Quant_II_hwk_05

吴温泉

目录

1 计数因变量分析

1.1 分析 CGSS2010 数据中 n35a 问题："请问你一共捐献过多少次?" 探讨性别、年龄、收入、共产党员对于捐献次数的关系。

```
dat <- read_dta("cgss2010_14.dta")

# summary(dat)
```

```
male <- ifelse(dat$a2==1, 1,
    ifelse(dat$a2==2, 0, NA)) # gender
a3a <- ifelse(dat$a3a < 17, NA, dat$a3a)  # age
age <- 2010-a3a
a62 <- ifelse(dat$a62 > 9999996, NA, dat$a62) # income
ccpmember <- ifelse(dat$a10 == 1, 1, 0) # member
probs <- c(0, 0.07, 0.25, 0.5, 0.75, 0.93, 1)
kpts <- quantile(a62, prob=probs, na.rm=TRUE)
hinc <- as.numeric(cut(a62, breaks=kpts, labels = 1:6, right=TRUE))
kpts2 <- quantile(a62, prob=c(0,1/6,2/6,3/6,4/6,5/6,1), na.rm=TRUE)
hinc2 <- as.numeric(cut(a62, breaks=kpts2, labels = 1:6, right=TRUE))
loghinc <- log(a62+1)

n35a <- ifelse(dat$n35a<0, 0, dat$n35a)

# dat1 <- cbind.data.frame("contribution"=factor(n35a), male, age, loghinc, ccpmember)
dat1 <- cbind.data.frame("contribution"=n35a, male, age, loghinc, ccpmember)
dat2 <- na.exclude(dat1)
```

1.2 使用 poisson 分析性别、年龄、收入、共产党员（自变量）对捐献次数（因变量）的关系，初步使用系数的正负关系解读因变量和自变量的关系，并说明 Peudo-R2 和似然值检验的结果。

```
M1 <- glm(contribution ~ male + age + loghinc + ccpmember, data=dat1, family=poisson(link="log"))
summary(M1)
```

```
##
## Call:
## glm(formula = contribution ~ male + age + loghinc + ccpmember,
##     family = poisson(link = "log"), data = dat1)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -4.034  -2.371  -1.774  -0.673  60.068
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)  1.748968   0.143660  12.174  < 2e-16 ***
## male        -0.074535   0.030916  -2.411 0.015913 *
## age          0.010596   0.001012  10.469  < 2e-16 ***
## loghinc     -0.043300   0.012294  -3.522 0.000428 ***
## ccpmember   -0.157643   0.054592  -2.888 0.003881 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 12263  on 789  degrees of freedom
## Residual deviance: 12099  on 785  degrees of freedom
##   (10993 observations deleted due to missingness)
## AIC: 14258
##
## Number of Fisher Scoring iterations: 7
```

```r
# pseudo R2
M0 <- update(M1, .~1)
devNull <- deviance(M0)
dev <- deviance(M1)
pR2 <- (devNull - dev) / devNull
pR2
```

```
## [1] 0.04932302
```

```r
# likelihood ratio test
LR <- devNull - dev
k <- length(coef(M1))
prob <- pchisq(LR, df=k-1, lower.tail = FALSE)
prob
```

```
## [1] 1.567043e-134
```

Peudo-R2：有自变量的模型较没有自变数的模型可以解释的偏差比为 4.9% 似然值检验显著，有自变量的模型较没有自变数的模型可以解释 y 更多的偏差，拟合优度 (goodness of fit) 显著性改善。## 使用事件发生率比 (incidence rate ratio) 解释因变量和自变量的关系。

```
# irr, incidence rate ratio
IRR <- function(fit, newdata){
    IR <- predict(fit, newdata, type = "response") #lambdas
    IRR <- IR[2]/IR[1]
    return(IRR)
}

IRR(M1, dat1)
```

```
##        2
## 1.032884
```

1.3 比较 30 岁、收入为均值的共产党党员男性和女性之间在捐献次数的差异。

```
# predictive difference between gender
newX <- with(dat2, cbind.data.frame(male = c(0,1), age = 30, loghinc = mean(loghinc),
                                    ccpmember = 1))
phat <- predict(M1, newdata=newX, type = "terms")
phat[2,] - phat[1,]
```

```
##         male         age     loghinc   ccpmember
## -0.07453545  0.00000000  0.00000000  0.00000000
```

比较除了性别以外相同的两个人，男性比女性少 7 % 的概率认为社会是公正的。

1.4 使用绘图的方式，呈现有自变量的模型和没有自变量的模型对于预测捐献次数概率的差别。

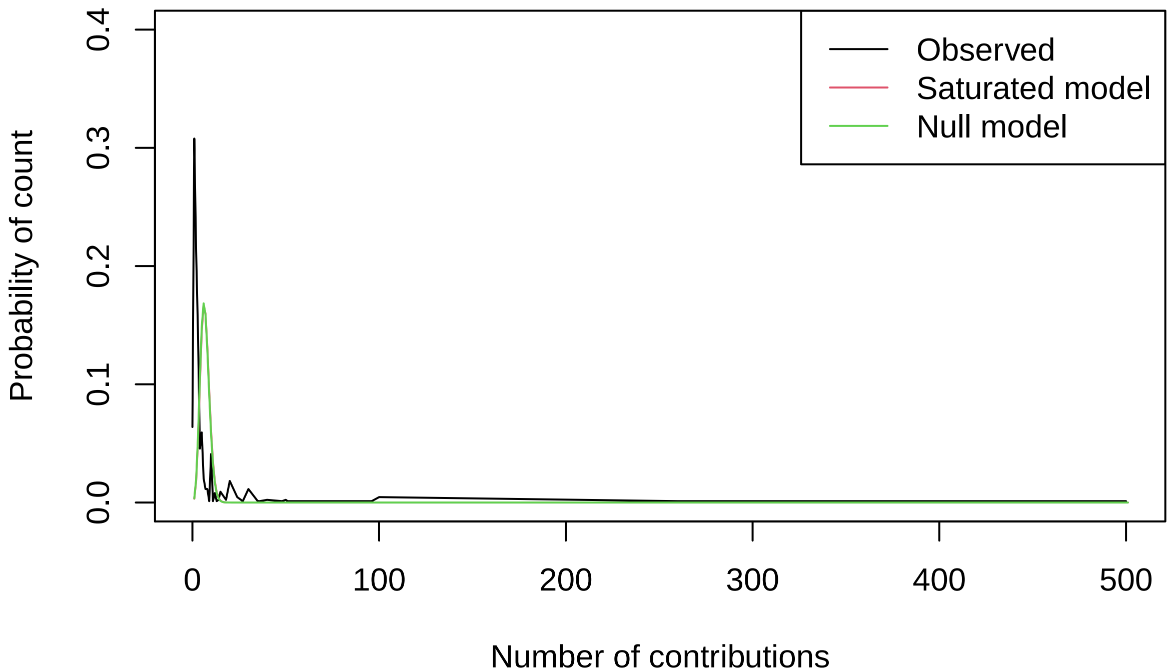
```
# newX <- dat2
newX <- with(dat2, cbind.data.frame(male = mean(male), age = mean(age),  loghinc = mean(loghinc), 


numContri <- seq(0, max(dat2$contribution))
K <- length(numContri)
phatN <- phatS <- rep(NA, K)
for(i in 1:K){
```

```
    lamS <- predict(M1, newdata=newX, type="response")
    phatS[i] <- dpois(numContri[i], lambda=lamS)
    lamN <- predict(M0, newdata=newX, type="response")
    phatN[i] <- dpois(numContri[i], lambda=lamN)
}

tabContri <- table(dat1$contribution)
pObserved <- tabContri / sum(tabContri)
xObserved <- names(pObserved)


plot(0, 0, xlim =c(0, K), ylim=c(0, 0.4), type="n", axes=FALSE, frame.plot=TRUE,
    ylab="Probability of count", xlab="Number of contributions")
lines(x = xObserved, y = pObserved)
#segments(x0=as.numeric(xObserved), y0=0, x1=as.numeric(xObserved), y1=pObserved)
lines(x = 1:K, y = phatS, col=2)
lines(x = 1:K, y = phatN, col=3)
legend("topright", col=1:3, lty=1, legend=c("Observed", "Saturated model", "Null model"))
axis(2)
axis(1)
```



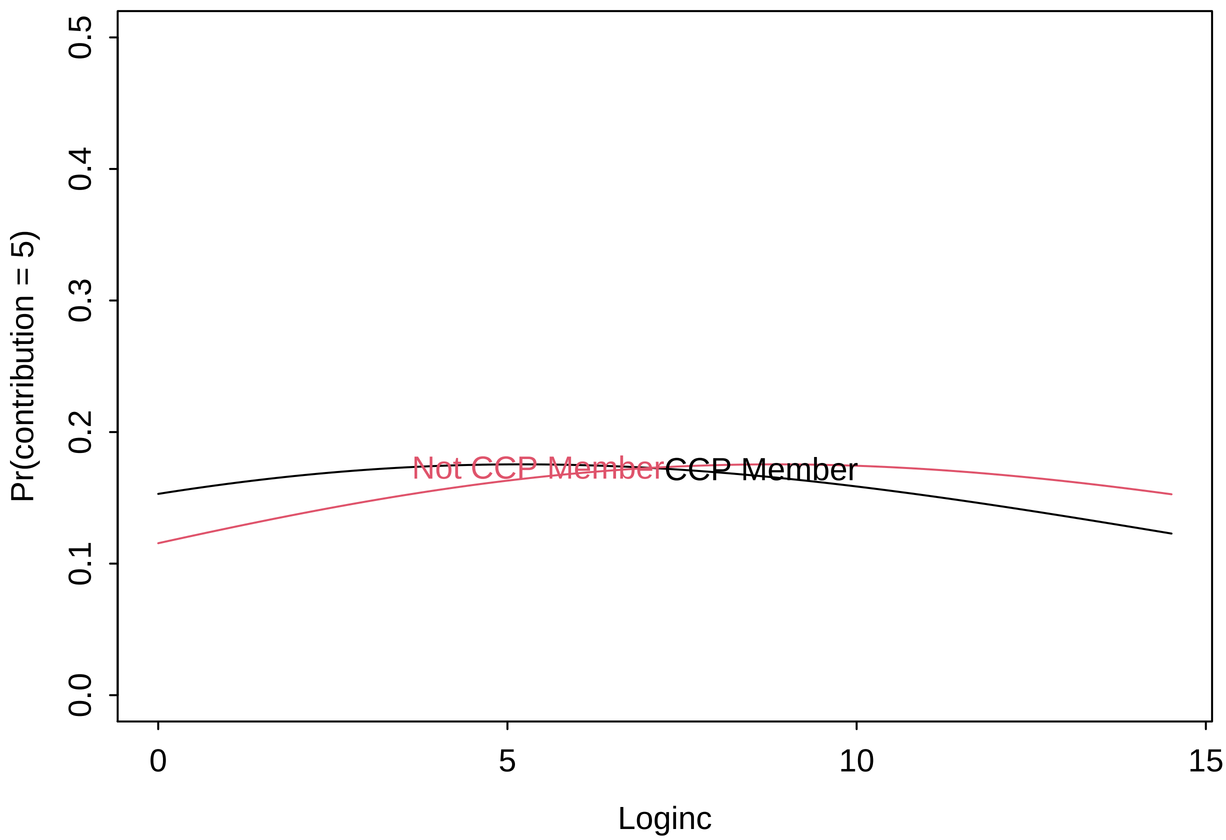
1.5 使用绘图的方式，呈现党员和非党员的 30 岁男性，他在不同收入捐献 5 次的概率差别。你观察到什么现象。

```
rangeInc <- with(dat2, range(loghinc))
incFake <- seq(rangeInc[1],rangeInc[2],length=1000)
newXY <- with(dat2, cbind.data.frame(male = 1, age = 30, ccpmember = 1, loghinc = incFake))
newXN <- with(dat2, cbind.data.frame(male = 1, age = 30, ccpmember = 0, loghinc = incFake))
lambdaY <- predict(M1, newdata=newXY, type = "response")
lambdaN <- predict(M1, newdata=newXN, type = "response")

N <- 5 # publish 2 papers
phatY <- dpois(N, lambdaY) # compute predictive probability given N = 2, and lambda predicted from
phatN <- dpois(N, lambdaN)

par(mar=c(3,3,1,1), mgp=c(2,0.5,0), tcl=-0.2)
plot(x=incFake, y=phatY, type="l", xlim=rangeInc, ylim=c(0,0.5),
    xlab = "Loginc", ylab = "Pr(contribution = 5)")
```

```
lines(x = incFake, y=phatN, col=2)
text(x = incFake[500], y = phatY[500], "CCP Member", adj=0)
text(x = incFake[500], y = phatN[500], "Not CCP Member", adj=1, col=2)
```




```
#legend("topright", lty=1, col=c(1,2), legend = c("Female", "Male"))
# legend(locator(1), lty=1, col=c(1,2), legend = c("CCP Member", "Not CCP Member"))
```

可以发现，比较除了政治面貌以外相同的两个人，在收入较低时，党员比非党员捐献五次的几率更大；随着收入的增加，非党员捐献五次的几率逐渐超过党员。

1.6 使用负二项回归重新检验上述关系，你认为 poisson 和负二项回归那个比较合适？你的主张根据是什么？（提示：α 检验）

```
# negative binomial
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.3
```

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

## The following object is masked from 'package:dplyr':
##
##     select
```

```
M3 <- glm.nb(contribution ~ male + age + ccpmember + loghinc, data=dat1)
summary(M3)
```

```
##
## Call:
## glm.nb(formula = contribution ~ male + age + ccpmember + loghinc,
##     data = dat1, init.theta = 0.644737165, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8368  -0.9584  -0.6537  -0.2605  10.4023
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.863680   0.475720   3.918 8.94e-05 ***
## male        -0.015511   0.097240  -0.160  0.87326
## age          0.009649   0.003272   2.949  0.00319 **
## ccpmember   -0.126392   0.163313  -0.774  0.43898
## loghinc     -0.052712   0.041059  -1.284  0.19921
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.6447) family taken to be 1)
##
##     Null deviance: 868.37  on 789  degrees of freedom
## Residual deviance: 853.81  on 785  degrees of freedom
##   (10993 observations deleted due to missingness)
## AIC: 4408.9
##
## Number of Fisher Scoring iterations: 1
##
##
```

```
##                  Theta:  0.6447
##              Std. Err.:  0.0324
##
##  2 x log-likelihood:  -4396.8550
```

```
# alpha test
G2 <- 2*(logLik(M3) - logLik(M1))
pchisq(G2, df=1, lower.tail=FALSE)
```

```
## 'log Lik.' 0 (df=6)
```

```
# H0: Negbin is the same as Poisson (alpha=0)
```

G2 显著，因此拒绝零假设，负二项回归合适

1.7 使用零膨胀计数回归重新检验上述关系（同时考虑零膨胀 poisson 和负二项回归）。

```
# zero inflated poisson
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
```

```
M5 <- zeroinfl(contribution ~ male + age + ccpmember + loghinc, data = dat1, dist = "poisson", lin
summary(M5)  #
```

```
##
## Call:
## zeroinfl(formula = contribution ~ male + age + ccpmember + loghinc, data = dat1,
##     dist = "poisson", link = "logit")
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
##  -2.1852  -1.6725  -1.3224  -0.5751 195.1810
```

```
##
## Count model coefficients (poisson with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.723421   0.145179  11.871  < 2e-16 ***
## male        -0.054866   0.031128  -1.763  0.07797 .
## age          0.011437   0.001021  11.200  < 2e-16 ***
## ccpmember   -0.121841   0.054767  -2.225  0.02610 *
## loghinc     -0.039841   0.012360  -3.223  0.00127 **
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.88828    1.62176  -2.398   0.0165 *
## male         0.21833    0.32359   0.675   0.4999
## age          0.01689    0.01098   1.538   0.1240
## ccpmember    0.36818    0.46423   0.793   0.4277
## loghinc      0.01459    0.13806   0.106   0.9159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 15
## Log-likelihood: -7007 on 10 Df
```

```r
library(parallel)
library(abind)
cl <- makeCluster(parallel::detectCores())

bootSE <- function(fit, data){
    n <- dim(data)[1]
    idx <- sample(1:n, n, replace=TRUE)
    newData <- data[idx,]
    fit <- update(fit, data = newData)
    betas <- coef(fit)
    return(betas)
}

foo2 <- function() {
    require(pscl)
    replicate(250, bootSE(fit = M5, data = dat1))
}
```

```r
#cl <- makeCluster(spec = 4)
clusterExport(cl = cl, c("M5", "dat1", "bootSE")) # export object to each thread
tryCatch(res <- clusterCall(cl=cl, fun = foo2), finally = stopCluster(cl))
res2 <- abind(res, along=2)
simSes <- apply(res2, 1, sd)




M6 <- zeroinfl(contribution ~ male + age + ccpmember + loghinc, data=dat1, dist = "negbin", link =
```

```
## Warning in value[[3L]](cond): system is computationally singular: reciprocal
## condition number = 2.25585e-39FALSE
```

```r
summary(M6)
```

```
##
## Call:
## zeroinfl(formula = contribution ~ male + age + ccpmember + loghinc, data = dat1,
##     dist = "negbin", link = "logit")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7731 -0.6202 -0.4893 -0.2331 71.0275
##
## Count model coefficients (negbin with log link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.863674         NA      NA       NA
## male        -0.015512         NA      NA       NA
## age          0.009649         NA      NA       NA
## ccpmember   -0.126392         NA      NA       NA
## loghinc     -0.052711         NA      NA       NA
## Log(theta)  -0.438912         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.8758         NA      NA       NA
## male          0.0814         NA      NA       NA
## age         -11.7908         NA      NA       NA
## ccpmember     0.3209         NA      NA       NA
```

```
## loghinc      -2.4706         NA       NA        NA
##
## Theta = 0.6447
## Number of iterations in BFGS optimization: 26
## Log-likelihood: -2198 on 11 Df
```

1.8 使用 AIC 和 BIC 判定何种计数回归模型更适合，并使用 vuong() 检验提出哪个模型更为合适。

```
AIC(M1)
```

```
## [1] 14258.36
```

```
AIC(M3)
```

```
## [1] 4408.855
```

```
AIC(M5)
```

```
## [1] 14034.67
```

```
AIC(M6)
```

```
## [1] 4418.855
```

```
BIC(M1)
```

```
## [1] 14281.72
```

```
BIC(M3)
```

```
## [1] 4436.887
```

```
BIC(M5)
```

```
## [1] 14081.39
```

```r
BIC(M6)
```

```
## [1] 4470.247
```

```r
vuong(M1, M3)
```

```
## NA or numerical zeros or ones encountered in fitted probabilities
## dropping these 1 cases, but proceed with caution
## 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                   -4.206954 model2 > model1 1.2942e-05
## AIC-corrected         -4.206954 model2 > model1 1.2942e-05
## BIC-corrected         -4.206954 model2 > model1 1.2942e-05
```

```r
vuong(M1, M5)
```

```
## NA or numerical zeros or ones encountered in fitted probabilities
## dropping these 1 cases, but proceed with caution
## 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                   -2.614198 model2 > model1 0.0044719
## AIC-corrected         -2.466139 model2 > model1 0.0068289
## BIC-corrected         -2.120365 model2 > model1 0.0169876
```

```r
vuong(M3, M5)
```

```
## NA or numerical zeros or ones encountered in fitted probabilities
## dropping these 1 cases, but proceed with caution
## 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                     4.214942 model1 > model2 1.2492e-05
## AIC-corrected           4.221761 model1 > model2 1.2120e-05
## BIC-corrected           4.237684 model1 > model2 1.1292e-05
```

```
vuong(M5, M6)
```

```
## NA or numerical zeros or ones encountered in fitted probabilities
## dropping these 1 cases, but proceed with caution
## 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                    -4.214942 model2 > model1 1.2492e-05
## AIC-corrected          -4.214942 model2 > model1 1.2492e-05
## BIC-corrected          -4.214942 model2 > model1 1.2492e-05
```

```
vuong(M3, M6)
```

```
## 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                 2.850674e-06 model1 > model2     0.5
## AIC-corrected       9.668177e+04 model1 > model2  <2e-16
## BIC-corrected       3.225320e+05 model1 > model2  <2e-16
```

从 AIC、BIC 可知，M3、M6 明显好与 M1、M2，同时 M3 比 M6 稍好。vuong 检验的正负号可知，M3 解释力大于 M1。M3 的解释力大于 M5。M6 与 M5 的对比中，调整后的 Vuong z-statistic 显著，M6 的解释力大于 M5。因此,M3 为最合适的模型。