

Poisson Regression model Demo

趙友誠

2024-10-10

Table of contents

資料簡介 DoctorVisits Dataset from package AER	1
建構模型	5
check for overdispersion	7
estimation of dispersion parameter	7
residual analysis	7
dispersion test	8
改用 Quasi_MLE	10

資料簡介 **DoctorVisits Dataset from package AER**

```
library(AER)
library(Hmisc)
library(ggplot2)
library(DataExplorer)
data(DoctorVisits)
```

Table 1: 變數解釋

變數	解釋	資料格式	備註
visits	過去兩週的看醫生 (諮詢) 的次數	num	counts:0~9
gender	性別	factor	1=male,2=female
age	年齡	num	years/100:0.19~0.72
income	年收入 (in 10,000 dollars)	num	income/10000:0.0~1.5
illness	過去兩週不舒服的次數	num	counts:0~5
reduced	過去兩週因生病或受傷的休養天數	num	counts:0~14
health	GHQ-12 心理健康問卷分數 (越低代表 心理狀態越健康)	num	0~12
private	有無私人醫療保險	factor	1=no,2=yes
freepoor	有無政府醫療保險 (低收入)	factor	1=no,2=yes
freerepat	有無政府醫療保險 (退伍軍人、高齡、 失能)	factor	1=no,2=yes
nchronic	有無不影響行動的慢性疾病	factor	1=no,2=yes
lchronic	有無限制行動的慢性疾病	factor	1=no,2=yes

```
str(DoctorVisits)
```

```
'data.frame': 5190 obs. of 12 variables: $ visits : num 1 1 1 1 1 1 1 1 1 1 ... $ gender : Factor w/ 2 levels
"male","female": 2 2 1 1 1 2 2 2 2 1 ... $ age : num 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 0.19 ... $ in-
come : num 0.55 0.45 0.9 0.15 0.45 0.35 0.55 0.15 0.65 0.15 ... $ illness : num 1 1 3 1 2 5 4 3 2 1 ... $ reduced : num
4 2 0 0 5 1 0 0 0 0 ... $ health : num 1 1 0 0 1 9 2 6 5 0 ... $ private : Factor w/ 2 levels "no","yes": 2 2 1 1 1 1 1 1
2 2 ... $ freepoor : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ... $ freerepat: Factor w/ 2 levels "no","yes": 1
1 1 1 1 1 1 1 1 1 1 ... $ nchronic : Factor w/ 2 levels "no","yes": 1 1 1 1 2 2 1 1 1 1 ... $ lchronic : Factor w/ 2 levels
"no","yes": 1 1 1 1 1 1 1 1 1 1 ...
```

```
latex(describe(DoctorVisits),title="",file="", caption="descriptive statistic of the dataset")
```

DoctorVisits														
12 Variables						5190 Observations								
visits														
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95		
5190	0	10	0.489	0.3017	0.5154	0	0	0	0	0	1	2		
Value	0	1	2	3	4	5	6	7	8	9				
Frequency	4141	782	174	30	24	9	12	12	5	1				
Proportion	0.798	0.151	0.034	0.006	0.005	0.002	0.002	0.002	0.001	0.000				
For the frequency table, variable is rounded to the nearest 0														
gender														
n	missing	distinct												
5190	0	2												
Value	male	female												
Frequency	2488	2702												
Proportion	0.479	0.521												
age														
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95		
5190	0	12	0.978	0.4064	0.2258	0.19	0.19	0.22	0.32	0.62	0.72	0.72		
Value	0.19	0.22	0.27	0.32	0.37	0.42	0.47	0.52	0.57	0.62	0.67	0.72		
Frequency	752	1213	523	301	146	126	181	222	273	316	315	822		
Proportion	0.145	0.234	0.101	0.058	0.028	0.024	0.035	0.043	0.053	0.061	0.061	0.158		
For the frequency table, variable is rounded to the nearest 0														
income														
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95		
5190	0	14	0.983	0.5832	0.4085	0.15	0.25	0.25	0.55	0.90	1.10	1.30		
Value	0.00	0.01	0.06	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.90	1.10	1.30	1.50
Frequency	79	35	80	249	1195	462	400	467	455	441	589	361	162	215
Proportion	0.015	0.007	0.015	0.048	0.230	0.089	0.077	0.090	0.088	0.085	0.113	0.070	0.031	0.041
For the frequency table, variable is rounded to the nearest 0														
illness														
n	missing	distinct	Info	Mean	Gmd									
5190	0	6	0.934	1.432	1.481									
Value	0	1	2	3	4	5								
Frequency	1554	1638	946	542	274	236								
Proportion	0.299	0.316	0.182	0.104	0.053	0.045								
For the frequency table, variable is rounded to the nearest 0														
reduced														
n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95		
5190	0	15	0.368	0.8618	1.592	0	0	0	0	0	2	7		
Value	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Frequency	4454	177	108	74	45	40	17	38	17	7	12	2	6	5
Proportion	0.858	0.034	0.021	0.014	0.009	0.008	0.003	0.007	0.003	0.001	0.002	0.000	0.001	0.001
Value	14													
Frequency	188													
Proportion	0.036													
For the frequency table, variable is rounded to the nearest 0														

health

	n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95	
	5190	0	13	0.797	1.218	1.84	0	0	0	0	2	4	6	
Value		0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency		3026	823	446	273	187	132	104	61	42	32	21	24	19
Proportion		0.583	0.159	0.086	0.053	0.036	0.025	0.020	0.012	0.008	0.006	0.004	0.005	0.004

For the frequency table, variable is rounded to the nearest 0

private

	n	missing	distinct
	5190	0	2
Value		no	yes
Frequency		2892	2298
Proportion		0.557	0.443

freepoor

	n	missing	distinct
	5190	0	2
Value		no	yes
Frequency		4968	222
Proportion		0.957	0.043

freerepat

	n	missing	distinct
	5190	0	2
Value		no	yes
Frequency		4099	1091
Proportion		0.79	0.21

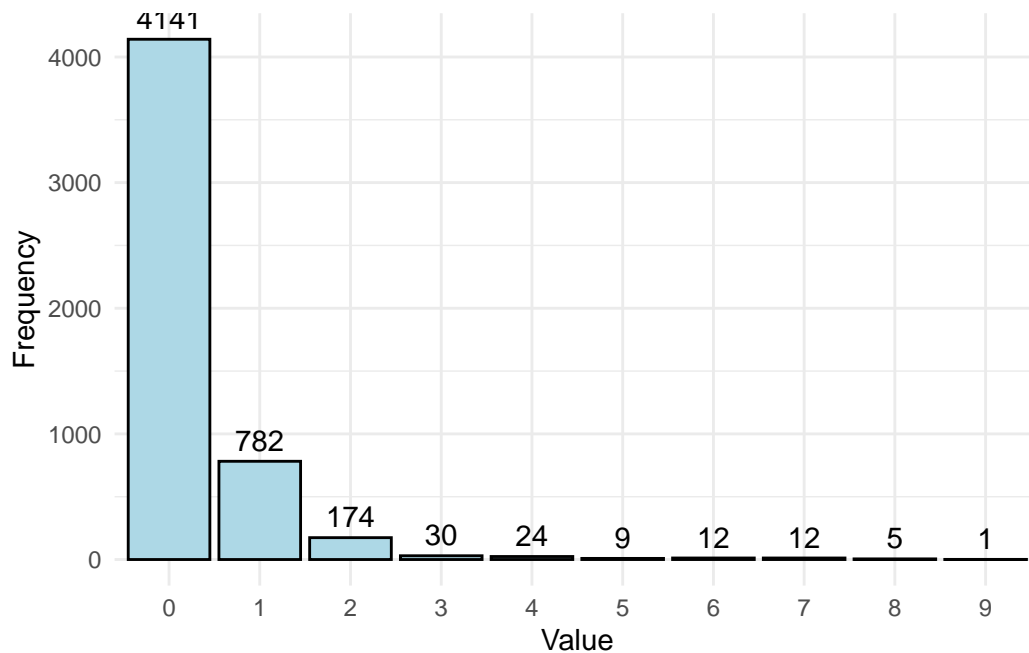
nchronic

	n	missing	distinct
	5190	0	2
Value		no	yes
Frequency		3098	2092
Proportion		0.597	0.403

lchronic

	n	missing	distinct
	5190	0	2
Value		no	yes
Frequency		4585	605
Proportion		0.883	0.117

```
library(ggplot2)
mytable <- data.frame(table(DoctorVisits$visits))
ggplot(mytable, aes(x = factor(Var1), y = Freq)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "black") +
  geom_text(aes(label = Freq), vjust = -0.5, hjust = 0.5) + # 調整 vjust 和 hjust
  labs(x = "Value", y = "Frequency") +
  theme_minimal()
```



觀察變數分布建議使用 `Hmisc::describe()`。

```
sum(DoctorVisits$private=="yes" & DoctorVisits$freepoor=="yes" & DoctorVisits$freerepat=="yes")
```

```
[1] 0
```

```
sum(DoctorVisits$private=="yes" & DoctorVisits$freepoor=="yes")
```

```
[1] 0
```

```
sum(DoctorVisits$private=="yes" & DoctorVisits$freerepat=="yes")
```

```
[1] 0
```

```
sum(DoctorVisits$freepoor=="yes" & DoctorVisits$freerepat=="yes")
```

```
[1] 0
```

```
DoctorVisits$insurance <- as.factor(apply(DoctorVisits[,8:10], MARGIN = 1, function(row){
  return(ifelse(row[1]=="yes", "P",ifelse(row[2]=="yes", "GP", ifelse(row[3]=="yes", "GR", "N"))))
}))
```

```
DoctorVisits <- DoctorVisits[,c(1:7,13,11,12)]
```

```
sum(DoctorVisits$nchronic=="yes" & DoctorVisits$lchronic=="yes")
```

```
[1] 0
```

```
sum(DoctorVisits$lchronic=="yes")
```

```
[1] 605
```

```
sum(DoctorVisits$nchronic=="yes")
```

```
[1] 2092
```

```
sum(DoctorVisits$nchronic=="no" & DoctorVisits$lchronic=="no")
```

```
[1] 2493
```

```
DoctorVisits$chronDis <- as.factor(apply(DoctorVisits[,9:10], MARGIN = 1, function(row){
  return(ifelse(row[1]=="yes", "nch",ifelse(row[2]=="yes", "lch", "N"))))
```

```
)))
DoctorVisits <- DoctorVisits[,c(1:8,11)]
```

將 private, freepoor, freerepat 三個變數合併成 insurance 類別變數:

P=private, GP=freepoor, GR=freerepat, N= 沒有保險

nchronic,lchronic 合併成 chronDis 類別變數:

nch= 有慢性疾病但不限制行動, lch= 有慢性疾病並且會限制行動, N= 沒有慢性疾病

建構模型

```
mydata <- model.matrix(~.*.-1, data = DoctorVisits[, -1])
mydata <- data.frame(visits = DoctorVisits$visits, mydata[, -1])
fullmodel <- glm(visits~., data=mydata, family = poisson())
nullmodel <- glm(visits~1, data=mydata, family = poisson())
fit.step <- step(
  nullmodel,
  scope = list(lower = nullmodel, upper=fullmodel),
  direction = "both", k = log(5190), trace = FALSE)
cat("The number of variables selected is:", length(fit.step$coefficients)-1)
```

The number of variables selected is: 9

```
summary(fit.step)
```

Call:

```
glm(formula = visits ~ reduced + illness + illness.reduced +
    age + age.reduced + age.health + reduced.health + genderfemale +
    genderfemale.age, family = poisson(), data = mydata)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-2.733401	0.106077	-25.768	< 2e-16 ***
reduced	0.243237	0.012260	19.839	< 2e-16 ***
illness	0.243573	0.021086	11.552	< 2e-16 ***
illness.reduced	-0.013430	0.002913	-4.610	4.03e-06 ***
age	1.428861	0.225245	6.344	2.24e-10 ***
age.reduced	-0.130323	0.021866	-5.960	2.52e-09 ***
age.health	0.128120	0.024048	5.328	9.95e-08 ***
reduced.health	-0.005169	0.001449	-3.569	0.000359 ***
genderfemale	0.601664	0.127675	4.712	2.45e-06 ***
genderfemale.age	-0.920576	0.255766	-3.599	0.000319 ***

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: 4295.5 on 5180 degrees of freedom
 AIC: 6647.1

Number of Fisher Scoring iterations: 6

Table 2: fitted model

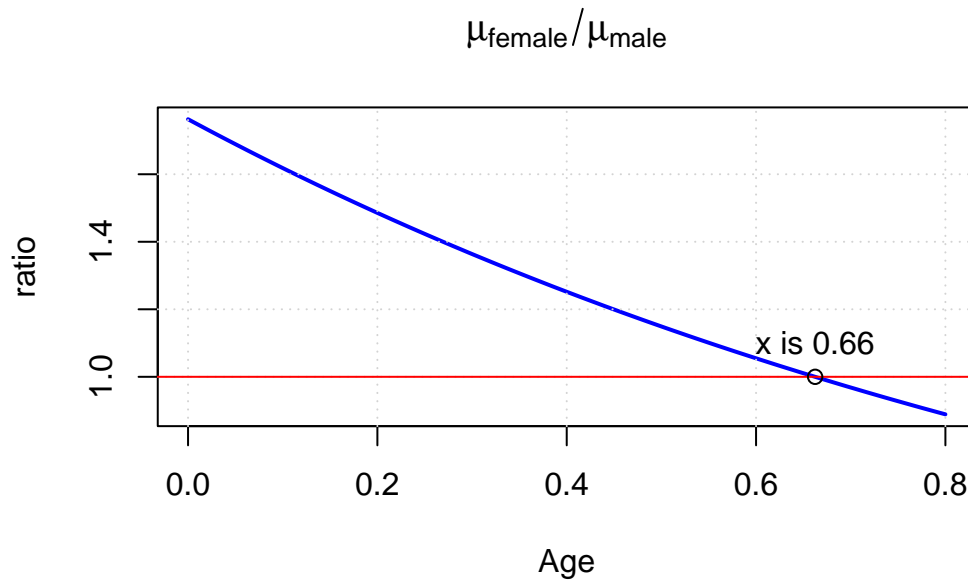
	Coefficients	Std Error	Z value	P value	Significance
(Intercept)	-2.816	0.107	-26.350	0.000	***
illness	0.241	0.021	11.257	0.000	***
genderfemale	0.567	0.128	4.437	0.000	***
age	1.603	0.224	7.146	0.000	***
reduced	0.236	0.012	19.123	0.000	***
health	0.065	0.013	4.988	0.000	***
illness.reduced	-0.013	0.003	-4.506	0.000	***
age.reduced	-0.115	0.021	-5.366	0.000	***
reduced.health	-0.005	0.001	-3.515	0.000	***
genderfemale.age	-0.856	0.256	-3.344	0.001	**

按照慣例做法，將未選中的主效應項加入模型：

發現加入 health 會使得 age.health 與 health 不顯著，因此傾向保留主效應。

```
oldformula <- fit.step$formula
newformula <- update(oldformula, .~.-age.health+health)
fit.step <- glm(newformula, data = mydata, family = poisson())
output <- summary(fit.step)$coefficients
output <- data.frame(
  coef = round(output[,1],3),
  sd = round(output[,2],3),
  z = round(output[,3],3),
  pvalue = round(output[,4],3)
)
output$sig <- sapply(output$pvalue, function(p) {
  ifelse(p < 0.001,"***",ifelse(p < 0.01,"**",ifelse(p < 0.05,"*","")))
})
colnames(output) <- c("Coefficients", "Std Error", "Z value", "P value", "Significance")
output <- output[c(1,3,8,5,2,10,4,6,7,9),]
latex(output, title="",file="",caption = "fitted model")

x <- seq(0, 0.8, by = 0.001)
curve(expr = exp(0.567 - 0.856 * x), from = 0, to = 0.8,
      xlab = "Age", ylab = "ratio",
      col = "blue", lwd = 2, main = expression(mu[female] / mu[male]))
grid()
abline(h=1, col="red")
points(x = 0.567/0.856, y = 1, col="black")
text(x = 0.567/0.856, y = 1.1,paste0("x is ",round(0.567/0.856,2)))
```



check for overdispersion

estimation of dispersion parameter

```
# Deviance Method
dispersion_deviance <- sum(resid(fit.step, type = "deviance")^2)/fit.step$df.residual
cat("estimated dispersion by deviance:", dispersion_deviance)
```

estimated dispersion by deviance: 0.8298426

```
# Pearson Method
dispersion_pearson <- sum(resid(fit.step, type = "pearson")^2)/fit.step$df.residual
cat("estimated dispersion by pearson:", dispersion_pearson)
```

estimated dispersion by pearson: 1.312561

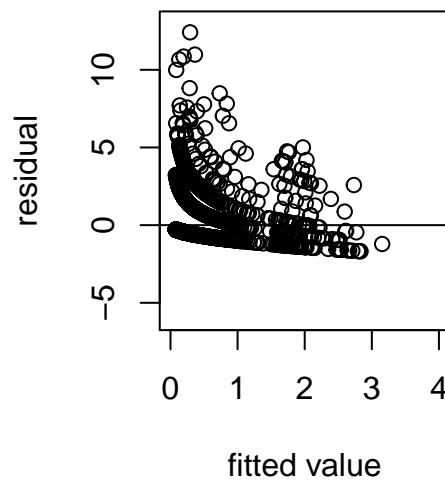
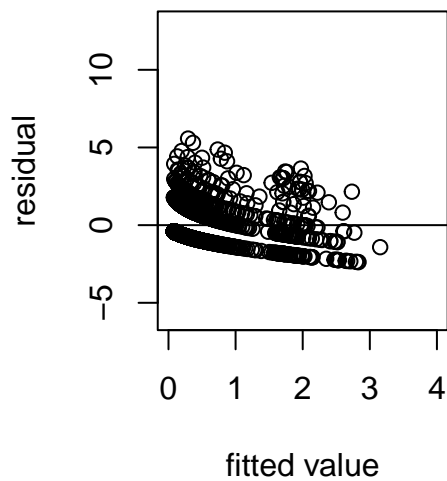
residual analysis

```
# 檢查 pearson residual 的變異數是否有等於 1
cat("Variance of Pearson residual:", var(resid(fit.step, type = "pearson")))
```

Variance of Pearson residual: 1.310178

```
# 殘差圖 (記得依 fitted value 大小排序)
par(mfrow = c(1, 2))
plot(sort(fit.step$fitted.values),
      resid(fit.step, type = "deviance")[order(fit.step$fitted.values, decreasing = FALSE)],
      main="Fig 2.1 Deviance residual plot", ylab = "residual", xlab="fitted value",
      ylim = c(-6,13), xlim=c(0,4))
abline(0,0)
plot(sort(fit.step$fitted.values),
      resid(fit.step, type = "pearson")[order(fit.step$fitted.values, decreasing = FALSE)],
      main="Fig 2.2 Pearson residual plot", ylab = "residual", xlab="fitted value",
      ylim = c(-6,13), xlim=c(0,4))
abline(0,0)
```

Fig 2.1 Deviance residual plot | **Fig 2.2 Pearson residual plot**



dispersion test

```
dispersiontest(fit.step, alternative = "greater",
               trafo = 1)
```

Overdispersion test

```
data: fit.step
z = 6.6542, p-value = 1.424e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
alpha
0.3801358
```

```
dispersiontest(fit.step, alternative = "greater",
               trafo = function(mu) mu)
```

Overdispersion test

```
data: fit.step
z = 6.6542, p-value = 1.424e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
alpha
0.3801358
```

```
dispersiontest(fit.step, alternative = "greater",
               trafo = 2)
```

Overdispersion test

```
data: fit.step
z = 8.4562, p-value < 2.2e-16
alternative hypothesis: true alpha is greater than 0
```


sample estimates:

alpha

1.008399

```
dispersiontest(fit.step, alternative = "greater",  
               trafo = function(mu) mu^2)
```

Overdispersion test

data: fit.step

z = 8.4562, p-value < 2.2e-16

alternative hypothesis: true alpha is greater than 0

sample estimates:

alpha

1.008399

```
dispersiontest(fit.step, alternative = "greater")
```

Overdispersion test

data: fit.step

z = 6.6542, p-value = 1.424e-11

alternative hypothesis: true dispersion is greater than 1

sample estimates:

dispersion

1.380136

```
dispersiontest(fit.step, alternative = "two.sided")
```

Dispersion test

data: fit.step

z = 6.6542, p-value = 2.848e-11

alternative hypothesis: true dispersion is not equal to 1

sample estimates:

dispersion

1.380136

```
dispersiontest(fit.step, alternative = "less")
```

Underdispersion test

data: fit.step

z = 6.6542, p-value = 1

alternative hypothesis: true dispersion is less than 1

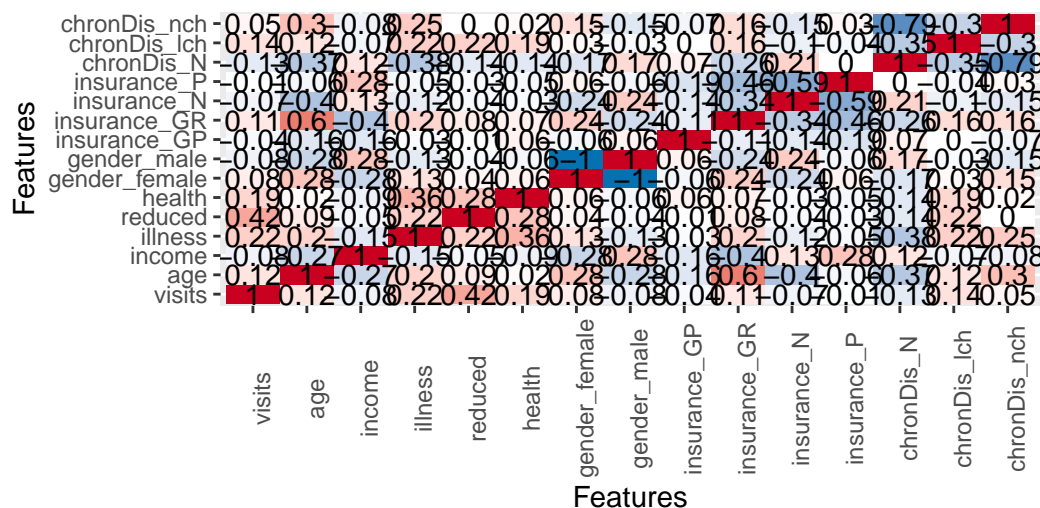
sample estimates:

dispersion

1.380136

改用 Quasi_MLE

```
plot_correlation(DoctorVisits)
```



```
fit.quasipoi <- glm(data = DoctorVisits, formula = visits ~ ., family = quasipoisson())
summary(fit.quasipoi)
```

Call:

```
glm(formula = visits ~ ., family = quasipoisson(), data = DoctorVisits)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-2.385566	0.224281	-10.637	< 2e-16 ***
genderfemale	0.156490	0.064683	2.419	0.01558 *
age	0.279123	0.191244	1.460	0.14448
income	-0.187416	0.098488	-1.903	0.05711 .
illness	0.186156	0.021043	8.847	< 2e-16 ***
reduced	0.126690	0.005796	21.857	< 2e-16 ***
health	0.030683	0.011607	2.644	0.00823 **
insuranceGR	0.522102	0.213801	2.442	0.01464 *
insuranceN	0.438462	0.207164	2.116	0.03435 *
insuranceP	0.564960	0.206266	2.739	0.00618 **
chronDisN	-0.150717	0.094780	-1.590	0.11186
chronDisnch	-0.033417	0.078897	-0.424	0.67191

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.327571)

Null deviance: 5634.8 on 5189 degrees of freedom
 Residual deviance: 4380.1 on 5178 degrees of freedom
 AIC: NA

Number of Fisher Scoring iterations: 6

```
dispersion_deviance_quasi <- sum(resid(fit.quasipoi, type = "deviance")^2)/fit.quasipoi$df.residual
dispersion_pearson_quasi <- sum(resid(fit.quasipoi, type = "pearson")^2)/fit.quasipoi$df.residual
cat("estimated dispersion by deviance:", dispersion_deviance_quasi,
    "\nestimated dispersion by pearson:", dispersion_pearson_quasi)
```

estimated dispersion by deviance: 0.8459121

estimated dispersion by pearson: 1.32757