Categorical+ Continuous(Covariates) ANCOVA, Poi GLM

SIT Wai Tang

Description: German health registry for the year 1984. A data frame with 3,874 observations on the following 16 variables.

docvis: number of visits to doctor during year (0-121)

hospvis: number of days in hospital during year (0-51)

edlevel: educational level (categorical: 1-4)

age: 25-64

outwork: out of work=1; 0=working

female: female=1; 0=male

married: married=1; 0=not married

kids: have children=1; no children=0

hhninc: household yearly income in marks (in Marks)

educ: years of formal education (7-18)

self: self-employed=1; not self employed=0

edlevel1: (1/0) not high school graduate

edlevel2: (1/0) high school graduate

edlevel3: (1/0) university/college

edlevel4: (1/0) graduate school

```
rwm1984 = read.csv("rwm1984.csv")

head(rwm1984)
```

```
##   X docvis hospvis edlevel age outwork female married kids hhninc educ self
## 1 1      1       0       3  54       0      0       1    0  3.050 15.0    0
## 2 2      0       0       1  44       1      1       1    0  3.050  9.0    0
## 3 3      0       0       1  58       1      1       0    0  1.434 11.0    0
## 4 4      7       2       1  64       0      0       0    0  1.500 10.5    0
## 5 5      6       0       3  30       1      0       0    0  2.400 13.0    0
## 6 6      9       0       3  26       1      0       0    0  1.050 13.0    0
##   edlevel1 edlevel2 edlevel3 edlevel4
## 1        0        0        1        0
## 2        1        0        0        0
## 3        1        0        0        0
## 4        1        0        0        0
## 5        0        0        1        0
## 6        0        0        1        0
```

Poission General Linear Modeling

log-linear mean function to predict docvis(number of visits to doctor)

```
glmrp = glm(docvis ~ outwork + factor(female)+factor(married) +factor(kids)+ age + factor(edlevel), family=poisson, data=rwm
1984)
glmrp
```

```
##
## Call:  glm(formula = docvis ~ outwork + factor(female) + factor(married) +
##     factor(kids) + age + factor(edlevel), family = poisson, data = rwm1984)
##
## Coefficients:
##      (Intercept)           outwork   factor(female)1  factor(married)1
##          0.16106           0.27256           0.26036          -0.09489
##    factor(kids)1               age  factor(edlevel)2  factor(edlevel)3
##         -0.12427           0.01941          -0.07110          -0.19504
## factor(edlevel)4
##         -0.26386
##
## Degrees of Freedom: 3873 Total (i.e. Null);  3865 Residual
## Null Deviance:       25790
## Residual Deviance: 23890     AIC: 30990
```

```
summary(glmrp)
```

```
##
## Call:
## glm(formula = docvis ~ outwork + factor(female) + factor(married) +
##     factor(kids) + age + factor(edlevel), family = poisson, data = rwm1984)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7235  -2.1676  -1.2563   0.3216  25.9023
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.1610617  0.0512422   3.143  0.00167 **
## outwork           0.2725570  0.0215072  12.673  < 2e-16 ***
## factor(female)1   0.2603575  0.0211447  12.313  < 2e-16 ***
## factor(married)1 -0.0948931  0.0226585  -4.188 2.81e-05 ***
## factor(kids)1    -0.1242692  0.0222559  -5.584 2.36e-08 ***
## age               0.0194091  0.0009722  19.964  < 2e-16 ***
## factor(edlevel)2 -0.0711018  0.0422948  -1.681  0.09274 .
## factor(edlevel)3 -0.1950417  0.0398849  -4.890 1.01e-06 ***
## factor(edlevel)4 -0.2638602  0.0480863  -5.487 4.08e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 25791  on 3873  degrees of freedom
## Residual deviance: 23891  on 3865  degrees of freedom
## AIC: 30992
##
## Number of Fisher Scoring iterations: 6
```

```
exp(coef(glmrp))
```

```
##      (Intercept)          outwork  factor(female)1 factor(married)1
##        1.1747574        1.3133184        1.2973938        0.9094702
##    factor(kids)1              age factor(edlevel)2 factor(edlevel)3
##        0.8831420        1.0195986        0.9313670        0.8228004
## factor(edlevel)4
##        0.7680809
```

The fitted model is:

$log(\hat{\mu}) = 0.1611 + 0.2726outwork + 0.2603I_{female=1} - 0.0949I_{married=1} - 0.1243I_{kids=1} + 0.0194age - 0.0711I_{edlevel=2} - 0.1950I_{edlevel=3} - 0.2$

the effect of edlevel on dovis

```
coef(glmrp)
```

```
##      (Intercept)          outwork  factor(female)1 factor(married)1
##       0.16106166       0.27255703       0.26035749      -0.09489309
##    factor(kids)1              age factor(edlevel)2 factor(edlevel)3
##      -0.12426924       0.01940907      -0.07110184      -0.19504168
## factor(edlevel)4
##      -0.26386018
```

```
exp(coef(glmrp))
```

```
##      (Intercept)          outwork  factor(female)1 factor(married)1
##        1.1747574        1.3133184        1.2973938        0.9094702
##    factor(kids)1              age factor(edlevel)2 factor(edlevel)3
##        0.8831420        1.0195986        0.9313670        0.8228004
## factor(edlevel)4
##        0.7680809
```

```
exp(coef(glmrp))-1
```

```
##      (Intercept)          outwork  factor(female)1 factor(married)1
##       0.17475740       0.31331836       0.29739380      -0.09052984
##    factor(kids)1              age factor(edlevel)2 factor(edlevel)3
##      -0.11685797       0.01959865      -0.06863296      -0.17719964
## factor(edlevel)4
##      -0.23191908
```

the effect of edlevel on dovis:

The coefficient for the edlevel effects are estimated as $ -0.0711(2), -0.1950(3), -0.2639(4)$, indicating that high school graduation will affect the dovis negatively, when the other variables are kept fixed. Higher of the education level, less the number of visits to doctors.

The average ratios are estimated as $0.9313670, 0.8228004, 0.7680809$, i.e. about 6.86%, 17,72%, 23.19% less in the expected number of visits to doctor for high school, university, graduate school education level, respectively, compared with edlevel=1 group, i.e., not high school graduate group.

the effect of sex on dovis:

The coefficient for the sex effect is estimated as $\hat{\beta}_{female=1} = 0.26035$, indicating that female visit doctors more often than male, when the other variables are kept fixed. The average ratio is estimated as $e^{0.26035} = 1.2974$, i.e.about 29.74% more in the expected number of visits to doctor of female than male.

the effect of age on dovis:

The coefficient for the age effect is estimated as $\hat{\beta}_{age} = 0.0194$. When the other variables are kept fixed, one more year in age is associated with an average ratio $e^{0.0194} = 1.0195$, i.e. about 1.95% increase in the expected number of visits to doctor. ### ——————

Predicte the number of visits to doctor for a woman of 35 years old, working at a company, married, having kids, and having graduate school degree

```
test1 = data.frame(outwork=0, female=1, married=1, age=35, kids=1 ,edlevel=4)
pred = predict(glmrp, newdata=test1, interval="confidence", se=TRUE)
pred
```

```
## $fit
##        1
## 0.617714
##
## $se.fit
## [1] 0.05054894
##
## $residual.scale
## [1] 1
```

95% Confidence Interval

```
exp(pred$fit-1.96*pred$se.fit)
```

```
##        1
## 1.679738
```

```
exp(pred$fit+1.96*pred$se.fit)
```

```
##        1
## 2.047849
```

---

numvisit: visits to MD office 3month prior

reform: 1=interview year post-reform: 1998; 0=pre-reform:1996

badh: 1=bad health; 0 = not bad health

age: Age(years 20-60)

educ: education(1: 7-10; 2=10.5-12; 3=High School gradudate+)

educ1: educ1= 7-10 years

educ2: educ2= 10.5-12 years

educ3: educ3= post secondary or high school

agegrp: age: 1=20-39; 2=40-49; 3=50-60

age1: age 20-39

age2: age 40-49

age3: age 50-60

loginc: log(household income in DM)

```
mdvis <- read.csv("mdvis.csv", header = T)
head(mdvis)
```

```
##   numvisit reform badh age educ educ1 educ2 educ3 agegrp age1 age2 age3
## 1       30      1    0  58    2     0     1     0      3    0    0    1
## 2       25      0    0  24    2     0     1     0      1    1    0    0
## 3       25      0    0  50    3     0     0     1      3    0    0    1
## 4       25      0    0  40    1     1     0     0      2    0    1    0
## 5       20      1    0  54    2     0     1     0      3    0    0    1
## 6       60      0    1  29    2     0     1     0      1    1    0    0
##     loginc
## 1 7.870875
## 2 7.672544
## 3 7.194270
## 4 8.104677
## 5 6.484581
## 6 7.664526
```

```
mdvis$reform <- as.factor(mdvis$reform)
mdvis$badh <- as.factor(mdvis$badh)
mdvis$educ <- factor(mdvis$educ)
mdvis$agegrp <- factor(mdvis$agegrp)
str(mdvis)
```

```
## 'data.frame':    2227 obs. of  13 variables:
##  $ numvisit: int  30 25 25 25 20 60 20 20 16 20 ...
##  $ reform  : Factor w/ 2 levels "0","1": 2 1 1 1 2 1 1 1 2 2 ...
##  $ badh    : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 1 1 1 1 ...
##  $ age     : int  58 24 50 40 54 29 24 25 44 57 ...
##  $ educ    : Factor w/ 3 levels "1","2","3": 2 2 3 1 2 2 2 2 3 2 ...
##  $ educ1   : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ educ2   : int  1 1 0 0 1 1 1 1 0 1 ...
##  $ educ3   : int  0 0 1 0 0 0 0 0 1 0 ...
##  $ agegrp  : Factor w/ 3 levels "1","2","3": 3 1 3 2 3 1 1 1 2 3 ...
##  $ age1    : int  0 1 0 0 0 1 1 1 0 0 ...
##  $ age2    : int  0 0 0 1 0 0 0 0 1 0 ...
##  $ age3    : int  1 0 1 0 1 0 0 0 0 1 ...
##  $ loginc  : num  7.87 7.67 7.19 8.1 6.48 ...
```

Poission General Linear Modeling

Filter the variables that can explain the response numvisit

```
fit <- glm(numvisit ~ reform+badh+loginc+educ+agegrp, data=mdvis, family = poisson(link = "log"))
anova(fit, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: numvisit
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                    2226     8848.8
## reform  1    49.11      2225     8799.7 2.422e-12 ***
## badh    1  1337.85      2224     7461.9 < 2.2e-16 ***
## loginc  1    17.59      2223     7444.3 2.735e-05 ***
## educ    2    29.20      2221     7415.1 4.558e-07 ***
## agegrp  2    16.80      2219     7398.3 0.0002249 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The anova test suggests that all the factors included are significant.

Estmate and standard errors of the log-linear mean function

```
summary(fit)
```

```
##
## Call:
## glm(formula = numvisit ~ reform + badh + loginc + educ + agegrp,
##     family = poisson(link = "log"), data = mdvis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1107  -1.9225  -0.6676   0.5588  12.2675
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.31496    0.27244  -1.156 0.247648
## reform1     -0.13850    0.02656  -5.215 1.84e-07 ***
## badh1        1.14177    0.02995  38.127  < 2e-16 ***
## loginc       0.13697    0.03583   3.822 0.000132 ***
## educ2        0.08179    0.03365   2.431 0.015066 *
## educ3       -0.08104    0.03692  -2.195 0.028180 *
## agegrp2      0.09076    0.03217   2.821 0.004784 **
## agegrp3      0.13525    0.03589   3.769 0.000164 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 8848.8  on 2226  degrees of freedom
## Residual deviance: 7398.3  on 2219  degrees of freedom
## AIC: 11880
##
## Number of Fisher Scoring iterations: 5
```

```
exp(coef(fit))
```

```
## (Intercept)     reform1       badh1      loginc       educ2       educ3
##   0.7298203   0.8706594   3.1323084   1.1467906   1.0852240   0.9221584
##     agegrp2     agegrp3
##   1.0950015   1.1448178
```

```
"______"
```

```
## [1] "______"
```

```
exp(coef(fit))-1
```

```
## (Intercept)     reform1       badh1      loginc       educ2       educ3
## -0.27017968 -0.12934059  2.13230840  0.14679059  0.08522397 -0.07784164
##     agegrp2     agegrp3
##  0.09500148  0.14481781
```

the effect of reform (health reform) and loginc (household income) on numvisit(number of patient visits to a physician's office)

The average ratio of health reform is estimated as e^(-0.13850) = 0.87066, i.e. about 12.93% reduced in the expected numvisit;

and household income is estimated as e^(0.13697) = 1.14679, i.e. about 14.68% more in the expected numvisit.

Validity on using linear regression model to fit response numvisit

```
fit_lm = lm(numvisit ~ reform+badh+loginc+educ+agegrp, data = mdvis)
summary(fit_lm)
```

```
##
## Call:
## lm(formula = numvisit ~ reform + badh + loginc + educ + agegrp,
##     data = mdvis)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.338 -1.919 -0.832  0.858 53.145
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.7288     1.6469  -0.443   0.6581
## reform1      -0.3561     0.1588  -2.243   0.0250 *
## badh1         4.5351     0.2537  17.876   <2e-16 ***
## loginc        0.3695     0.2172   1.701   0.0891 .
## educ2         0.2159     0.2074   1.041   0.2980
## educ3        -0.1987     0.2194  -0.905   0.3653
## agegrp2       0.2146     0.1958   1.096   0.2734
## agegrp3       0.3740     0.2284   1.638   0.1017
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.731 on 2219 degrees of freedom
## Multiple R-squared:  0.1398, Adjusted R-squared:  0.137
## F-statistic:  51.5 on 7 and 2219 DF,  p-value: < 2.2e-16
```

```
shapiro.test(fit_lm$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  fit_lm$residuals
## W = 0.68494, p-value < 2.2e-16
```

Since the p-value < 2.2e-16, we reject the H0 and conclude that the residuals do not follow normal distribution,and the linear regression model deviates from normal distribution a lot. Hence, it is not valid to use a linear regression model.