness of fit for Poisson regression model

```
X2 <- 2 * (logLik(model1) - logLik(model3))
X2
```

```
## 'log Lik.' -1501.642 (df=4)
```

```
pchisq(X2, df = 1, lower.tail=FALSE)
```

```
## 'log Lik.' 1 (df=4)
```

4.zero+NB

- https://www.rdocumentation.org/packages/pscl/versions/1.5.5/topics/zeroinfl (https://www.rdocumentation.org/packages/pscl/versions/1.5.5/topics/zeroinfl)

```
model4 <- zeroinfl(teeth2 ~ offset(log(day2)) + (age + gender + income) | offset(log(day2)) +
(age + gender + income), data = middata, dist = "negbin")
```

```
## Warning in value[[3L]](cond): 系統計算上是奇異的: 互反條件數 = 2.17872e-37FALSE
```

```
summary(model4)
```

```
##
## Call:
## zeroinfl(formula = teeth2 ~ offset(log(day2)) + (age + gender + income) |
##     offset(log(day2)) + (age + gender + income), data = middata, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.5555 -0.4573 -0.3575 -0.1825 23.4806
##
## Count model coefficients (negbin with log link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.49023         NA      NA       NA
## age          0.01662         NA      NA       NA
## gender       0.36701         NA      NA       NA
## income      -0.05845         NA      NA       NA
## Log(theta)  -1.08708         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.5884         NA      NA       NA
## age         -15.5878         NA      NA       NA
## gender       -0.5043         NA      NA       NA
## income       -0.4051         NA      NA       NA
##
## Theta = 0.3372
## Number of iterations in BFGS optimization: 29
## Log-likelihood: -3260 on 9 Df
```

```
AIC(model4)
```

```
## [1] 6538.664
```

Vuong non-nested hypothesis testing to compare different models

```
# 基於對兩個不嵌套模型的預測概率的比較

vuong(model1, model2) # 普通泊松 vs 零膨脹泊鬆
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishable)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   -8.923464 model2 > model1 < 2.22e-16
## AIC-corrected         -8.832112 model2 > model1 < 2.22e-16
## BIC-corrected         -8.549159 model2 > model1 < 2.22e-16
```

```
vuong(model3, model4) # 普通負二項式與零膨脹負二項式
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishable)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A p-value
## Raw                2.382037e-04 model1 > model2  0.4999
## AIC-corrected      6.536035e+03 model1 > model2  <2e-16
## BIC-corrected      2.678069e+04 model1 > model2  <2e-16
```

結果

- 第一個結果顯示 普通泊松 < 零膨脹泊鬆(好)
- 第二個結果顯示 普通負二項式(好) > 零膨脹負二項式
- 負二項式分佈迴歸的模型更加擬合數據
- 由 AIC: 8030.3/ AIC: 7256.85/ AIC: 6530.7/ AIC: 6538.664 也能得出 普通負二項式較好
- 推測可能是因為整體 0 的數量(500)不像治療前&後那麼極端(660/780)，所以分數差不多

以下變數解釋方式皆相同，直接最後模型比較

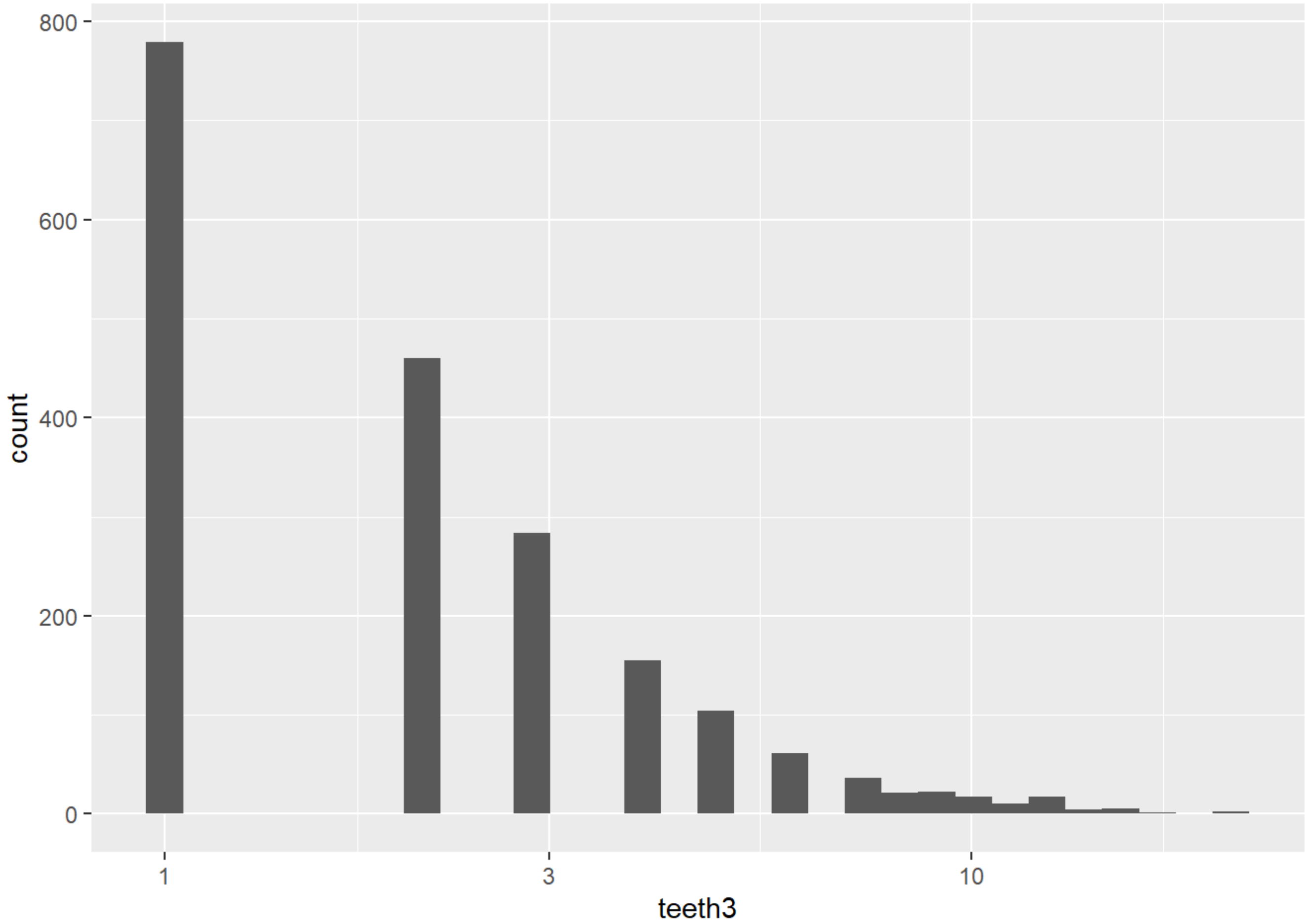
治療後Model(with offset)

```
ggplot(afterdata, aes(teeth3)) +
  geom_histogram() +
  scale_x_log10()
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 2038 rows containing non-finite values (stat_bin).
```



1.poisson

```
head(afterdata)
```

```
##     age gender income teeth3 year3
## 1 35.72      0      0      0 11.58
## 2 44.11      1      0      1 11.93
## 3 21.94      0      0      0  7.03
## 4 22.66      0      0      0  8.94
## 5 24.46      1      0      0  7.44
## 6 25.19      1      0      0  9.93
```

```
model1 = glm(teeth3 ~ offset(log(year3)) + (age + gender + income), family = poisson, data =
afterdata)

summary(model1)
```

```
##
## Call:
## glm(formula = teeth3 ~ offset(log(year3)) + (age + gender + income),
##     family = poisson, data = afterdata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6531  -1.5384  -1.1138   0.4193   9.0211
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.540008   0.065173 -38.974  < 2e-16 ***
## age          0.012691   0.001143  11.103  < 2e-16 ***
## gender       0.281390   0.027978  10.058  < 2e-16 ***
## income      -0.094943   0.016507  -5.752 8.84e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10425  on 4015  degrees of freedom
## Residual deviance: 10132  on 4012  degrees of freedom
## AIC: 15393
##
## Number of Fisher Scoring iterations: 6
```

泊松回歸的擬合優度偏差檢驗

- https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/ (https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/)

```
# overall goodness of fit test for Poisson model
# pchisq(model1$deviance, df=model1$df.residual, lower.tail=FALSE)

with(model1, cbind(res.deviance = deviance, df = df.residual,
  p = pchisq(deviance, df.residual, lower.tail=FALSE)))
```

```
##      res.deviance   df p
## [1,]     10131.73 4012 0
```

結果

零假設是我們的模型被正確指定，我們有強有力的證據拒絕該假設。
所以我們有強有力的證據表明我們的模型擬合不佳。

- 但也許我們只是運氣不好——即使原假設為真，檢驗也有 5% 的機率會被拒絕。

2.zero+poisson

- https://stats.oarc.ucla.edu/r/dae/zip/ (https://stats.oarc.ucla.edu/r/dae/zip/)

```r
model2 <- zeroinfl(teeth3 ~ offset(log(year3)) + (age + gender + income) | offset(log(year3))
+ (age + gender + income), data = afterdata)

summary(model2)
```

```
##
## Call:
## zeroinfl(formula = teeth3 ~ offset(log(year3)) + (age + gender + income) |
##     offset(log(year3)) + (age + gender + income), data = afterdata)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.1637 -0.7844 -0.6044  0.3556 13.7473
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.530386   0.078144 -19.584  < 2e-16 ***
## age          0.005034   0.001367   3.681 0.000232 ***
## gender       0.163296   0.032433   5.035 4.78e-07 ***
## income      -0.109687   0.018765  -5.845 5.06e-09 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.272108   0.188675  -6.742 1.56e-11 ***
## age         -0.019270   0.003490  -5.522 3.36e-08 ***
## gender      -0.317088   0.079475  -3.990 6.61e-05 ***
## income      -0.009144   0.047193  -0.194    0.846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 12
## Log-likelihood: -6754 on 8 Df
```

```r
AIC(model2)
```

```
## [1] 13524.12
```

Over dispersion:

One of the important assumptions of the Poisson model is equi-dispersion.

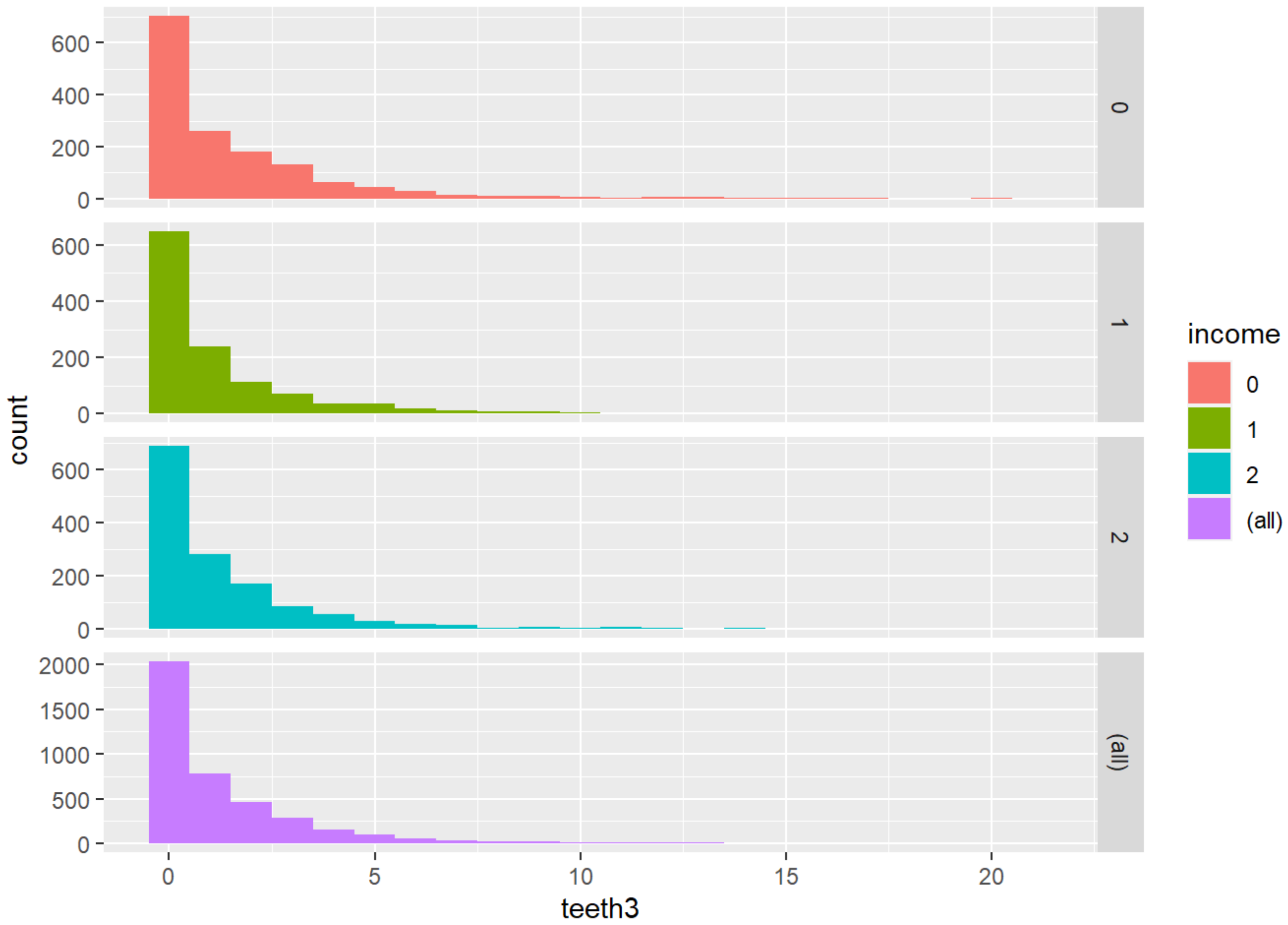
That is,
the mean and variance are equal:

One way to solve the over dispersion problem is to use an alternative distribution for count data.
Negative Binomial regression model:

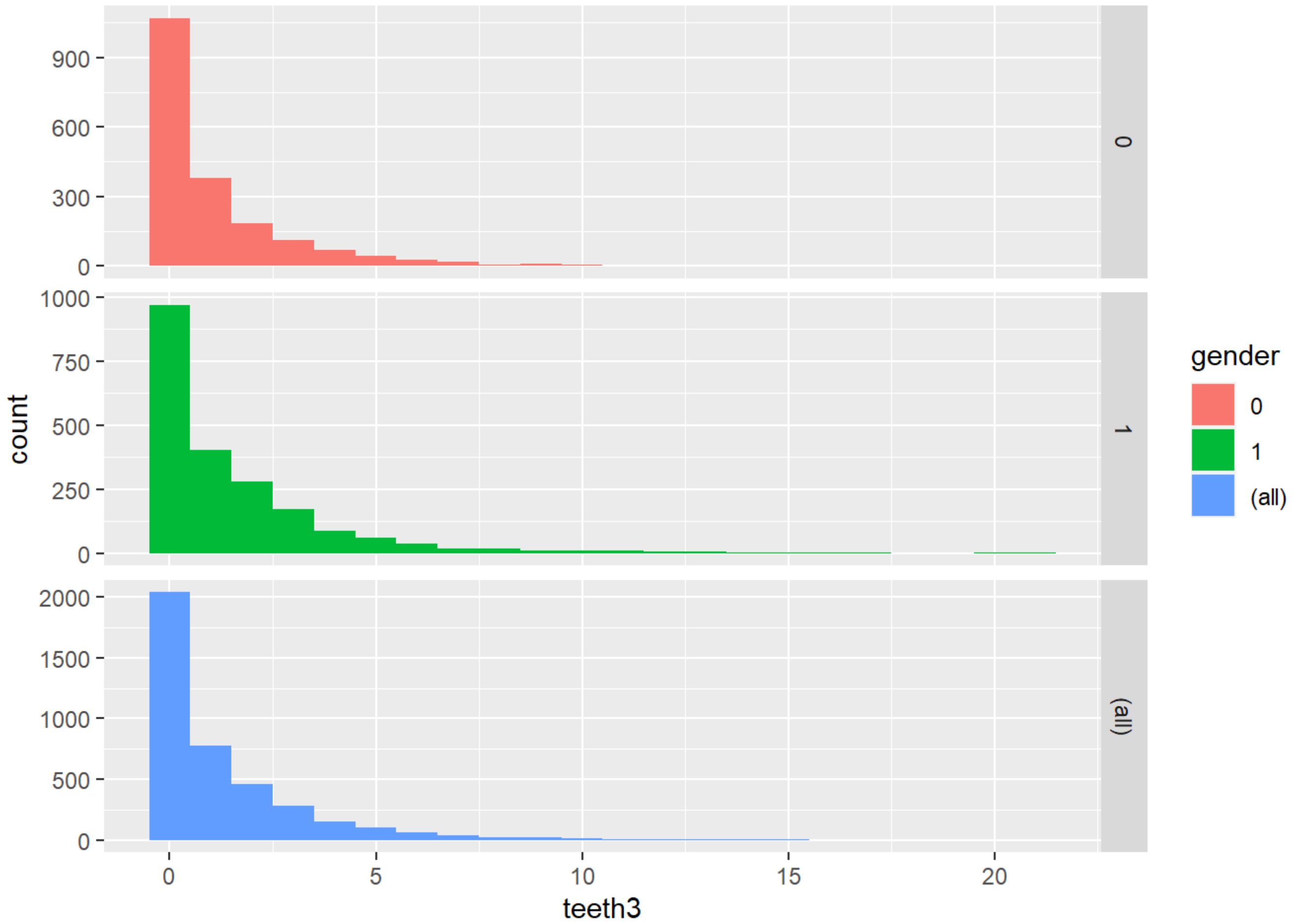
3.NB

- https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/ (https://stats.idre.ucla.edu/r/dae/negative-binomial-regression/)

```
ggplot(afterdata, aes(teeth3, fill = income)) +
  geom_histogram(binwidth=1) +
  facet_grid(income ~ ., margins=TRUE, scales="free")
```




```
ggplot(afterdata, aes(teeth3, fill = gender)) +
  geom_histogram(binwidth=1) +
  facet_grid(gender ~ ., margins=TRUE, scales="free")
```




```
with(afterdata, tapply(teeth3, income, function(x) {
  sprintf("M (SD^2) = %1.2f (%1.2f)", mean(x), (sd(x))^2)
}))
```

```
##                         0                         1                         2
## "M (SD^2) = 1.58 (6.09)" "M (SD^2) = 1.19 (4.21)" "M (SD^2) = 1.20 (3.50)"
```

```
with(afterdata, tapply(teeth3, gender, function(x) {
  sprintf("M (SD^2) = %1.2f (%1.2f)", mean(x), (sd(x))^2)
}))
```

```
##                         0                         1
## "M (SD^2) = 1.13 (3.70)" "M (SD^2) = 1.53 (5.54)"
```

```
model3 <- glm.nb(teeth3 ~ offset(log(year3)) + (age + gender + income), data = afterdata)

summary(model3)
```

```
##
## Call:
## glm.nb(formula = teeth3 ~ offset(log(year3)) + (age + gender +
##     income), data = afterdata, init.theta = 0.6172817126, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5599  -1.1455  -0.9003   0.2305   3.6285
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.646134   0.117508 -22.519  < 2e-16 ***
## age          0.014431   0.002114   6.828 8.62e-12 ***
## gender       0.315857   0.049724   6.352 2.12e-10 ***
## income      -0.097038   0.029747  -3.262  0.00111 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.6173) family taken to be 1)
##
##     Null deviance: 3903.6  on 4015  degrees of freedom
## Residual deviance: 3802.4  on 4012  degrees of freedom
## AIC: 12412
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.6173
##           Std. Err.:  0.0259
##
##  2 x log-likelihood:  -12402.4450
```

Checking goodness of fit for Poisson regression model

```
X2 <- 2 * (logLik(model1) - logLik(model3))
X2
```

```
## 'log Lik.' -2982.317 (df=4)
```

```
pchisq(X2, df = 1, lower.tail=FALSE)
```

```
## 'log Lik.' 1 (df=4)
```

4.zero+NB

- https://www.rdocumentation.org/packages/pscl/versions/1.5.5/topics/zeroinfl (https://www.rdocumentation.org/packages/pscl/versions/1.5.5/topics/zeroinfl)

```
model4 <- zeroinfl(teeth3 ~ offset(log(year3)) + (age + gender + income) | offset(log(year3))
+ (age + gender + income), data = afterdata, dist = "negbin")

summary(model4)
```

```
##
## Call:
## zeroinfl(formula = teeth3 ~ offset(log(year3)) + (age + gender + income) |
##     offset(log(year3)) + (age + gender + income), data = afterdata, dist = "negbin")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7398 -0.6450 -0.4116  0.2525 10.7399
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.282147   0.148537 -15.364  < 2e-16 ***
## age          0.008644   0.002509   3.445 0.000572 ***
## gender       0.277893   0.051547   5.391    7e-08 ***
## income      -0.111318   0.030732  -3.622 0.000292 ***
## Log(theta)  -0.414330   0.048892  -8.474  < 2e-16 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.0153369  1.7249302   2.328 0.019921 *
## age         -0.2238538  0.0613857  -3.647 0.000266 ***
## gender      -1.1224174  0.5708744  -1.966 0.049283 *
## income      -0.0002924  0.3253673  -0.001 0.999283
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.6608
## Number of iterations in BFGS optimization: 32
## Log-likelihood: -6184 on 9 Df
```

```
AIC(model4)
```

```
## [1] 12385.3
```

Vuong non-nested hypothesis testing to compare different models

```
# 基於對兩個不嵌套模型的預測概率的比較

vuong(model1, model2) # 普通泊松 vs 零膨脹泊鬆
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishable)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A    p-value
## Raw                   -14.90796 model2 > model1 < 2.22e-16
## AIC-corrected         -14.84440 model2 > model1 < 2.22e-16
## BIC-corrected         -14.64428 model2 > model1 < 2.22e-16
```

```
vuong(model3, model4) # 普通負二項式與零膨脹負二項式
```

```
## Vuong Non-Nested Hypothesis Test-Statistic:
## (test-statistic is asymptotically distributed N(0,1) under the
##  null that the models are indistinguishable)
## -------------------------------------------------------------
##               Vuong z-statistic             H_A   p-value
## Raw                  -2.9007984 model2 > model1 0.0018611
## AIC-corrected        -2.2404868 model2 > model1 0.0125297
## BIC-corrected        -0.1611519 model2 > model1 0.4359869
```

結果

- 第一個結果顯示 普通泊松 < 零膨脹泊鬆(好)
- 第二個結果顯示 普通負二項式 < 零膨脹負二項式(好)
- 負二項式分佈迴歸的模型更加擬合數據
- 由 AIC: 15393/ AIC: 13524.12/ AIC: 12412/ AIC: 12385.3 也能得出零膨脹負二項式較好

總結果

- 前 : AIC: 9068.3/ AIC: 8075.904/ AIC: 7782.7/ AIC: 7754.268
- 中 : AIC: 8030.3/ AIC: 7256.85 / AIC: 6530.7/ AIC: 6538.664
- 後 : AIC: 15393 / AIC: 13524.12/ AIC: 12412 / AIC: 12385.3

不管哪個模型哪個時間點，
三個變數對平均拔牙顆數都有顯著影響(age增加 、gender增加、income減少)
兩個變數(age減少、gender減少)幾乎顯著影響 Zero-inflation model

- 負二項式分佈迴歸的模型更加擬合數據
