Generalized Linear Model

Contents

1 Exponential Family

Define a random variable Y with probability density function $f(y;\theta,\phi)$ where θ is an unknown canonical parameter and ϕ is an known scale parameter. We say that the distribution of Y belongs to exponential family if $f(y;\theta,\phi)$ can be written as

$$f(y;\theta,\phi) = exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)\right)$$

1.1 Properties of the exponential family

For a single observation, its likelihodd function with score and information function are:

$$lf(\theta,\phi;y) = log(f(y;\theta,\phi)) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y;\phi)$$

$$S(\theta) = \frac{\partial l}{\partial \theta} = \frac{y' - b(\theta)}{a(\phi)}$$

$$I(\theta) = -\frac{\partial^2 l}{\partial \theta^2} = \frac{b''(\theta)}{a(\phi)}$$

So we have some general results base on those:

- $E(S(\theta)) = 0$
- $E(I(\theta)) = Var(S(\theta))$
- Mean of the Y : $\mu = E(Y) = b'(\theta)$
- Variance of the Y: $Var(Y) = b''(\theta)a(\phi) = V(\mu)a(\phi)$

1.2 Link Function

Link function is a function that relates the linear model $\eta = X'\beta$ to the mean μ: $g(\mu) = \eta = X'\beta$

1.2.1 Example of Poisson

The pmf of Poisson distribution is:

$$f(y;\lambda) = \frac{\lambda^y e^{-\lambda}}{y!} = e^{(ylog\lambda - \lambda - log(y!))}$$

- $\theta = log\lambda$

- $b(\theta) = e^{\theta}$
- $\mu = E(Y) = b'(\theta) = e^{\theta}$
- $log(\mu) = \eta = X'\beta$

2 Analysis of the Binary Data

2.1 Multiple Logistic Regression

2.1.1 Example of Post-Surgery Drug

Consider a study that is to investigate the effect of a new drug in reducing mortality after an abdominal surgery. Patients were assigned to treatment or control groups randomly. And each surgery was devided into 3 groups: low risk, medium risk and high risk. We summarize the result as follows:

```r
library(kableExtra)

sur <- read.csv('surgical_table.csv', header = F)

kable(sur,align = 'c')%>%
  column_spec(1,bold = T, width = "2.5cm")%>%
  column_spec(2,bold = T, width = "2.5cm")%>%
  column_spec(3, width = "3cm")%>%
  column_spec(4, width = "3cm")%>%
  column_spec(5, width = "3cm")%>%
  row_spec(1,bold = T)
```

V1	**V2**	V3	V4	V5
treatment	**outcome**	**surgical risk: low**	**surgical risk: medium**	**surgical risk: high**
control	**died**	1	2	3
	survived	4	5	6
treatment	**died**	3	4	5
	survived	55	44	33

2.1.2 Model Selection

In this section, we will consider which link function to use and whether there is any interaction in the model.

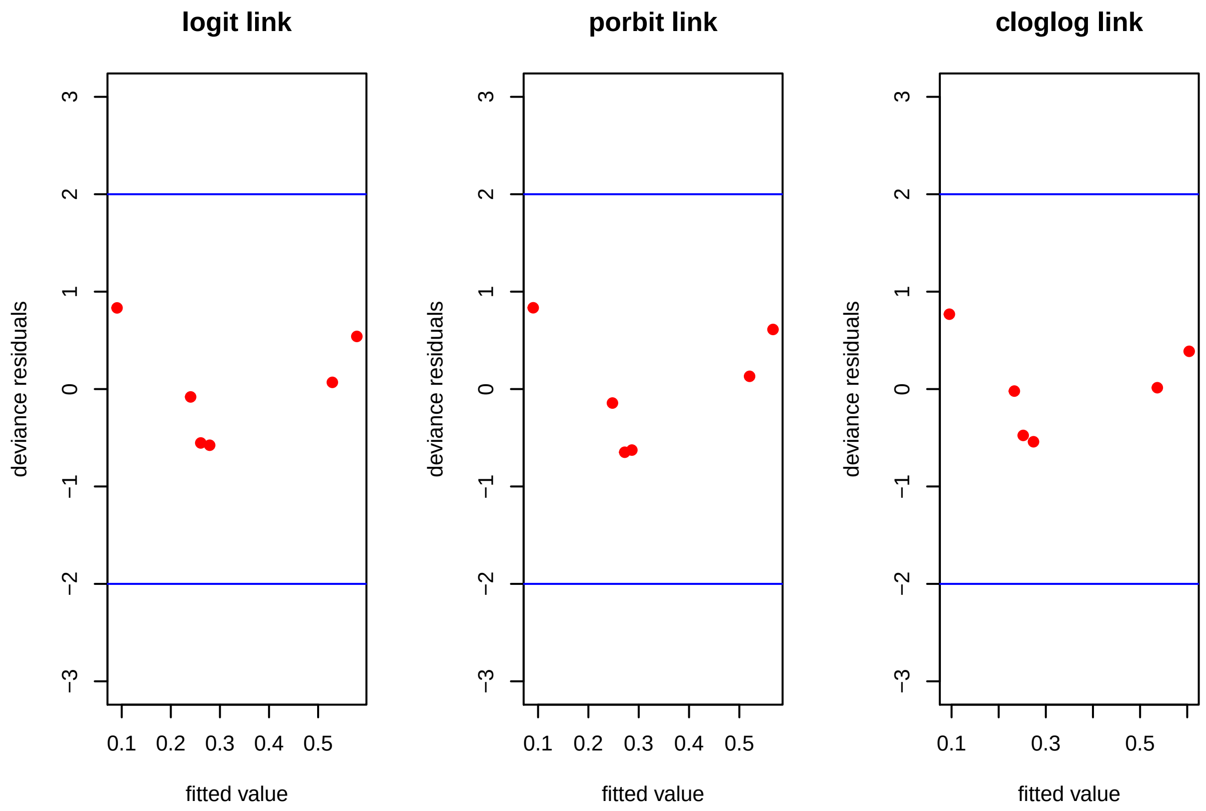
2.1.2.1 Link Functions

First, we will use surgical data to fit several different link functions. And we will choose a best one to fit the data by looking at the plot of 'fitted value vs residuals'.

The link functions that are taken into concerns are:

- Logistic Link: $logit(\pi) = X'\beta$, with estimated probability $\hat{\pi} = \frac{e^{X'\beta}}{1+e^{X'\beta}}$
- Probit Link: $\Phi^{-1}(\pi) = X'\beta$, with estimated probability $\hat{\pi} = \Phi(X'\beta)$, where Φ is the cdf of N(0,1).
- Cloglog Link: $log(-log(1-\pi)) = X'\beta$, with estimated probability $\hat{\pi} = 1 - e^{(-e^{X'\beta})}$

```r
sur <- read.csv('surgical_data.csv', header = T)
sur$risk = relevel(sur$risk,'low')
sur$resp <- cbind(sur$died, sur$survived)


sur.logit <- glm(resp ~ treatment + risk, family= binomial(logit), data = sur)

sur.logit.fv <- sur.logit$fitted.values
sur.logit.dr <- residuals.glm(sur.logit,'deviance')



sur.probit <- glm(resp ~ treatment + risk, family= binomial(probit), data = sur)

sur.probit.fv <- sur.probit$fitted.values
sur.probit.dr <- residuals.glm(sur.probit,'deviance')



sur.cloglog <- glm(resp ~ treatment + risk, family= binomial(cloglog), data = sur)

sur.cloglog.fv <- sur.cloglog$fitted.values
sur.cloglog.dr <- residuals.glm(sur.cloglog,'deviance')

par(mfrow = c(1,3))

plot(sur.logit.fv, sur.logit.dr, main= 'logit link', xlab = 'fitted value', ylab = 'devi
abline(h = c(-2,2), col= 'blue')

plot(sur.probit.fv, sur.probit.dr, main= 'porbit link', xlab = 'fitted value', ylab = 'd
abline(h = c(-2,2), col= 'blue')


plot(sur.cloglog.fv, sur.cloglog.dr, main= 'cloglog link', xlab = 'fitted value', ylab =
abline(h = c(-2,2), col= 'blue')
```



All the models fit the data well. In convenience of the interpretation of the impact of the parameters, we choose logistic regression (ie, logit link) to use.

2.1.2.2 Interaction Detection

Knowing that the the model we fit so far is the main effect model which doesn't involve interactions. But we will double check the assumption we made by doing a deviance test:

- Saturated model: $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x1 + \beta_2 x2 + \beta_3 x3 + \beta_4(x_1 * x_2) + \beta_5(x_1 * x_3)$, where $x_1 = I(teatment = 1)$ and x_2 is the high risk level of the surgery, x_3 is the medium level of the surgery and $(x_1 * x_2)$ and $(x_1 * x_3)$ represent the interaction between the two.
- Main effect model: $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
- H_0: The main effect model is adequate compared to saturated model.
- $\Delta D = D_{main} - D_{saturated} \sim \chi^2_{df=2}$

```
sur.logit.dev <- sur.logit$deviance

sur.logit.main.pv <- 1 - pchisq(sur.logit.dev,2)

sur.logit.main.pv
```

```
## [1] 0.44132
```

From the deviance test we know that the ΔD is 1.6359701 and the corresponding p-value is $0.44132 > 0.05$. So we do not reject the null hypothesis that the main effect model is adequate.

2.1.3 Inference

```
summary(sur.logit)
```

```
##
## Call:
## glm(formula = resp ~ treatment + risk, family = binomial(logit),
##     data = sur)
##
## Deviance Residuals:
##        1         2         3         4         5         6
## -0.55313   0.06879   0.54048   0.83315  -0.08070  -0.57664
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.0410     0.5275  -1.973   0.0484 *
## treatmenttreated  -1.2670     0.5630  -2.251   0.0244 *
## riskhigh           1.3589     0.7151   1.900   0.0574 .
## riskmedium         1.1569     0.6670   1.734   0.0828 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10.928  on 5  degrees of freedom
## Residual deviance:  1.636  on 2  degrees of freedom
## AIC: 25.615
##
## Number of Fisher Scoring iterations: 4
```

Now we get the model of $log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x1 + \beta_2 x2$. And we can make some inference about the effect of the new drug on reducing the mortality of the patients, or we can predict the probability of the mortality of the patients.

2.1.3.1 Effect of Treatment with 95% Confidence Interval

Base on the above model we get, we first test if there is any effect of the new drug. So we will perform an estimate and confidence interval on the treatment effect β_1:

$$log(\frac{\pi_t}{1-\pi_t}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where π_t is the probability of the death of the patients in the treatment group using the new drug.

$$log(\frac{\pi_c}{1-\pi_c}) = \beta_0 + \beta_1 \times 0 + \beta_2 x_2 + \beta_3 x_3$$

where π_c is the probability of the death of the patients in the control group which don't use the new durg, while holding every other factors the same as treatment group.

These two formular minus each other, we get:

$$log(OR) = log\left(\frac{\pi_t/1-\pi_t}{\pi_c/1-\pi_c}\right) = \beta_1$$

$$O.R. = \frac{\pi_t/1-\pi_t}{\pi_c/1-\pi_c} = e^{\beta_1}$$

where O.R. is the Odds Ratio of the death of the patients using the new drugs vs not using the new drugs, while holding any other fators the same. If there is no effect of the new drugs, the Odds Ratio is supposed to be close to one.

```r
sur.beta_1 <- sur.logit$coefficients[2]
sur.OR <- exp(sur.beta_1)
```

Now we know that the Odds Ratio of the death of the patients using the new drugs vs not using the new drugs is 0.2816759, which seems that the new drug reduces the probability of death of the patients after the surgery.

With 95% confidence interval of the Odds Ratio:

$$C.I.of.O.R. = e^{(\beta_1 \pm 1.96 \times se(\beta_1))}$$

```r
sur.beta_1.se <- 0.5630
sur.OR.CI <- exp(sur.beta_1 + c(-1,1)*qnorm(0.975)*sur.beta_1.se)
```

While getting the 95% confidence interval (0.09, 0.85) without 1 included in, we confirmed that the Odds Ratio isn't close to 1, which means that the reduction of the mortality of the patients using this new drug is significant.

2.1.3.2 Probability of Death with 95% Confidence Interval

Also based on the data, we can predict the probability of death of a patients of some specific charactoristic.

For example, consider a patient who reveived a srugery of risk high and used the new drug after the surgery. The probability of the death of her will be:

$$\hat{\pi} = \frac{e^{\beta_0+\beta_1+\beta_2}}{1+e^{\beta_0+\beta_1+\beta_2}}$$

```
sur.beta_0 <- sur.logit$coefficients[1]
sur.beta_2 <- sur.logit$coefficients[2]
sur.pi <- exp(sur.beta_0 + sur.beta_1 +sur.beta_2)/(1+exp(sur.beta_0 + sur.beta_1 +sur.b
```

As calculated above the expected probability of death of "the patient who reveived a srugery of risk high and used the new drug after the surgery" is 2%.

We can also calculate the 95% confidence interval for the probability of death.

We know that the 95% CI for $\beta_0 + \beta_1 + \beta_2$ is:

$$(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2) \pm 1.96 \times se(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2)$$

$$(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2) = [1,1,1,0]\beta = C'\beta$$

$$se(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2) = \sqrt{Var(C'\beta)} = \sqrt{C'Var(\beta)C}$$

```
C <- c(1,1,1,0)

betas.hat <- C%*%sur.logit$coefficients

sur.var.cov <- summary(sur.logit)$cov.unscaled
betas.se <- sqrt(C%*%sur.var.cov%*%C)

betas.CI <- betas.hat +c(-1,1)*qnorm(0.975)*betas.se
```

```
## Warning in c(-1, 1) * qnorm(0.975) * betas.se: Recycling array of length 1 in vector-
##   Use c() or as.vector() instead.
```

```
## Warning in betas.hat + c(-1, 1) * qnorm(0.975) * betas.se: Recycling array of length
##   Use c() or as.vector() instead.
```

$$CI = (\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2) \pm 1.96 \times se(\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2) = (-2.065, 0.167)$$

which includes 2% (0.02) in the interval.

2.1.4 Testing Redundant factors

```r
summary(sur.logit)
```

```
##
## Call:
## glm(formula = resp ~ treatment + risk, family = binomial(logit),
##     data = sur)
##
## Deviance Residuals:
##        1         2         3         4         5         6
## -0.55313   0.06879   0.54048   0.83315  -0.08070  -0.57664
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.0410     0.5275  -1.973   0.0484 *
## treatmenttreated  -1.2670     0.5630  -2.251   0.0244 *
## riskhigh           1.3589     0.7151   1.900   0.0574 .
## riskmedium         1.1569     0.6670   1.734   0.0828 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10.928  on 5  degrees of freedom
## Residual deviance:  1.636  on 2  degrees of freedom
## AIC: 25.615
##
## Number of Fisher Scoring iterations: 4
```

Although we have analysised the probability of mortality or Odds Ratio of different groups. But looking the case in a general way, we noticed from the p-values in R-output that the effect of high or medium risk surgery on the mortality of patients seems to be no different than the effect of low risk surgery.

We can do a formal deviance test to test this hypothesis:

```r
sur.logit2 <- glm(resp ~ treatment, family= binomial(logit), data = sur)

summary(sur.logit2)
```

```
##
## Call:
## glm(formula = resp ~ treatment, family = binomial(logit), data = sur)
##
## Deviance Residuals:
##       1        2        3        4        5        6
## -1.9004   0.7671   1.4146  -0.2952   0.2724   0.0000
```

```
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.2719     0.3318  -0.819   0.4125
## treatmenttreated  -1.1144     0.5373  -2.074   0.0381 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10.9278  on 5  degrees of freedom
## Residual deviance:  6.3623  on 4  degrees of freedom
## AIC: 26.341
##
## Number of Fisher Scoring iterations: 4
```

We first construct a model that doesn't distinguish the differences of the impact of surgical risk on the mortality of patients, which is:

$$log(\frac{\pi_r}{1-\pi_r}) = \beta_0 + \beta_1 x_1$$

where r is for distinguishing this reduced model from the main effect model:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Now the hypotheses are:

$$H_0 : \beta_2 = \beta_3 = 0 \quad vs \quad H_a : at\ least\ one\ of\ them \neq 0$$

Deviance Test becomes:

$$\Delta D = D_0 - D_a \quad \sim \chi^2_2$$

```
D <- sur.logit2$deviance - sur.logit$deviance

1- pchisq(D, sur.logit2$df.residual- sur.logit$df.residual)
```

```
## [1] 0.0941203
```

We get that the $\Delta D \approx 4.73$, and p-value $Pr(\chi^2_2 > 4.73) = 0.09 > 0.05$

Therefore, we do not reject the null hypothesis that there is no differences in the effect of the surgical risk on the mortality of patients. The mortality of patients is only influenced by the drugs in this case. And the model can be reduced to

$$log(\frac{\pi_r}{1-\pi_r}) = \beta_0 + \beta_1 x_1$$

3 Analysis of The Counts (Poisson) Data

3.1 Poisson Log-Linear Regression Model

Remember we have derived the exponential family. For a variable

$$Y \sim Poisson(\mu)$$

$$f(y) = exp(ylog(\mu) - \mu - log(y!))$$

with

$$E(Y) = \mu, \quad Var(Y) = \mu$$

For a time homogeneous poisson process ($\mu = \lambda t$), under the canonical link, we have:

$$log\mu = log\lambda + logt = X'\beta + logt$$

We are still modelling the mean of the response variable, but this time, more exactly, we are modelling the rate of the event that will happen.

And this is the Poisson Log-Linear Regression Model we will be using.

3.1.1 Example of Ship Damage Incidents

McCullagh and Nelder (1989) have used a data set of ship damages in which it records the number of a certain type of damage occurs in cargo ships.

The factors taken into corncern that may related to damages are:

- Ship Type: A, B, C, D, E
- Year of Construction: 1960-1964, 1965-1969, 1970-1974, 1975-1979 (coded as 1, 2, 3, 4)
- Period of Operation: 1960-1974, 1975-1979 (coded as 1 and 2)
- Month: total number of months of operation and construction
- Y: total number of damages occurred during operations.

And the first a few rows of our data looks like:

```r
ship <- read.csv('ship_data.csv', header = T)

ship$cyr <- factor(ship$cyr)
ship$oyr <- factor(ship$oyr)

head(ship, align='c')
```

```
##   type cyr oyr months  y
## 1    A   1   1    127  0
## 2    A   1   2     63  0
## 3    A   2   1   1095  3
## 4    A   2   2   1095  4
## 5    A   3   1   1512  6
## 6    A   3   2   3353 18
```

3.1.2 Model Selection

We first look at the main effect model with their nested models. The method used to compare the models is by doing the deviance test:

- Model1 (Main Effect Model): $type + cyr + oyr + offset(log(months))$

  $$log(\mu_i) = \beta_0+\beta_1(type_2)+\beta_2(type_3)+\beta_3(type_4)+\beta_4(type_5)+\beta_5(cyr_2)+\beta_6(cyr_3)+\beta_7(cyr_4)+\beta_8(oyr_2)+log$$

- Model2 (without type effect): $cyr + oyr + offset(log(months))$
- Model3 (without year of construction effect): $type + oyr + offset(log(months))$
- Model4 (without year of operation effect): $type + cyr + offset(log(months))$

$$Model1 \quad vs \quad Model2$$

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad vs \quad H_a : at\ \ least\ \ one\ \ of\ \ them \neq 0$$

```
model1 <- glm(y ~ type + cyr+ oyr+ offset(log(months)),family=poisson(log), data= ship)

model2 <- glm(y ~ cyr + oyr+ offset(log(months)), family=poisson(log),data = ship)

D21 <- model2$deviance - model1$deviance

pv21 <- 1- pchisq(D21,4)
```

Deviance Statistics becomes:

$$\Delta D = D_2 - D_1 = 23.67 \quad \sim \chi^2_4$$

$$p - value = Pr(\chi^2_4 > \Delta D) = 9.30 \times 10^{-5} << 0.05$$

We reject the null hypothesis that the Model2 is adequate than Model1 at 95% confidence level.

$$Model1 \quad vs \quad Model3$$

$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0 \quad vs \quad H_a : at\ \ least\ \ one\ \ of\ \ them \neq 0$$

```
model3 <- glm(y ~ type + oyr+ offset(log(months)), family=poisson(log),data = ship)

D31 <- model3$deviance - model1$deviance

pv31 <- 1- pchisq(D31,3)
```

Deviance Statistics becomes:

$$\Delta D = D_3 - D_1 = 31.41 \quad \sim \chi^2_3$$

$$p - value = Pr(\chi^2_3 > \Delta D) = 6.97 \times 10^{-7} << 0.05$$

We reject the null hypothesis that the Model3 is adequate than Model1 at a 95% confidence level.

$$Model1 \quad vs \quad Model4$$

$$H_0 : \beta_8 = 0 \quad vs \quad H_a : \beta_8 \neq 0$$

```
model4 <- glm(y ~ type + cyr + offset(log(months)), family=poisson(log),data = ship)

D41 <- model4$deviance - model1$deviance

pv41 <- 1- pchisq(D41,1)
```

Deviance Statistics becomes:

$$\Delta D = D_4 - D_1 = 10.66014 \quad \sim \chi^2_1$$

$$p - value = Pr(\chi^2_1 > \Delta D) = 0.001 < 0.05$$

We reject the null hypothesis that the Model4 is adequate than Model1 at a 95% confidence level.

So Model1 is the best model so far.

3.1.3 Interaction Detection

We will test if there should be interaction added into the mdoel by defining the model as:

- Model5 (with interaction): $type + cyr + oyr + type * cyr + type * oyr + oyr * cyr + offset(log(months))$

Again, we will perform deviance test:

$$Model1 \quad vs \quad Model5$$
$$H_0 : Main\ effect\ model\ is\ adequate \quad vs \quad H_a : Interaction\ model\ is\ adequate$$

```
model5 <- glm(y ~ type + cyr +oyr + type*cyr + type*oyr + cyr*oyr + offset(log(months))

D15 <- model1$deviance - model5$deviance

pv15 <- 1- pchisq(D15,model1$df.residual - model5$df.residual)
```

Deviance Statistics becomes:

$$\Delta D = D_1 - D_5 = 31.8386 \quad \sim \chi^2_{18}$$

$$p - value = Pr(\chi^2_{18} > \Delta D) = 0.023 < 0.05$$

p-value is smaller than 0.05, suggesting that diagnostic test reject the null hypothesis that the Model1 is adequate than Model5 at a 95% confidence level. Interaction model is more adequate.

But looking at the standard errors, we found that thoses are very large:

```
summary(model5)
```

```
##
## Call:
## glm(formula = y ~ type + cyr + oyr + type * cyr + type * oyr +
##     cyr * oyr + offset(log(months)), family = poisson(log), data = ship)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.27162  -0.06844  -0.00003   0.06804   1.29954
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -25.0958 10984.9041  -0.002   0.9982
## typeB          18.0300 10984.9041   0.002   0.9987
## typeC          18.5429 10984.9041   0.002   0.9987
## typeD          -0.6610 15531.1431   0.000   1.0000
## typeE           0.9866 19032.0164   0.000   1.0000
## cyr2           19.2419 10984.9041   0.002   0.9986
## cyr3           19.5426 10984.9041   0.002   0.9986
## cyr4           19.1794 10984.9041   0.002   0.9986
## oyr2            0.5983     0.5130   1.166   0.2435
## typeB:cyr2    -18.4077 10984.9041  -0.002   0.9987
## typeC:cyr2    -19.5621 10984.9041  -0.002   0.9986
## typeD:cyr2    -19.4154 19024.0155  -0.001   0.9992
## typeE:cyr2      0.2506 19032.0164   0.000   1.0000
## typeB:cyr3    -18.6702 10984.9041  -0.002   0.9986
## typeC:cyr3    -18.1464 10984.9041  -0.002   0.9987
## typeD:cyr3      1.0523 15531.1431   0.000   0.9999
## typeE:cyr3     -1.0517 19032.0164   0.000   1.0000
## typeB:cyr4    -18.7985 10984.9041  -0.002   0.9986
## typeC:cyr4    -17.3361 10984.9042  -0.002   0.9987
## typeD:cyr4     -0.3900 15531.1431   0.000   1.0000
## typeE:cyr4     -2.1194 19032.0164   0.000   0.9999
## typeB:oyr2      0.1068     0.4535   0.235   0.8139
## typeC:oyr2     -1.5018     0.7639  -1.966   0.0493 *
## typeD:oyr2      0.1293     0.8721   0.148   0.8821
## typeE:oyr2      0.1556     0.5646   0.276   0.7828
## cyr2:oyr2      -0.3925     0.3031  -1.295   0.1954
## cyr3:oyr2      -0.2645     0.3609  -0.733   0.4636
## cyr4:oyr2           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 146.3283  on 33  degrees of freedom
## Residual deviance:   6.8565  on  7  degrees of freedom
## AIC: 158.72
##
## Number of Fisher Scoring iterations: 18
```

And the correlation matrix of the regression parameter reveals that there exist strong positive or strong negative relationship (close to $+1 \ \ or \ \ -1$).

```
co <- summary(model5,correlation = T)$correlation
co[1:11,1:4]
```

```
##               (Intercept)         typeB         typeC         typeD
## (Intercept)  1.0000000000 -1.000000e+00 -1.000000e+00 -7.072824e-01
## typeB       -0.9999999999  1.000000e+00  1.000000e+00  7.072824e-01
## typeC       -0.9999999979  1.000000e+00  1.000000e+00  7.072824e-01
## typeD       -0.7072823939  7.072824e-01  7.072824e-01  1.000000e+00
## typeE       -0.5771802551  5.771803e-01  5.771803e-01  4.082294e-01
## cyr2        -0.9999999991  1.000000e+00  1.000000e+00  7.072824e-01
## cyr3        -0.9999999994  1.000000e+00  1.000000e+00  7.072824e-01
## cyr4        -0.9999999985  1.000000e+00  1.000000e+00  7.072824e-01
## oyr2         -0.0000221541  1.764375e-05  1.941425e-05  9.149212e-06
## typeB:cyr2    0.9999999990 -1.000000e+00 -1.000000e+00 -7.072824e-01
## typeC:cyr2    0.9999999929 -1.000000e+00 -1.000000e+00 -7.072824e-01
```

This means that the model is over-parameterized:

- For type*cyr:

The reason can be found from the inspectation of the data. If we check the ship type D whose construction year is 1 or 2, we found that there is no damages occurred. While, type*cyr is trying to explain that the effect of the type of ship on the response depends on the year of construction. So the interaction of type and cyr is not significant.

We can also check if the other interactions (type with oyr, and cyr with oyr) seperately are significant or not by doing deviance test:

- For type*oyr:

$$H_0 : the\ type * oyr\ is\ not\ significant\ compared\ to\ main\ effect\ model$$

```
model6 <- glm(y ~ type + cyr+ oyr+type*oyr + offset(log(months)), family = poisson(log)

pv6 <- 1- pchisq(model1$dev - model6$dev, model1$df.residual - model6$df.residual)
```

p-value = 0.29 > 0.05. We do not reject than the type*oyr is not significant compared to main effect model.

- For cyr*oyr:

$$H_0 : the\ cyr * oyr\ is\ not\ significant\ compared\ to\ main\ effect\ model$$

```
model7 <- glm(y ~ type + cyr+ oyr+cyr*oyr + offset(log(months)), family = poisson(log),

pv7 <- 1- pchisq(model1$dev - model7$dev, model1$df.residual - model7$df.residual)
```

p-value = 0.4091268 > 0.05. We do not reject that the cyr*oyr is not significant compared to main effect model.

- In all, we conclude that main effect model is the best fitting model.

3.1.4 Expected Number of Damages

Now we can estimate the expected number of damage that will occur for a specific type of ship.

$$log(\mu) = log(\lambda) + log(t) = X'\beta + log(t)$$

$$Since\ \ Poisson\ \ regression, \quad E(N(t)) = \mu = exp(X'\beta + log(t))$$

```r
betas_poisson <- model1$coefficients
log.lambda <- betas_poisson[1] + betas_poisson[7]
log.t <- log(15*12)
E <- exp(log.lambda + log.t)
```

For example, the expected number of damage for a type A ship built in 1970 and operated during the entire period of 1960-1974 is 0.6740047.

3.2 Analysis of the 3-Way Contingency Table

3.2.1 Example of Pneumoconiosis

Let's see an study of Pneunoconiosis conducted by Ashford and Sowden (1970). They did a random sampling of a group of smoking coal miners to study the relation between age, coughing status, breathing status and a lung disorder called pneumoconiosis. And the data are collected in the following table:

```r
library(kableExtra)
pneu_table <- read.csv('pneu_table.csv',header=F)

kable(pneu_table,align ='c')%>%
  row_spec(1:4,bold =T)%>%
  column_spec(1,bold = T)
```

V1	V2	V3	V4	V5
	Breathless		**Breathless**	
	Yes		**No**	
	Coughed		**Coughed**	
Age	**Yes**	**No**	**Yes**	**No**
20-24	9	7	95	1841
25-29	23	9	105	1654
30-34	54	19	177	1863
35-39	121	48	257	2357
40-44	169	54	273	1778
45-49	269	88	324	1712
50-54	404	117	245	1324
55-59	406	152	225	967
60-64	372	106	132	526

3.2.2 Log-Linear Model Selection

In order not to ignore any potential interaction between age (A), breathlessness (B) and coughing (C), we start the model from the saturated one:

$$log\mu_{ijk} = \mu + \mu_i^A + \mu_j^B + \mu_k^C + \mu_{ij}^{AB} + \mu_{ik}^{AC} + \mu_{jk}^{BC} + \mu_{ijk}^{ABC}$$

$$where \ \ \mu_1^A = \mu_1^B = \mu_1^C = 0$$
$$\mu_{1j}^{AB} = \mu_{j1}^{AB} = \mu_{1k}^{AC} = \mu_{j1}^{AC} = \mu_{1k}^{BC} = \mu_{j1}^{BC} = 0$$

$$\mu_{1jk}^{ABC} = \mu_{i1k}^{ABC} = \mu_{ij1}^{ABC}$$
$$for \ \ i = 1, ...9, \ \ j = 1, 2, \ \ k = 1, 2$$

```
freq = c(9,23,54,121,169,269,404,406,372,7,9,19,48,54,88,117,152,106,95,105,177,257,273,

names = list(
  age = c('20-24', '25-29', '30-34', '35-39', '40-44', '45-49', '50-54', '55-59', '60-64
  coughed = c('yes','no'),
  breathless = c('yes', 'no')
  )

cfb.3d = array(freq, c(9, 2, 2), dimnames= names)

library(plyr)

cfb = count(as.table(cfb.3d))
cfb = cfb[,1:4]
names(cfb) = c('age', 'coughed', 'breathless', 'freq')

cfb$coughed<- relevel(cfb$coughed,'no')
cfb$breathless<- relevel(cfb$breathless,'no')

model.saturated = glm(freq ~age*breathless*coughed,family = poisson,data = cfb)
summary(model.saturated)
```

```
##
## Call:
## glm(formula = freq ~ age * breathless * coughed, family = poisson,
##     data = cfb)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [24]  0  0  0  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                        7.51806    0.02331 322.577  < 2e-16 ***
## age25-29                          -0.10711    0.03388  -3.162  0.00157 ** 
## age30-34                           0.01188    0.03286   0.361  0.71774    
## age35-39                           0.24708    0.03110   7.944 1.96e-15 ***
## age40-44                          -0.03482    0.03325  -1.047  0.29501    
## age45-49                          -0.07265    0.03358  -2.164  0.03049 *  
## age50-54                          -0.32965    0.03603  -9.148  < 2e-16 ***
## age55-59                          -0.64387    0.03972 -16.212  < 2e-16 ***
## age60-64                          -1.25276    0.04944 -25.339  < 2e-16 ***
## breathlessyes                     -5.57215    0.37868 -14.715  < 2e-16 ***
## coughedyes                        -2.96419    0.10521 -28.174  < 2e-16 ***
## age25-29:breathlessyes             0.35843    0.50509   0.710  0.47793    
## age30-34:breathlessyes             0.98665    0.44336   2.225  0.02606 *  
## age35-39:breathlessyes             1.67821    0.40578   4.136 3.54e-05 ***
## age40-44:breathlessyes             2.07789    0.40309   5.155 2.54e-07 ***
## age45-49:breathlessyes             2.60407    0.39414   6.607 3.92e-11 ***
## age50-54:breathlessyes             3.14592    0.39077   8.051 8.24e-16 ***
## age55-59:breathlessyes             3.72184    0.38860   9.577  < 2e-16 ***
## age60-64:breathlessyes             3.97029    0.39336  10.093  < 2e-16 ***
## age25-29:coughedyes                0.20720    0.14559   1.423  0.15471    
## age30-34:coughedyes                0.61039    0.13136   4.647 3.37e-06 ***
## age35-39:coughedyes                0.74812    0.12404   6.031 1.62e-09 ***
## age40-44:coughedyes                1.09041    0.12367   8.817  < 2e-16 ***
## age45-49:coughedyes                1.29951    0.12141  10.704  < 2e-16 ***
## age50-54:coughedyes                1.27703    0.12612  10.125  < 2e-16 ***
## age55-59:coughedyes                1.50609    0.12864  11.708  < 2e-16 ***
## age60-64:coughedyes                1.58169    0.14334  11.035  < 2e-16 ***
## breathlessyes:coughedyes           3.21550    0.51482   6.246 4.21e-10 ***
## age25-29:breathlessyes:coughedyes  0.47976    0.65556   0.732  0.46427    
## age30-34:breathlessyes:coughedyes  0.18284    0.58513   0.312  0.75468    
## age35-39:breathlessyes:coughedyes -0.07484    0.54631  -0.137  0.89103    
## age40-44:breathlessyes:coughedyes -0.20081    0.54194  -0.371  0.71097    
## age45-49:breathlessyes:coughedyes -0.43345    0.53272  -0.814  0.41584    
## age50-54:breathlessyes:coughedyes -0.28911    0.53000  -0.545  0.58542    
## age55-59:breathlessyes:coughedyes -0.77493    0.52873  -1.466  0.14275    
## age60-64:breathlessyes:coughedyes -0.57755    0.53538  -1.079  0.28070    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
```

```
##     Null deviance:  2.5889e+04  on 35  degrees of freedom
## Residual deviance: -7.4118e-13  on  0  degrees of freedom
## AIC: 327.52
##
## Number of Fisher Scoring iterations: 3
```

We fit the saturated model as above. Noticing that the u_{ijk}^{ABC} in the saturated model is insignificant through individual t-test, we may want to check the homogeneous association model:

$$log\mu_{ijk} = \mu + \mu_i^A + \mu_j^B + \mu_k^C + \mu_{ij}^{AB} + \mu_{ik}^{AC} + \mu_{jk}^{BC}$$

```
model.homo = glm(freq ~ age + coughed + breathless + age*coughed + age*breathless + coug
1- pchisq(model.homo$deviance - model.saturated$deviance , model.homo$df.residual - mode
```

```
## [1] 0.0007994639
```

H_0: model.homo is as adequate as the saturated model.

After doing the Deviance-tests, we can find that p-value is $0.0007994639 < 0.05$. So we reject the model.homo is as adequate as saturated model.

So the saturated model is adequate. And that means that there is an association between all three factors. The effect of each individual factor has to associated with all the other two factors together.

3.2.3 Odds Ratio Estimate

Now we can use the data to do lots of estimate. Let's see an example about how to calculate the Odds Ratio.

The OR of coughing among coal miners aged 40-44 who became breathless compared to those were not breathless can be calculated by:

$$log(OR) = log\left(\frac{\pi_{c_1b_1a_{40}}/(\pi_{c_0b_0a_{40}})}{\pi_{c_1b_0a_{40}}/(\pi_{c_0b_0a_{40}})}\right)$$
$$OR = e^{log(OR)}$$

```
or <- exp(model.saturated$coefficients[28]+model.saturated$coefficients[32])
```

OR of coughing among coal miners aged 40-44 who became breathless compared to those were not breathless is 20.38272

4 Overdispersion Analysis

Frequently, the model doesn't fit quite well to the data because of the restriction of the model. For example, a poisson model assume that Var(Y) = E(Y), but the data frequently shows that Var(Y) > E(Y), which we say the data is overdispersed. In order to handle the over dispersion of the data, we use Ad Hoc method or Mixed model.

4.1 Overdispersion for Counts Data

4.1.1 Ad Hoc Method

Recall Poisson Distribution:

$$f(y;\theta,\phi) = exp\Big(ylog\mu - \mu - logy!\Big)$$

$$where \;\; E(Y) = b'(\theta) = \mu$$

$$Var(Y) = b''(\theta)a(\phi) = \mu, \quad \phi = 1, \quad a(\phi) = 1$$

In order to handle the case when Var(Y) > E(Y), we let $\phi \neq 1$, and

$$Var(Y) = \phi\mu$$

We say there exist overdispersion when $D >> n - p$, and we estimate dispersion parameter as: $hat\phi = \frac{D}{n-p}$.

4.1.2 Mixed model Method

We introduce a new random variable μ, with $E(\mu) = 1$, $Var(\mu) = \phi$, making the conditional distribution of Y: $Y_i|\mu_i \sim Poisson(\mu_i\lambda)$.

Interestingly, we will have some new property of the unconditional distribution of Y:

$$E(Y_i) = E(E(Y_i|\mu)) = E(\mu_i\lambda) = \lambda$$

$$Var(Y_i) = Var(E(Y_i|\mu_i)) + E(Var(Y_i|\mu)) = \lambda(\lambda\phi + 1)$$

$$Variance \;\; is \;\; inflated \;\; by \;\; a \;\; factor \;\; of \;\; (\lambda\phi + 1)$$

If we assume

$$\mu \sim Gamma(\alpha, \beta)$$

$$Y_i \sim NB(a = \frac{1}{\phi}, b = \lambda\phi)$$

$$E(Y_i) = ab = \lambda, \quad Var(Y_i) = ab(1 + b) = \lambda(1 + \lambda\phi) \quad still \;\; holds$$

4.1.3 Example of Cystic Brosis

Let's see an example of Cystic Brosis study.

People have Cysitic Brosis often lead to dterioration in lung function. There is a study that gives rhDNase (trt) to patients as treatment group. And patients in the control group receive placebo. Treatment are assigned randomly, and roughly equally for similar days. The original lung function level (fev) and the occurrences of exacerbations (count) during the period of study were recorded.

There will be overdispersion happens in the data, and we want to account for that and draw some useful inference base on that.

```
rhD <- read.table('rhDNase.txt',header =T)
head(rhD)
```

```
##       id trt  fev count time
## 1 493301   1 28.8     0  168
## 2 493303   1 64.0     0  169
## 3 493305   0 67.2     2  168
## 4 493309   1 57.6     0  168
## 5 493310   0 57.6     0  171
## 6 493311   1 25.6     0  166
```

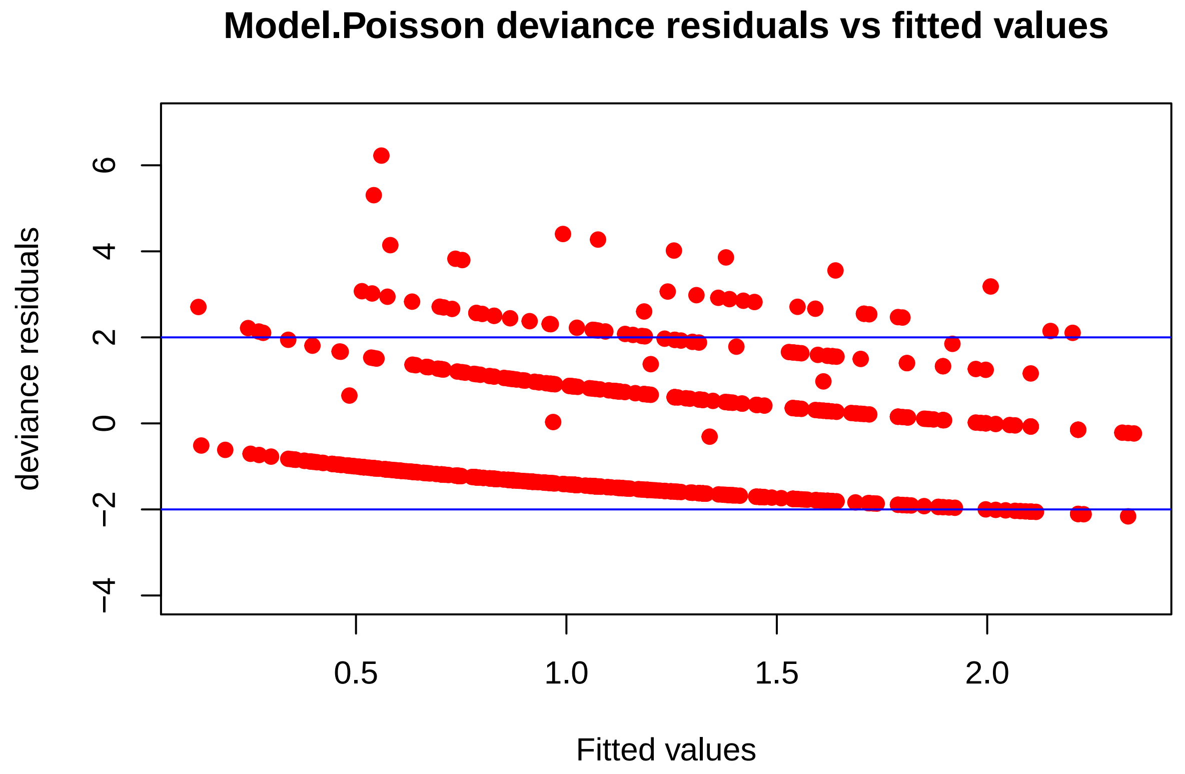
4.1.3.1 Overdispersion Detection

```
model.poisson <- glm(count ~ as.factor(trt) + fev+ offset(log(time)),family = poisson, 

summary(model.poisson)
```

```
##
## Call:
## glm(formula = count ~ as.factor(trt) + fev + offset(log(time)),
##     family = poisson, data = rhD)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1609  -1.4610  -1.0753   0.6093   6.2252
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -3.968839   0.095654 -41.491  < 2e-16 ***
## as.factor(trt)1 -0.253609   0.075140  -3.375 0.000738 ***
## fev             -0.016309   0.001601 -10.187  < 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: 1594.0  on 644  degrees of freedom
## Residual deviance: 1468.5  on 642  degrees of freedom
## AIC: 2179
##
## Number of Fisher Scoring iterations: 6
```

```
fv.poisson <- model.poisson$fitted.values
dr.poisson <- residuals.glm(model.poisson, 'deviance')

plot(fv.poisson,dr.poisson, main = 'Model.Poisson deviance residuals vs fitted values',
abline(h = -2,col='blue')
abline(h = 2,col='blue')
```



We fit the ordinary poisson log-linear model to the data, then find the deviance residuals plot is not well behaved. There is pattern exists in the plot, and no constant variance exists. This suggest that the model doesn't count some variance that it should do.

```
mean_var <- data.frame(
   Group = c('Overall', 'Placebo Group', 'rhDNase Group'),
  Mean = c(mean(rhD$count),mean(rhD$count[rhD$trt==0]),mean(rhD$count[rhD$trt==1])),

  Variance = c(var(rhD$count),var(rhD$count[rhD$trt==0]),var(rhD$count[rhD$trt==1]))
  )


kable(mean_var, align= 'c')%>%
  column_spec(1:3, width='4 cm')%>%
  column_spec(1, bold=T)%>%
  column_spec(2, background = 'green')%>%
  column_spec(3, background = 'red')
```

Group	Mean	Variance
Overall	1.1178295	3.001623
Placebo Group	1.2592593	3.313382
rhDNase Group	0.9750779	2.655627

$$E(Y) << Var(Y)$$

Noticing that the actual variance of the response is larger than the mean of the response. It's clear that there exists overdispersion.

4.1.3.2 Ad Hoc Method

We estimate the $\hat{\phi}$ and adjust variance to be $Var(Y) = Var(Y)\hat{\phi}$:

```
phi = model.poisson$dev / model.poisson$df.residual
```

$\hat{\phi} = 2.287424$. When estimating any confidence interval or using the variance, we just need to use $Var(Y)' = 2.287424 * Var(Y)$

4.1.3.3 Mixed Model Medthod

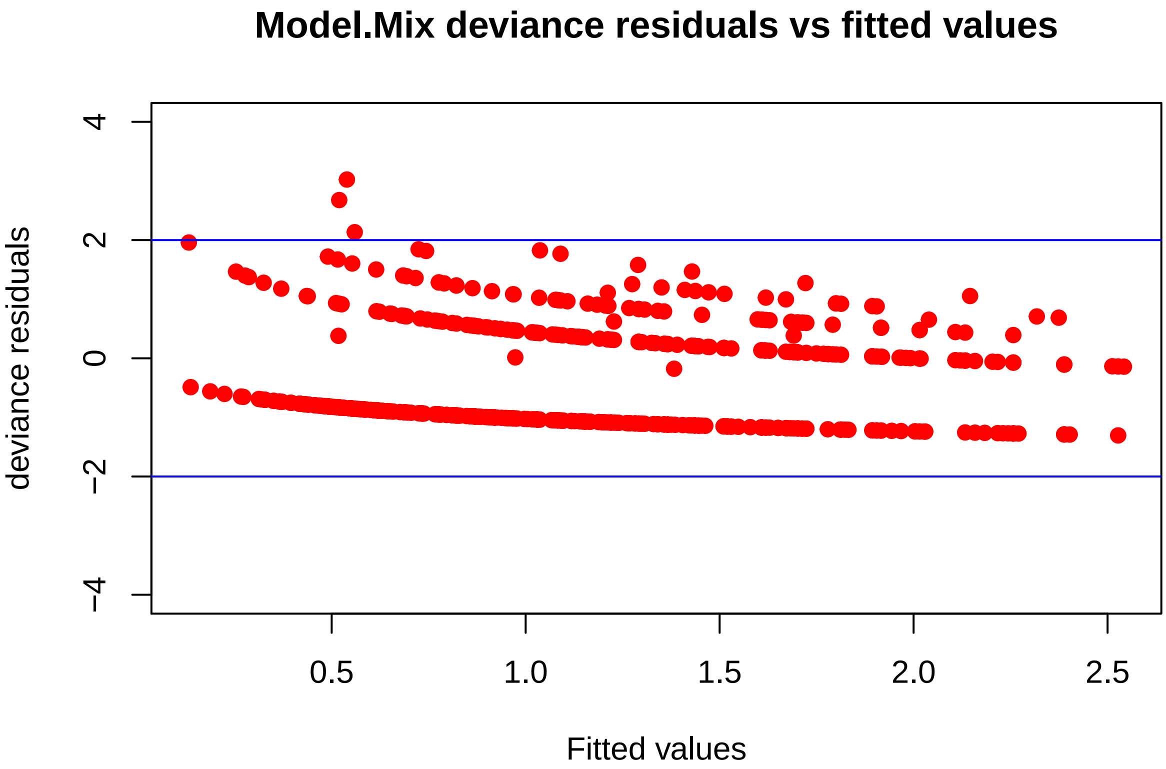
Since we can assume $\mu \sim Gamma(1, \phi)$, $Y \sim NB(\frac{1}{\phi}, \lambda\phi)$.

```
library(MASS)

model.mix = glm.nb(count ~ trt +fev + offset(log(time)), link = log,init.theta=1, data =

summary(model.mix)
```

```
##
## Call:
## glm.nb(formula = count ~ trt + fev + offset(log(time)), data = rhD,
##     init.theta = 0.4507489332, link = log)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3047  -1.0504  -0.8512   0.2774   3.0239
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.863457   0.190680 -20.261  < 2e-16 ***
## trt         -0.275605   0.141964  -1.941   0.0522 .
## fev         -0.017671   0.002835  -6.233 4.59e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.4507) family taken to be 1)
##
##     Null deviance: 595.02  on 644  degrees of freedom
## Residual deviance: 553.86  on 642  degrees of freedom
## AIC: 1808.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.4507
##           Std. Err.:  0.0531
##
##  2 x log-likelihood:  -1800.9170
```

```
dr.m <- residuals.glm(model.mix,'deviance')
fv.m <- model.mix$fitted.values
plot(fv.m,dr.m, main = 'Model.Mix deviance residuals vs fitted values', xlab = 'Fitted v
abline(h = -2,col='blue')
abline(h = 2,col='blue')
```



Now, we fit the GLM based on Y follows a negative binomial distribution. The deviance residuals are well distributed, which means that we have include the information of the overdispersed data now.