GLM_HW2_Poisson

Charlotte Li

4/29/2020

Part 1 of 1 - Models for Count Data In part 1 you will analyze data on cigarette smoking (smoke.csv) The variable "cigs" (the number of cigarettes smoked per day) is the outcome variable. We will consider the following explanatory variables: "lcigpric": the log of the price of cigarettes in the state (cents/pack). "lincome": the log of income (in $) "restauran": 1 if there are restaurant smoking restrictions, 0 otherwise "white": white=1 "nonwhite"=0 "educ": years of education "age" and "agesq": age and age squared.

```
library(pscl)
```

```
## Warning: package 'pscl' was built under R version 4.0.2
```

```
## 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
```

```
setwd("~/Documents/nyu/1stGradSpring/Generalized Linear Model/dataset")
smoke<-read.csv("smoke.csv")
attach(smoke)
```

1. What's the length of exposure for each observation in this dataset

One day

2. Fit a Poisson regression using all of the following variables: lcigpric, lincome, restauran, white, educ, age and agesq. Which one of the following predictors significantly decrease the risk of cigarette smoking?

```
summary(m1<-glm(cigs~lcigpric+lincome+restaurn+white+educ+age+agesq,family="poisson",data=smoke))
```

```
##
## Call:
## glm(formula = cigs ~ lcigpric + lincome + restaurn + white +
##     educ + age + agesq, family = "poisson", data = smoke)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.329  -4.224  -3.275   2.245  13.976
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.964e-01  6.139e-01   0.646    0.518
## lcigpric    -1.060e-01  1.434e-01  -0.739    0.460
## lincome      1.037e-01  2.028e-02   5.115 3.14e-07 ***
## restaurn    -3.636e-01  3.122e-02 -11.646  < 2e-16 ***
## white       -5.520e-02  3.742e-02  -1.475    0.140
## educ        -5.942e-02  4.256e-03 -13.961  < 2e-16 ***
## age          1.143e-01  4.969e-03  22.994  < 2e-16 ***
## agesq       -1.371e-03  5.695e-05 -24.070  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 15821  on 806  degrees of freedom
## Residual deviance: 14752  on 799  degrees of freedom
## AIC: 16239
##
## Number of Fisher Scoring iterations: 6
```

restauran and educ

3. In one sentence,explain the effect of age on the risk of cigarette smoking.

age has significantly positive impact on log(lam) we expect to see more risk of cigarette smoking for those who are older. eˆ0.1143=1.12. on average, the risk of cigareet smoking for older subjects is 1.12 times of younger subjects.

4. Based on the previous model, report the coefficient for restaurant restrictions (in terms of relative risk ratio) and compute a 95% confidence interval for the effect of such restrictions on mean number of cigarettes smoked. [hint: you can use 'mean +-2*SE formula to construct CI] The risk ratio associated with restaurant restrictions is -0.3636, the lower bound 95% CI is -0.4149569 and the upper bound -0.3122431

```
exp(-0.3636)
```

```
## [1] 0.6951692
```

```
exp(-0.3636+2*0.03122)
```

```
## [1] 0.7399594
```

```
exp(-0.3636-2*0.03122)
```

```
## [1] 0.6530902
```

5 In one sentence, interpret the coefficient (in terms of relative risk ratio) for restaurant restrictions

restaurants with restrictions have significantly negative impact on the risk of cigarette smoking, the log (lam) which is 0.36 less than that of restaurants without restrictions.

6.Predict the mean number of cigarettes smoked (also the rate of occurrence) for each individual per day based on this model. And further compute the predicted probabilities of smoking zero cigarettes based on the rate of occurrence using a Poisson probability distribution function. On average, the model predicts % of subjects who smoke 0 cigarettes a day

```
## predict lambda for each observation
pred1<-predict(m1) ##=log(lam)
## exponentiate pred1 to get lam
lambda1<-exp(pred1)
## get mean of Y=0 which is -lam
mean(exp(-lambda1))
```

```
## [1] 0.008704774
```

7 Tabulate the outcome variable, actually what percentage of subjects smoker 0 cigarette?

```
table(smoke$cigs)/nrow(smoke)
```

```
##
##           0           1           2           3           4           5
## 0.615861214 0.008674102 0.006195787 0.006195787 0.002478315 0.008674102
##           6           7           8           9          10          11
## 0.003717472 0.002478315 0.003717472 0.002478315 0.034696406 0.002478315
##          12          13          14          15          16          18
## 0.004956629 0.002478315 0.001239157 0.028500620 0.001239157 0.003717472
##          19          20          25          28          30          33
## 0.001239157 0.125154895 0.008674102 0.003717472 0.052044610 0.001239157
##          35          40          50          55          60          80
## 0.002478315 0.045848823 0.007434944 0.001239157 0.009913259 0.001239157
```

61.586%

8 Briefly comment on the possible issues of this poisson regression model.

There are over 60% of the people who don't smoke while the model predicts only less than 1% non-smokers. There could be a problem of zero-inflation (namely, some people are not at risk of smoking), or there could exist substantial individual variation after accounting for the observed covariates.

9 T/F

In this data, the variance is close to 20 times the mean. T Overdispersion: after fitting model, poisson as a desired model $Yi\ Poisson(\lambda)$ $log(lambda_i) = X_ibeta$ –Overdispersion: scale up SE proportionally by a factor of sqrt(18)= –14752/799=18 deviance resi/df=var

#10 (continued from Q 9) This suggests that the standard errors in the Poisson Regression is severely 1 Points underestimated. To correct this problem, approximately, standard errors should be multiplied by a factor of 4.

SE=sqrt(20) T

11 After correcting the standard errors, the log income is still a significant predictor to cigarette smoking. (hint: In R, use quasipoisson distribution; In Stata, use scale(x2) option

```r
summary(glm(cigs~lcigpric+lincome+restaurn+white+educ+age+agesq,family="quasipoisson",data=smoke))
```

```
##
## Call:
## glm(formula = cigs ~ lcigpric + lincome + restaurn + white +
##     educ + age + agesq, family = "quasipoisson", data = smoke)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.329  -4.224  -3.275   2.245  13.976
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3964489  2.7673851   0.143  0.88612
## lcigpric    -0.1059606  0.6463348  -0.164  0.86982
## lincome      0.1037276  0.0914154   1.135  0.25685
## restaurn    -0.3636059  0.1407348  -2.584  0.00995 **
## white       -0.0552011  0.1686704  -0.327  0.74355
## educ        -0.0594225  0.0191852  -3.097  0.00202 **
## age          0.1142571  0.0223981   5.101 4.22e-07 ***
## agesq       -0.0013708  0.0002567  -5.340 1.21e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 20.31782)
##
##     Null deviance: 15821  on 806  degrees of freedom
## Residual deviance: 14752  on 799  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

False

Q12 Fit a negative binomial model using the same set of predictors. Explain the key rationale behind this model. [Note if you use R, you will encounter a technical problem. The R glm.nb function has an issue estimating the nb regression for this data. See attached explanation and a patch so you can complete questions 12-15, as well as the question on AIC and BIC)

```
library(MASS)
library(foreign)
neg.bin.model <- glm(cigs~lcigpric+lincome+restaurn+white+educ+age+agesq, data = smoke,family = negativ
summary(neg.bin.model)
```

```
##
## Call:
## glm(formula = cigs ~ lcigpric + lincome + restaurn + white +
##     educ + age + agesq, family = negative.binomial(theta = 0.1357),
##     data = smoke)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1988  -1.0622  -0.9793   0.2290   1.7501
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.9632627  3.1540268   0.305  0.76014
## lcigpric    -0.2040665  0.7525856  -0.271  0.78634
## lincome      0.1024844  0.0951551   1.077  0.28179
## restaurn    -0.4703740  0.1458408  -3.225  0.00131 **
## white       -0.1822397  0.1900767  -0.959  0.33797
## educ        -0.0944611  0.0218044  -4.332 1.66e-05 ***
## age          0.1360158  0.0211620   6.427 2.23e-10 ***
## agesq       -0.0016186  0.0002314  -6.995 5.61e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.1357) family taken to be 0.4054058)
##
##     Null deviance: 635.73  on 806  degrees of freedom
## Residual deviance: 612.62  on 799  degrees of freedom
## AIC: 3875.7
##
## Number of Fisher Scoring iterations: 16
```

restaurant restrictions, education, age and age squared are significantly associated with risk of cigrette smoking. The negative-binomial model allows the variance of the distribution to be bigger than the mean (which is useful in cases of over-dispersion). It does this by adding an individual-level factor, so that each individual has a different rate of occurrence. We can't estimate the individual-level factor (since we only have one observation per individual), but we can estimate from our sample the variance of the individual level factors, which is denoted alpha; this variance, along with lambda, are the two parameters of the negative binomial distribution.

13 Compare with the Poisson Regression, in the neg. binomial model, we observe:

–the direction of the coefficients remain the same –the standard errors are much larger

14 In this negative binomial model, the variance of individual multiplying factors is 7.36. If there is no individual heterogeneity, we should expect the variance to be close to 0

15 Predict the percent of respondents who smoke zero cigarettes per day.The percent of respondents who smoke zero cigarettes per day is

```r
## predict lambda for each observation
pred.nb<-predict(neg.bin.model,data=smoke,type = "response")
alpha<-1/7.369
pn0<-(alpha/(pred.nb+alpha))^alpha
zero<-sum(pn0)
p.zero<-100*zero/dim(smoke)[1]
```

16 Now fit a zero-inflated Poisson Model. First briefly explain the rationale behind this model.

(For R: you will notice that the fit of zeroinfl is also a little bit wierd, with warning of covariance matrix not estimated correctly. To improve the stability of model fitting, you can try the following: instead of using age and agesq, try mean centered age (age-mean(age)) and its squared term in the model. This way the coefficients of other covariates are the same, except the coefficient of centered age and its squared are different –but they are still fitting the same quadratic relationship!)

```r
library(pscl)
age.center=age-mean(age)
age.center.sq<-age.center*age.center
zip1<-zeroinfl(cigs~lcigpric+lincome+restaurn+white+educ+age.center+age.center.sq|1, data = smoke)
summary(zip1)
```

```
##
## Call:
## zeroinfl(formula = cigs ~ lcigpric + lincome + restaurn + white + educ +
##     age.center + age.center.sq | 1, data = smoke)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -0.7711 -0.7646 -0.7551  0.7622  5.6099
##
## Count model coefficients (poisson with log link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.700e+00  6.111e-01   2.781  0.00541 **
## lcigpric       1.017e-01  1.422e-01   0.715  0.47464
## lincome        9.031e-02  1.994e-02   4.530 5.90e-06 ***
## restaurn      -1.267e-01  3.146e-02  -4.027 5.64e-05 ***
## white         -1.475e-02  3.735e-02  -0.395  0.69289    
## educ           2.410e-02  4.791e-03   5.029 4.93e-07 ***
## age.center     7.115e-03  8.871e-04   8.020 1.05e-15 ***
## age.center.sq -5.586e-04  5.485e-05 -10.185  < 2e-16 ***
## 
## Zero-inflation model coefficients (binomial with logit link):
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  0.47202    0.07237   6.522 6.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Number of iterations in BFGS optimization: 1 
## Log-likelihood: -2349 on 9 Df
```

Sometimes, mean of the count is under/over estimated. ZIP allows assumption of not all people smoke, those intend to smoke may not smoke as well. In ZIP model(who don't have intention to smoke), no explanatory variable is used. it is a binomial with logit link. log odds of smoke over not smoke is 0.427. For each subject ,there is a common probability of falling into cig=0: P=exp(0.472)/(1+exp(0.472))=0.605

17 Fit a zero-inflated poisson model using the same set of predictors for the Poisson part of the model, while only include the intercept for the zero-inflation part (the logistic regression part). This model estimate percent of subjects who are not at risk of smoking.

```
#prob of being non-smoking 0.615
exp(0.47202)/(1+exp(0.47202))
```

```
## [1] 0.6158618
```

```
#intercept for the inflate part is 0.47,also the log odds of non-smoking
```

#18 Now include the same set of explanatory variables into the inflate part of the model. Which variables significantly predict whether one is at risk of smoking at all.

```
summary(zip3<-zeroinfl(cigs~lcigpric+lincome+restaurn+white+educ+age+agesq|lcigpric+lincome+restaurn+whi
```

```
## Warning in sqrt(diag(object$vcov)): NaNs produced
```

```
## 
## Call:
## zeroinfl(formula = cigs ~ lcigpric + lincome + restaurn + white + educ + 
##     age + agesq | lcigpric + lincome + restaurn + white + educ + age.center + 
##     age.center.sq, data = smoke)
## 
## Pearson residuals:
##     Min      1Q  Median      3Q     Max 
## -1.4371 -0.7750 -0.5905  0.6338  5.6077 
## 
## Count model coefficients (poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  0.4564177  0.6080393   0.751    0.453    
## lcigpric     0.1016759  0.1421280   0.715    0.474    
## lincome      0.0903061  0.0192098   4.701 2.59e-06 ***
## restaurn    -0.1267156  0.0310735  -4.078 4.54e-05 ***
## white       -0.0147493  0.0360687  -0.409    0.683
## educ         0.0240952  0.0046483   5.184 2.18e-07 ***
## age          0.0531856  0.0002268 234.537  < 2e-16 ***
## agesq       -0.0005586         NA      NA       NA
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.7044314  3.8491166  -0.703   0.4823
## lcigpric       0.3395214  0.9080756   0.374   0.7085
## lincome       -0.0472620  0.1158289  -0.408   0.6832
## restaurn       0.4572348  0.1828355   2.501   0.0124 *
## white          0.1133361  0.2335092   0.485   0.6274
## educ           0.1350802  0.0277349   4.870 1.11e-06 ***
## age.center     0.0087192  0.0050483   1.727   0.0841 .
## age.center.sq  0.0013623  0.0003121   4.366 1.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -2322 on 16 Df
```

#19 Based on the zero-inflated model in Q19, predict the proportion of respondents who smoke zero cigarettes. The percentage of respondents who smoke zero cigarettes is %

```
## predict the probability of Y=y
predict.zip3<-predict(zip3,type = "prob")
## predict the prob of belonging to never smoke
zero.zip3<-predict(zip3,type = "zero")
## marginally how many 0 predicted
sum(predict.zip3[,1])/nrow(predict.zip3)
```

```
## [1] 0.6158613
```

```
## compare to observed 0s
sum(smoke$cigs==0)/nrow(smoke)
```

```
## [1] 0.6158612
```

20 Based on the AIC and BIC criterion, which model appears to fit the data the best?

```
# possion
AIC(m1)
```

```
## [1] 16239.04
```

```
# neg. bi
AIC(neg.bin.model)
```

```
## [1] 3875.7
```

```
# zip with no predictors
AIC(zip1)
```

```
## [1] 4716.756
```

```r
# zip with full set of predictor
AIC(zip3)
```

```
## [1] 4676.135
```