Analysis of the `quine` Dataset Using Gamma Regression, Negative Binomial and Log Normal Models

Tsu-Hao Fu

1. Gamma regression

The quine dataset was utilized to fit a Gamma regression model, with the number of days absent from school during the year (Days) serving as the response variable and a log link employed. To determine the maximal model, interactions were included up to the third order, while the null model was considered the minimal model. To account for zero counts, a small constant was added to Days when it equaled zero, with values of 0.01, 0.05, and 0.1 each used. By comparing the additive model of 0.01, 0.05, and 0.1, we can see that there is small difference between the residual deviances and AICs. Therefore, I decided to choose 0.1 which has lower residual deviance to handle zero counts and proceed the analysis.

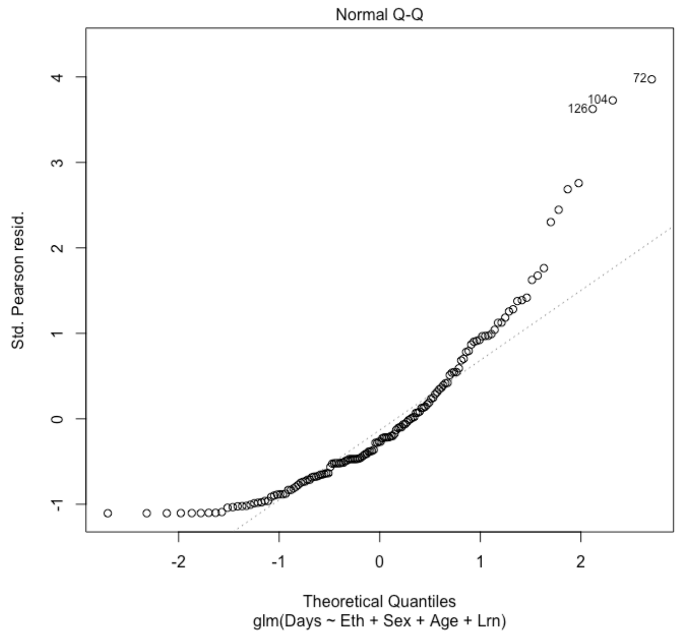
```
Call:
glm(formula = Days ~ Eth + Sex + Age + Lrn, family = Gamma(log)
    data = quine_01)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7063  -0.8256  -0.2631   0.3384   2.0119

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.91075    0.22811  12.760  < 2e-16 ***
EthN        -0.57258    0.15313  -3.739 0.000269 ***
SexM         0.07250    0.15956   0.454 0.650262
AgeF1       -0.45479    0.23793  -1.911 0.058005 .
AgeF2        0.07939    0.23654   0.336 0.737658
AgeF3        0.35200    0.24870   1.415 0.159192
LrnSL        0.28202    0.18501   1.524 0.129700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.8505192)

    Null deviance: 234.04  on 145  degrees of freedom
Residual deviance: 210.78  on 139  degrees of freedom
AIC: 1101.2

Number of Fisher Scoring iterations: 9
```

```
Call:
glm(formula = Days ~ Eth + Sex + Age + Lrn, family = Gamma(log),
    data = quine_05)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2435  -0.8255  -0.2632   0.3384   2.0109

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.91064    0.22800  12.766  < 2e-16 ***
EthN        -0.57224    0.15305  -3.739 0.000269 ***
SexM         0.07275    0.15948   0.456 0.648961
AgeF1       -0.45464    0.23781  -1.912 0.057963 .
AgeF2        0.07933    0.23642   0.336 0.737714
AgeF3        0.35186    0.24857   1.416 0.159156
LrnSL        0.28209    0.18492   1.525 0.129416
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.8496672)

    Null deviance: 205.11  on 145  degrees of freedom
Residual deviance: 181.87  on 139  degrees of freedom
AIC: 1104.3

Number of Fisher Scoring iterations: 8
```

```
Call:
glm(formula = Days ~ Eth + Sex + Age + Lrn, family = Gamma(log),
    data = quine_1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.0228  -0.8254  -0.2634   0.3385   2.0096

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.91046    0.22786  12.773  < 2e-16 ***
EthN        -0.57180    0.15296  -3.738  0.00027 ***
SexM         0.07311    0.15938   0.459  0.64717
AgeF1       -0.45448    0.23766  -1.912  0.05789 .
AgeF2        0.07925    0.23627   0.335  0.73782
AgeF3        0.35173    0.24842   1.416  0.15905
LrnSL        0.28220    0.18480   1.527  0.12902
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.848599)

    Null deviance: 192.69  on 145  degrees of freedom
Residual deviance: 169.47  on 139  degrees of freedom
AIC: 1104.6

Number of Fisher Scoring iterations: 8
```

A Stepwise AIC procedure was then used to fit a model in each case, with the multiple of the number of degrees of freedom for the penalty set to Log(146) due to the large size of our dataset. After employing a Stepwise search based on AIC, the selected model was the Additive model, consisting of Days predicted by Eth, Sex, Age, and Lrn. Therefore, our minimal model is as same as the maximal model. However, the Q-Q plot and Chi-square test shows that the residuals have a deviation from normality.

```
> qu.gm1.step = step(qu.gm1,scope=list(lower=~.,upper=~.^3), k=log(146))
Start:  AIC=1125.51
Days ~ Eth + Sex + Age + Lrn

          Df Deviance    AIC
<none>         169.47 1125.5
+ Eth:Lrn  1   166.83 1127.4
+ Sex:Age  3   158.86 1128.0
+ Eth:Sex  1   168.77 1129.7
+ Sex:Lrn  1   169.33 1130.3
+ Eth:Age  3   160.98 1130.5
+ Age:Lrn  2   168.37 1134.2
> 1  - pchisq(deviance(qu.gm1),  qu.gm1$df.resid)
[1] 0.04021188
```



2. Negative binomial & Log-normal

Upon comparing the Gamma regression model to the Negative Binomial model after applying StepAIC(), it becomes evident that the Gamma regression model is a more suitable fit for this dataset. Although the Negative Binomial model has a lower Deviance than the fitted Gamma model, it contains significantly more terms than the Gamma regression model. As a result, the interpretation of the fitted Negative Binomial Model is much more complicated than the relatively straightforward interpretation of the fitted Gamma regression model. Given this complexity and the minor reduction in Deviance, it was determined that the Gamma regression model is a better fit for the data.

```
glm.nb(formula = Days ~ Eth + Sex + Age + Lrn + Eth:Sex + Eth:Lrn +
    Sex:Age + Sex:Lrn + Eth:Sex:Lrn, data = quine, init.theta = 1.597990735,
    link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8950  -0.8827  -0.2299   0.5669   2.1071

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      3.01919    0.29706  10.163  < 2e-16 ***
EthN            -0.07312    0.26539  -0.276 0.782908
SexM            -0.47541    0.39550  -1.202 0.229355
AgeF1           -0.70887    0.32321  -2.193 0.028290 *
AgeF2           -0.61486    0.37141  -1.655 0.097826 .
AgeF3           -0.34235    0.32717  -1.046 0.295388
LrnSL            0.94358    0.32246   2.926 0.003432 **
EthN:SexM       -0.60586    0.36896  -1.642 0.100572
EthN:LrnSL      -1.35849    0.37719  -3.602 0.000316 ***
SexM:AgeF1      -0.01486    0.46225  -0.032 0.974353
SexM:AgeF2       1.24328    0.46134   2.695 0.007040 **
SexM:AgeF3       1.49319    0.45337   3.294 0.000989 ***
SexM:LrnSL      -0.70467    0.46536  -1.514 0.129966
EthN:SexM:LrnSL  2.11991    0.58056   3.651 0.000261 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.598) family taken to be 1)

    Null deviance: 234.56  on 145  degrees of freedom
Residual deviance: 167.56  on 132  degrees of freedom
AIC: 1093

Number of Fisher Scoring iterations: 1


              Theta:  1.598
          Std. Err.:  0.213

 2 x log-likelihood:  -1063.025
```

```
Call:
lm(formula = log(Days) ~ Eth + Sex + Lrn + Eth:Sex + Eth:Lrn +
    Sex:Lrn + Eth:Sex:Lrn, data = quine_1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3782 -0.5415  0.2162  0.9312  2.7757

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       2.0756     0.3238   6.411 2.13e-09
EthN              0.1005     0.4468   0.225   0.8224
SexM              0.8378     0.4468   1.875   0.0629
LrnSL             0.8361     0.4579   1.826   0.0700
EthN:SexM        -1.5554     0.6205  -2.507   0.0133
EthN:LrnSL       -1.6189     0.6319  -2.562   0.0115
SexM:LrnSL       -1.2203     0.7097  -1.719   0.0878
EthN:SexM:LrnSL   2.4184     0.9680   2.498   0.0136

(Intercept)     ***
EthN
SexM            .
LrnSL           .
EthN:SexM       *
EthN:LrnSL      *
SexM:LrnSL      .
EthN:SexM:LrnSL *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.411 on 138 degrees of freedom
Multiple R-squared:  0.1496,    Adjusted R-squared:  0.1065
F-statistic: 3.468 on 7 and 138 DF,  p-value: 0.001877
```

After comparing the fitted Gamma regression and fitted Log-Normal models, it is evident that the Gamma regression model is the better choice. This is because the Deviance of the final fitted Log-Normal regression model is significantly higher than that of the fitted Gamma regression model. Moreover, the final fitted Log-Normal model contains a substantially greater number of terms than the Gamma regression model, making it more challenging to interpret. As a result, it was determined that the Gamma regression model provides the best fit for the given data.