Poisson Regression model Demo

趙友誠

10/14/24

Table of contents

資料簡介 DoctorVisits Dataset from package AER

```
library(AER)
library(Hmisc)
library(ggplot2)
library(DataExplorer)
data(DoctorVisits)
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 ...
 $ income   : 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 ...
 $ 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 ...
```

```
#check how to combine the variables
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$chronDis <- as.factor(apply(DoctorVisits[,9:10], MARGIN = 1, function(row){
  return(ifelse(row[1]=="yes", "nch",ifelse(row[2]=="yes", "lch", "N")))
}))
DoctorVisits$isfemale <- ifelse(DoctorVisits$gender=="female",1,0)
DoctorVisits <- DoctorVisits[,c(1,12,3:8,11)]
head(DoctorVisits)
```

```
  visits isfemale  age income illness reduced health insurance chronDis
1      1        1 0.19   0.55       1       4      1         P        N
2      1        1 0.19   0.45       1       2      1         P        N
3      1        0 0.19   0.90       3       0      0         N        N
4      1        0 0.19   0.15       1       0      0         N        N
5      1        0 0.19   0.45       2       5      1         N      nch
6      1        1 0.19   0.35       5       1      9         N      nch
```

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

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
isfemale	性別	num	1,0
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	num	mentally (healthy)0~12(unhealthy)
insurance	醫療保險種類	factor	P: 私人, GP: 政府低收, GR: 政府高齡與其他, N: 沒有保險
chronDis	慢性疾病種類	factor	nch: 不限制行動, lch: 限制行動, N: 沒有慢性疾病

```
latex(describe(DoctorVisits),title="",file="")
```

DoctorVisits
9 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

isfemale

n	missing	distinct	Info	Sum	Mean	Gmd
5190	0	2	0.749	2702	0.5206	0.4992

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      1.840          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
```

insurance

```
     n    missing    distinct
  5190          0          4

Value         GP    GR     N     P
Frequency    222  1091  1579  2298
Proportion 0.043 0.210 0.304 0.443
```

chronDis

```
     n    missing    distinct
  5190          0          3

Value        lch     N   nch
Frequency    605  2493  2092
Proportion 0.117 0.480 0.403
```

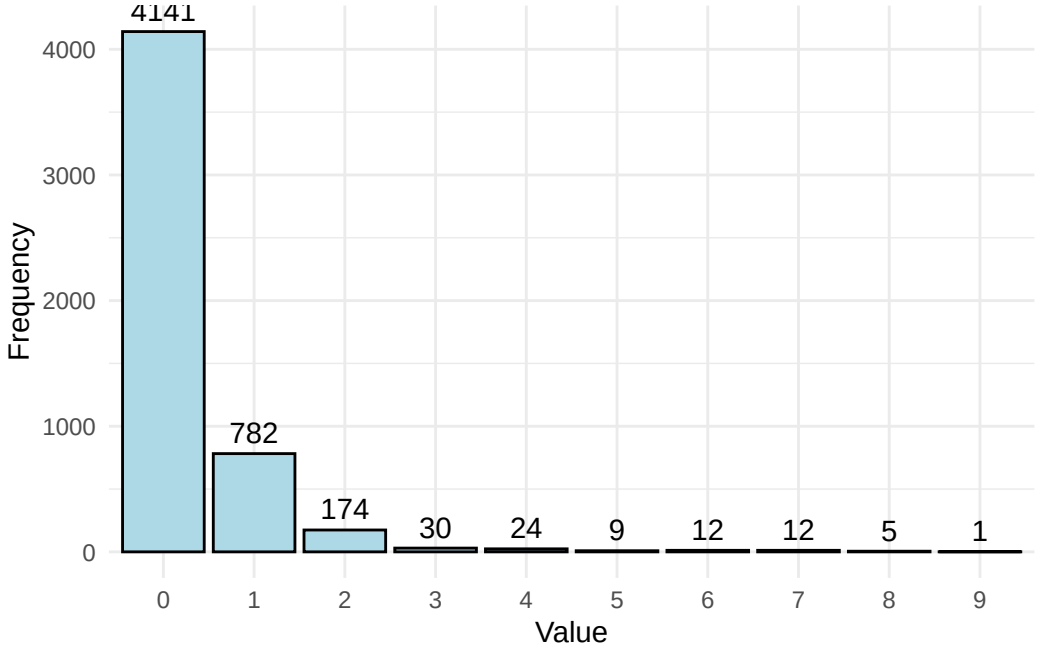
```r
library(ggplot2)
mytable <- data.frame(table(DoctorVisits$visits))
ggplot(mytable, aes(x = factor(Var1), y = Freq)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "black") +
  geom_text(aes(label = Freq), vjust = -0.5, hjust = 0.5) +  # 調整 vjust 和 hjust
  labs(x = "Value", y = "Frequency")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

建構模型

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

```
The number of variables selected is: 9
```

```r
summary(myGLM)
```




```
Call:
glm(formula = visits ~ reduced + illness + illness.reduced +
    age + age.reduced + age.health + income.insuranceN + insuranceGP +
    reduced.health, family = poisson(), data = mydata)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.246650   0.087223 -25.757  < 2e-16 ***
reduced            0.248641   0.012224  20.341  < 2e-16 ***
illness            0.247100   0.021081  11.722  < 2e-16 ***
illness.reduced   -0.013566   0.002924  -4.639 3.50e-06 ***
age                0.693581   0.161334   4.299 1.72e-05 ***
age.reduced       -0.138728   0.021995  -6.307 2.84e-10 ***
age.health         0.126539   0.024041   5.264 1.41e-07 ***
income.insuranceN -0.379462   0.094568  -4.013 6.01e-05 ***
insuranceGP       -0.586419   0.174576  -3.359 0.000782 ***
reduced.health    -0.005062   0.001455  -3.480 0.000501 ***
---
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: 4294.3  on 5180  degrees of freedom
AIC: 6645.9

Number of Fisher Scoring iterations: 6
```

因為建立模型是以解釋為目的，所以將未選中的主效應項加入模型:

若加入之後導致不顯著，可能發生共線性，這時傾向保留主項。

```r
oldformula <- myGLM$formula
newformula <- update(oldformula, .~.+health+income+insuranceGP+insuranceN)
myGLM <- glm(newformula, data = mydata, family = poisson())
summary(myGLM)
```

```
Call:
glm(formula = newformula, family = poisson(), data = mydata)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.240148   0.124987 -17.923  < 2e-16 ***
reduced            0.245012   0.012461  19.662  < 2e-16 ***
illness            0.239337   0.021496  11.134  < 2e-16 ***
illness.reduced   -0.013113   0.002934  -4.469 7.86e-06 ***
age                0.779130   0.187681   4.151 3.31e-05 ***
age.reduced       -0.131301   0.022353  -5.874 4.26e-09 ***
age.health         0.065778   0.045767   1.437 0.150650
income.insuranceN -0.439233   0.189245  -2.321 0.020288 *
insuranceGP       -0.612192   0.177855  -3.442 0.000577 ***
reduced.health    -0.005549   0.001504  -3.689 0.000225 ***
health             0.037524   0.024783   1.514 0.130000
income            -0.103220   0.089158  -1.158 0.246982
insuranceN         0.080865   0.134028   0.603 0.546279
---
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: 4289.2  on 5177  degrees of freedom
AIC: 6646.7

Number of Fisher Scoring iterations: 6
```

```r
oldformula <- myGLM$formula
newformula <- update(oldformula, .~.-age.health-income.insuranceN)
myGLM <- glm(newformula, data = mydata, family = poisson())
summary(myGLM)
```

```
Call:
glm(formula = newformula, family = poisson(), data = mydata)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -2.198041   0.114001 -19.281  < 2e-16 ***
reduced          0.239242   0.012244  19.540  < 2e-16 ***
illness          0.239675   0.021454  11.171  < 2e-16 ***
illness.reduced -0.013264   0.002929  -4.528 5.95e-06 ***
age              0.808066   0.169650   4.763 1.91e-06 ***
age.reduced     -0.119195   0.021527  -5.537 3.08e-08 ***
insuranceGP     -0.653518   0.177211  -3.688 0.000226 ***
reduced.health  -0.005386   0.001494  -3.605 0.000312 ***
health           0.067381   0.013026   5.173 2.31e-07 ***
```

```
income          -0.205167   0.079147  -2.592 0.009536 **
insuranceN      -0.189168   0.069931  -2.705 0.006829 **
---
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: 4296.6  on 5179  degrees of freedom
AIC: 6650.2

Number of Fisher Scoring iterations: 6
```

```r
lmtest::lrtest(object = myGLM,
  glm(visits~.-chronDis-isfemale,
      DoctorVisits,
      family=poisson())
)
```

```
Likelihood ratio test

Model 1: visits ~ reduced + illness + illness.reduced + age + age.reduced +
    insuranceGP + reduced.health + health + income + insuranceN
Model 2: visits ~ (isfemale + age + income + illness + reduced + health +
    insurance + chronDis) - chronDis - isfemale
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  11 -3314.1
2   9 -3362.1 -2 95.962  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Goodness-of-fit test 的結果顯示，加入交互項的模型顯著優於僅有主項的模型。

```r
output <- summary(myGLM)$coefficients
output <- data.frame(
  coef = round(output[,1],4),
  sd = round(output[,2],4),
  z = round(output[,3],4),
  pvalue = round(output[,4],4)
)
output$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(1,5,3,10,7,11,2,9,4,6,8),]
latex(output, title="",file="",caption = "fitted model")
```

Table 2: fitted model

	Coefficients	Std Error	Z value	P value	Significance
(Intercept)	−2.1980	0.1140	−19.2809	0.0000	***
age	0.8081	0.1696	4.7631	0.0000	***
illness	0.2397	0.0215	11.1713	0.0000	***
income	−0.2052	0.0791	−2.5922	0.0095	**
insuranceGP	−0.6535	0.1772	−3.6878	0.0002	***
insuranceN	−0.1892	0.0699	−2.7050	0.0068	**
reduced	0.2392	0.0122	19.5398	0.0000	***
health	0.0674	0.0130	5.1728	0.0000	***
illness.reduced	−0.0133	0.0029	−4.5283	0.0000	***
age.reduced	−0.1192	0.0215	−5.5369	0.0000	***
reduced.health	−0.0054	0.0015	−3.6052	0.0003	***

check for overdispersion

estimation of dispersion parameter

```
phi_est<-sum(
  resid(myGLM, type = "pearson")^2
  )/myGLM$df.residual
cat("estimated dispersion by pearson:", phi_est)
```

```
estimated dispersion by pearson: 1.322044
```

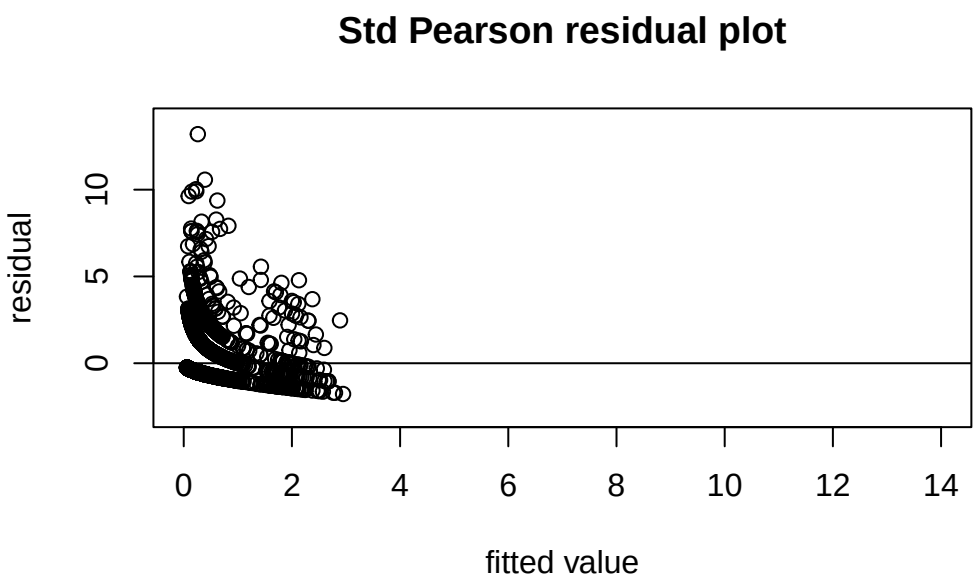
residual analysis

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

```
Variance of Pearson residual: 1.319388
```

```
# 殘差圖 (記得依 fitted value 大小排序)
plot(sort(myGLM$fitted.values),
     rstandard(myGLM, type = "pearson")[order(myGLM$fitted.values, decreasing = FALSE)],
     main="Std Pearson residual plot", ylab = "residual", xlab="fitted value",
     ylim = c(-3,14), xlim=c(0,14))
abline(0,0)
```

```
plot(sort(myGLM$fitted.values),
     resid(myGLM, type = "pearson")[order(myGLM$fitted.values, decreasing = FALSE)],
     main="Pearson residual plot", ylab = "residual", xlab="fitted value",
     ylim = c(-3,14), xlim=c(0,14))
abline(0,0)
abline(0,1)
```



overdispersion test

```
  #g(mu) = mu
  dispersiontest(myGLM,alternative = "greater", trafo = 1)
```

```
    Overdispersion test

data:  myGLM
z = 6.5053, p-value = 3.878e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
    alpha
0.3905122
```

```
  dispersiontest(myGLM,alternative = "greater", trafo = function(mu) mu)
```

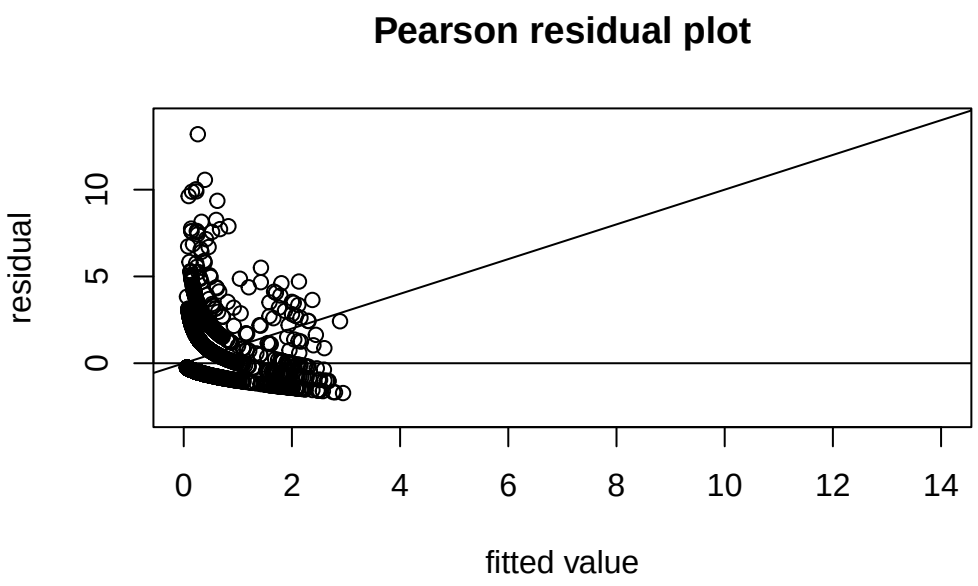
```
    Overdispersion test

data:  myGLM
z = 6.5053, p-value = 3.878e-11
alternative hypothesis: true alpha is greater than 0
sample estimates:
    alpha
0.3905122
```

```
  dispersiontest(myGLM,alternative = "greater")
```

```
    Overdispersion test

data:  myGLM
z = 6.5053, p-value = 3.878e-11
```




```
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.390512
```

```
#g(mu)=mu^2
dispersiontest(myGLM,alternative = "greater", trafo = 2)
```

```
    Overdispersion test

data:  myGLM
z = 7.8987, p-value = 1.409e-15
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
0.988608
```

```
dispersiontest(myGLM,alternative = "greater", trafo = function(mu) mu^2)
```

```
    Overdispersion test

data:  myGLM
z = 7.8987, p-value = 1.409e-15
alternative hypothesis: true alpha is greater than 0
sample estimates:
   alpha
0.988608
```

```
dispersiontest(myGLM,alternative = "greater")
```

```
    Overdispersion test
```

```
data:  myGLM
z = 6.5053, p-value = 3.878e-11
alternative hypothesis: true dispersion is greater than 1
sample estimates:
dispersion
  1.390512
```

```r
overdisp::overdisp(
  x = myGLM$model,
  dependent.position = 1,
  predictor.position = 2:dim(myGLM$model)[2])
```

```
    Overdispersion Test - Cameron & Trivedi (1990)

data:  myGLM$model
Lambda t test score: = 7.8987, p-value = 3.414e-15
alternative hypothesis: Overdispersion
```