STAT5703 HW3 Ex4

Chao Huang (ch3474), Wancheng Chen (wc2687), Chengchao Jin (cj2628)

Problem 4

Question 1. Fit a Poisson model

```
library(SDMTools)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data <- read.table("./docvisits.asc", header = T) %>% select(1:13)
```

```
pmodel = glm(dvisits ~. , data = data, family = poisson)
summary(pmodel)
```

```
##
## Call:
## glm(formula = dvisits ~ ., family = poisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9170  -0.6862  -0.5743  -0.4839   5.7005
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.223848   0.189816 -11.716   <2e-16 ***
## sex          0.156882   0.056137   2.795   0.0052 **
## age          1.056299   1.000780   1.055   0.2912
## agesq       -0.848704   1.077784  -0.787   0.4310
## income      -0.205321   0.088379  -2.323   0.0202 *
## levyplus     0.123185   0.071640   1.720   0.0855 .
## freepoor    -0.440061   0.179811  -2.447   0.0144 *
## freerepa     0.079798   0.092060   0.867   0.3860
## illness      0.186948   0.018281  10.227   <2e-16 ***
## actdays      0.126846   0.005034  25.198   <2e-16 ***
## hscore       0.030081   0.010099   2.979   0.0029 **
## chcond1      0.114085   0.066640   1.712   0.0869 .
## chcond2      0.141158   0.083145   1.698   0.0896 .
## ---
## 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: 4379.5  on 5177  degrees of freedom
## AIC: 6737.1
##
## Number of Fisher Scoring iterations: 6
```

I don't think this model fits the data well, since the residual deviance is not much different from the null deviance so it shows sign of over diversion.

Question 2,3,4 (with handwritten solutions)

2. Denote I_i as an indicator R.V., where $I_i = 1$ if and only if $y_i \geq 0$. Also, denote $\pi_i \triangleq \pi(x_i)$, $\lambda_i \triangleq \lambda(x_i)$.

$$\ell(\pi, \lambda; y, x) = \sum_i \log(1-\pi_i)\cdot(1-I_i) + I_i \log(\pi_i) + I_i\left(-\lambda_i + y_i \log\lambda_i - \log(y_i!) - \log(1-e^{-\lambda_i})\right)$$

$$= \underbrace{\sum_i \log(1-\pi_i) + I_i \cdot \log\frac{\pi_i}{1-\pi_i}}_{\ell_1(\pi; y, x)} + \underbrace{\sum_i I_i\left(-\lambda_i + y_i \log\lambda_i - \log(y!) - \log(1-e^{-\lambda_i})\right)}_{\ell_2(\lambda; y, x)}$$

3. ℓ_1 corresponds to a binomial model with all data samples

ℓ_2 corresponds to a truncated poisson model with $y_i > 0$.

so they can be optimized seperately.

Therefore, for ℓ_1, we have. $\phi = 1$. $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$.

$\mu_i = \frac{e^\theta}{1+e^\theta}$. $V(\mu_i) = \mu_i(1-\mu_i)$.

since $\mu_i = \pi_i = \pi(x_i^T\beta)$. link function $g = \pi^{-1}$

Using (10.18) from textbook. we have,

Figure 1: Q2 - Q4 Handwritten Solutions (Page 1)

$$\frac{\partial l_1(\beta)}{\partial \beta} = X^T u(\beta). \quad \text{where} \quad u_j = \frac{y_j - \mu_j}{g'(\mu_j) V(\mu_j)}$$

where $g'(\mu_j) = \frac{d g(\mu_j)}{d \mu_j}$

for l_2, similarly. we have. $\theta_i = \log \lambda_i$. $\phi = 1$

$b(\theta_i) = e^{\theta_i} + \log(1 - e^{-e^{\theta_i}})$. $\mu_i = \frac{\lambda_i}{1 - e^{-\lambda_i}}$. $V(\mu_i) = \mu_i (1 + \lambda - \mu_i)$

since $\mu_i = \frac{\lambda(x_i^T \gamma)}{1 - e^{-\lambda(x_i^T \gamma)}} = f(x_i^T \gamma)$. the link function is.

$g = f^{-1}$.

so $\frac{\partial l_2(\gamma)}{\partial \gamma} = X^T u(\gamma)$. where $u_j = \frac{y_j - \mu_j}{g'(\mu_j) V(\mu_j)}$

4. $\begin{bmatrix} \hat{\beta}_{mle} \\ \hat{\gamma}_{mle} \end{bmatrix} \overset{a.s.}{\sim} N\left(\begin{bmatrix} \hat{\beta}_{mle} \\ \hat{\gamma}_{mle} \end{bmatrix}, \frac{1}{n} \begin{bmatrix} I_1 & 0 \\ 0 & I_2 \end{bmatrix}^{-1} \right)$

where $I_1 = -E\left[\frac{\partial^2 l_1(\beta)}{\partial \beta \, \partial \beta^T} \right]$. $I_2 = -E\left[\frac{\partial^2 l_2(\gamma)}{\partial \gamma . \partial \gamma^T} \right]$

The variance matrix can be split to 2 blocks since the likelihood function can be split to two seperate parts.

Figure 2: Q2 - Q4 Handwritten Solutions (Page 2)

This model set aside $y = 0$ as a special case and model all the other values of y using a truncated model. So in this way, we will not let the large number of $y = 0$ samples affect our estimation of parameters of the Poisson distribution. Also the parameters can be estimated seperately since two groups of parameters don't influence each other. In other words, we can first fit a binomial model using binary data, then fit a truncated poisson model using truncated Poisson model.

Question 5. Fit hurdle model

```
source("./truncpoisson.R")
```

```
summary(VGAM::vglm(formula = dvisits ~ ., data = data %>% filter(dvisits > 0), family = VGAM::pospoisso
```

```
## Length  Class   Mode
##      1   vglm     S4
```

```
trunc_model <- glm(formula = dvisits ~ ., data = data %>% filter(dvisits > 0), family = truncpoisson)
summary(trunc_model)
```

```
##
## Call:
## glm(formula = dvisits ~ ., family = truncpoisson, data = data %>%
##     filter(dvisits > 0))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2895  -0.7389  -0.6420  -0.5169   4.9121
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.313877   0.444832  -2.954  0.00321 **
## sex          0.005410   0.128402   0.042  0.96640
## age          4.185247   2.277770   1.837  0.06643 .
## agesq       -4.487536   2.420408  -1.854  0.06402 .
## income      -0.526491   0.217412  -2.422  0.01562 *
## levyplus    -0.176421   0.169666  -1.040  0.29867
## freepoor     0.036869   0.377409   0.098  0.92220
## freerepa    -0.463598   0.207714  -2.232  0.02583 *
## illness      0.080932   0.042566   1.901  0.05754 .
## actdays      0.122332   0.010135  12.071  < 2e-16 ***
## hscore       0.004612   0.020927   0.220  0.82560
## chcond1      0.022394   0.165417   0.135  0.89234
## chcond2      0.010164   0.186391   0.055  0.95652
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for truncpoisson family taken to be 2.038733)
##
##     Null deviance: 1506.3  on 1048  degrees of freedom
## Residual deviance: 1170.8  on 1036  degrees of freedom
## AIC: 2236.5
##
## Number of Fisher Scoring iterations: 9
```

```
binary_data <- data %>%
  mutate(bin_visits = ifelse(dvisits > 0, 1 , 0)) %>%
```

```
  select(-dvisits)

bin_model <- glm(bin_visits ~ . , data = binary_data, family = binomial)
summary(bin_model)
```

```
##
## Call:
## glm(formula = bin_visits ~ ., family = binomial, data = binary_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3131  -0.6334  -0.4949  -0.3810   2.5232
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.289901   0.277246  -8.259  < 2e-16 ***
## sex          0.260689   0.082349   3.166 0.001547 **
## age         -1.976087   1.527118  -1.294 0.195666
## agesq        2.736668   1.680658   1.628 0.103455
## income       0.007457   0.127383   0.059 0.953316
## levyplus     0.267007   0.100618   2.654 0.007962 **
## freepoor    -0.680383   0.261069  -2.606 0.009157 **
## freerepa     0.416240   0.139871   2.976 0.002922 **
## illness      0.263485   0.028966   9.096  < 2e-16 ***
## actdays      0.158077   0.011922  13.259  < 2e-16 ***
## hscore       0.063430   0.017405   3.644 0.000268 ***
## chcond1      0.102007   0.091346   1.117 0.264117
## chcond2      0.266798   0.125975   2.118 0.034187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5224.5  on 5189  degrees of freedom
## Residual deviance: 4556.4  on 5177  degrees of freedom
## AIC: 4582.4
##
## Number of Fisher Scoring iterations: 5
```

```
hurdle_model <- pscl::hurdle(dvisits ~. , data = data, dist = "poisson", zero.dist = "binomial", link =
summary(hurdle_model)
```

```
##
## Call:
## pscl::hurdle(formula = dvisits ~ ., data = data, dist = "poisson",
##     zero.dist = "binomial", link = "logit")
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.7084 -0.4363 -0.3343 -0.2551 11.5458
##
## Count model coefficients (truncated poisson with log link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.1738733  0.3122239  -3.760 0.000170 ***
```

```
## sex          0.0004136  0.0900614   0.005 0.996335
## age          3.9598238  1.5979992   2.478 0.013213 *
## agesq       -4.2481265  1.6976740  -2.502 0.012338 *
## income      -0.5177801  0.1529759  -3.385 0.000713 ***
## levyplus    -0.1524321  0.1192158  -1.279 0.201030
## freepoor     0.0350603  0.2643832   0.133 0.894501
## freerepa    -0.4389010  0.1458957  -3.008 0.002627 **
## illness      0.0787933  0.0298629   2.639 0.008327 **
## actdays      0.1143761  0.0071079  16.091  < 2e-16 ***
## hscore       0.0045545  0.0146969   0.310 0.756638
## chcond1      0.0237003  0.1160931   0.204 0.838237
## chcond2     -0.0001276  0.1309931  -0.001 0.999223
## Zero hurdle model coefficients (binomial with logit link):
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.289901   0.277246  -8.259  < 2e-16 ***
## sex          0.260689   0.082349   3.166 0.001547 **
## age         -1.976087   1.527118  -1.294 0.195666
## agesq        2.736668   1.680658   1.628 0.103455
## income       0.007457   0.127383   0.059 0.953316
## levyplus     0.267007   0.100618   2.654 0.007962 **
## freepoor    -0.680383   0.261070  -2.606 0.009157 **
## freerepa     0.416240   0.139871   2.976 0.002922 **
## illness      0.263485   0.028966   9.096  < 2e-16 ***
## actdays      0.158077   0.011922  13.259  < 2e-16 ***
## hscore       0.063430   0.017405   3.644 0.000268 ***
## chcond1      0.102007   0.091346   1.117 0.264117
## chcond2      0.266798   0.125975   2.118 0.034187 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Number of iterations in BFGS optimization: 23
## Log-likelihood: -3213 on 26 Df
```

The binomial part (zero hurdle model) is exactly the same. However, the truncated poisson part (count model) is different. However, if I use pospoisson as the family in VGAM package, the result is consistent with the hurdle model. A possible reason is that the implementation of truncated poisson is somehow different.

Question 6. Model selection

The hurdle model seems to provide a better fit, since it has a smaller AIC as $2 * 26 - 2 * (-3213) = 6478$. The high insurance doesn't seem to increase the number of consultations significantly. However, it's significant that people with levyplus and freerepa will have higher probability of consulations compared with ones using freepoor, since they both have positive slopes and high significance in the zero hurdle model.