MAST30027 Assignment 2

Name: Lang (Ron) Chen

Student ID: 1181506

Subject: MAST30027

Assignment number: 2

Tutorial time and tutor: Tues 11:00 Yidi Deng

Investigating Factors Affecting Fertility Rate of Indian Women in Fiji

This dataset contains information about the number of children and number of mothers in each combination of {marriage duration, place of residence, education level} for the Indian race in Fiji. In total, there are 70 groups (observations) of data - 70 combination of 5 marriage duration levels, 3 residence levels and 4 education levels with two combinations having 0 mothers and thereby 0 children.

The variables nChildren and nMother are jointly target variables, whilst marriage duration, residence and education are predictive variables which are treatment levels (ordinal/nominal categorical variables). To investigate the factors and two way interaction of factors affecting fertility rate of mothers (number of children per mother) in each family group, the response variable used will be taken to be nChildren/nMother for each group.

```
library(faraway)
setwd("/Users/tg.chenny/Desktop/1. University/1. Undergraduate/20. Modern Applied Statistics/Asmt/Asmt 2")

data = read.table(file='assignment2_prob1.txt', header=TRUE)

data$duration = factor(data$duration, levels = c("0-4", "5-9", "10-14", "15-19", "20-24", "25-29"), ordered=TRUE)
data$residence = factor(data$residence, levels = c("Suva", "urban", "rural"))
data$education = factor(data$education, levels=c("none", "lower", "upper", "sec+"))

# number of observations
(size = dim(data)[1])

## [1] 70

ftable(xtabs(cbind(nChildren, nMother) ~ duration + residence + education, data))

##                               nChildren nMother
## duration residence education
## 0-4      Suva      none               4       8
##                    lower             24      21
##                    upper             38      42
##                    sec+              37      51
##          urban     none              14      12
##                    lower             23      27
##                    upper             41      39
##                    sec+              35      51
##          rural     none              60      62
##                    lower             98     102
##                    upper            104     107
##                    sec+              35      47
```

```
## 5-9      Suva      none              31      10
##                    lower             80      30
##                    upper             49      24
##                    sec+              38      22
##          urban     none              59      13
##                    lower             98      37
##                    upper            118      44
##                    sec+              48      21
##          rural     none             171      70
##                    lower            317     117
##                    upper            200      81
##                    sec+              47      21
## 10-14    Suva      none              49      12
##                    lower             99      27
##                    upper             58      20
##                    sec+              24      12
##          urban     none              75      18
##                    lower            143      43
##                    upper            105      29
##                    sec+              50      15
##          rural     none             364      88
##                    lower            546     132
##                    upper            197      50
##                    sec+              30       9
## 15-19    Suva      none              59      14
##                    lower            153      31
##                    upper             41      13
##                    sec+              11       4
##          urban     none             108      23
##                    lower            225      42
##                    upper             92      20
##                    sec+              19       5
##          rural     none             577     114
##                    lower            481      86
##                    upper            135      30
##                    sec+               2       1
## 20-24    Suva      none             118      21
##                    lower             91      18
##                    upper             47      12
##                    sec+              13       5
##          urban     none             118      22
##                    lower            147      25
##                    upper             65      13
##                    sec+              16       3
##          rural     none             756     117
##                    lower            431      68
##                    upper            132      23
##                    sec+               5       2
## 25-29    Suva      none             310      47
##                    lower            182      27
```

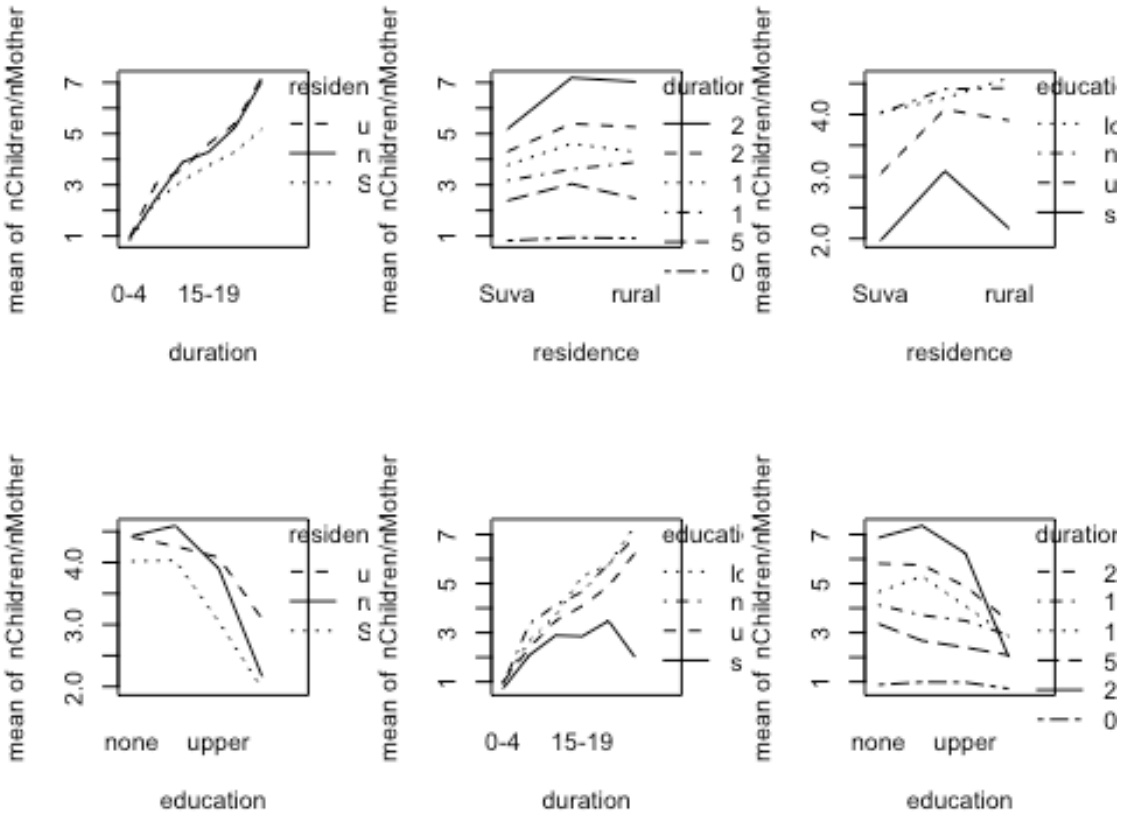
```
##                   upper              43       8
##                   sec+               2        1
##          urban    none             300       46
##                   lower            338       45
##                   upper             98       13
##                   sec+               0        0
##          rural    none            1459      195
##                   lower            461       59
##                   upper             58       10
##                   sec+               0        0
```

Given the response variable is a rate, the poisson regression (with logarithmic link function) was chosen because its sole parameter is precisely the rate (also being the mean of its distribution). In Poisson regression, the log of rate parameter lambda will be predicted by the linear combination of predictive variables.

Although there are 0s in nMother in the display of the table above, this will not cause an log(0) issue as those 2 rows are only displayed due to the need of representing every combination, and won't actually be inputted to the regression.

We begin by plotting the interaction tables to witness whether there exists interaction between the factors.

```
# plot interaction plots
par(mfrow = c(2, 3))
with(data, interaction.plot(duration, residence, nChildren/nMother))
with(data, interaction.plot(residence, duration, nChildren/nMother))
with(data, interaction.plot(residence, education, nChildren/nMother))
with(data, interaction.plot(education, residence, nChildren/nMother))
with(data, interaction.plot(duration, education, nChildren/nMother))
with(data, interaction.plot(education, duration, nChildren/nMother))
```



Observing the 6 plots, in every interaction plot except for duration and residence, there exists at least one line that crosses over others, suggesting the potential for statistically significance of interaction terms. However, as only one line (thus one level) of the variable typically cross over, whether the interaction term for these pairs of predictive variables are overall statistically significant will need to be evaluated using anova tests.

On a side note, although 3-way interaction is not required, it is not valid regardless because to do because the model residuals will run out of degrees of freedom (i.e. become a saturated model).

```
# interaction model - poisson

## the use of 'offset' is effectively the same as regressing with nChildren/n
Mother as the predictor variable
imodel = glm(nChildren ~ offset(log(nMother)) + (duration + residence + educa
tion)^2, data = data, family=poisson())
summary(imodel)

##
## Call:
## glm(formula = nChildren ~ offset(log(nMother)) + (duration +
##     residence + education)^2, family = poisson(), data = data)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -1.7572  -0.3222   0.0414   0.3298   2.8134
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.262560   0.054120  23.329  < 2e-16 ***
## duration.L                     1.322461   0.109693  12.056  < 2e-16 ***
## duration.Q                    -0.475204   0.099868  -4.758 1.95e-06 ***
## duration.C                     0.310979   0.090042   3.454 0.000553 ***
## duration^4                    -0.123519   0.082325  -1.500 0.133514
## duration^5                     0.003130   0.077310   0.040 0.967704
## residenceurban                 0.004121   0.066846   0.062 0.950846
## residencerural                 0.048692   0.054980   0.886 0.375822
## educationlower                -0.015048   0.064926  -0.232 0.816718
## educationupper                -0.284101   0.081056  -3.505 0.000457 ***
## educationsec+                 -0.665426   0.152905  -4.352 1.35e-05 ***
## duration.L:residenceurban      0.147030   0.109403   1.344 0.178971
## duration.Q:residenceurban     -0.101429   0.096908  -1.047 0.295260
## duration.C:residenceurban      0.049790   0.090883   0.548 0.583798
## duration^4:residenceurban     -0.059840   0.086231  -0.694 0.487714
## duration^5:residenceurban      0.084682   0.082494   1.027 0.304646
## duration.L:residencerural      0.232160   0.094578   2.455 0.014100 *
## duration.Q:residencerural     -0.112487   0.084271  -1.335 0.181937
## duration.C:residencerural     -0.038218   0.078852  -0.485 0.627904
## duration^4:residencerural      0.020052   0.075060   0.267 0.789356
## duration^5:residencerural     -0.037891   0.072443  -0.523 0.600943
## duration.L:educationlower      0.063735   0.093908   0.679 0.497332
## duration.Q:educationlower      0.020680   0.087169   0.237 0.812474
## duration.C:educationlower     -0.048863   0.076118  -0.642 0.520921
## duration^4:educationlower      0.074274   0.065747   1.130 0.258605
## duration^5:educationlower      0.091940   0.057318   1.604 0.108704
## duration.L:educationupper     -0.066616   0.102487  -0.650 0.515696
## duration.Q:educationupper      0.103240   0.096634   1.068 0.285355
## duration.C:educationupper     -0.033646   0.086988  -0.387 0.698916
## duration^4:educationupper      0.080111   0.078622   1.019 0.308232
## duration^5:educationupper     -0.025175   0.073140  -0.344 0.730700
## duration.L:educationsec+      -0.481404   0.444798  -1.082 0.279120
## duration.Q:educationsec+      -0.310273   0.410113  -0.757 0.449317
## duration.C:educationsec+      -0.161468   0.299016  -0.540 0.589197
## duration^4:educationsec+      -0.042075   0.198420  -0.212 0.832068
## duration^5:educationsec+      -0.043235   0.157360  -0.275 0.783506
## residenceurban:educationlower  0.014568   0.078828   0.185 0.853377
## residencerural:educationlower  0.036396   0.066889   0.544 0.586350
## residenceurban:educationupper  0.258773   0.099801   2.593 0.009517 **
## residencerural:educationupper  0.201583   0.089264   2.258 0.023928 *
## residenceurban:educationsec+   0.318915   0.144496   2.207 0.027308 *
## residencerural:educationsec+   0.244863   0.147421   1.661 0.096717 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3731.852  on 69  degrees of freedom
## Residual deviance:   30.856  on 28  degrees of freedom
## AIC: 544.33
##
## Number of Fisher Scoring iterations: 4
```

```
anova(imodel, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: nChildren
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                    69     3731.9
## duration             5   3565.8        64      166.1 < 2.2e-16 ***
## residence            2     45.4        62      120.7 1.391e-10 ***
## education            3     50.0        59       70.7 7.930e-11 ***
## duration:residence  10     13.5        49       57.1   0.19551
## duration:education  15     14.5        34       42.7   0.48923
## residence:education  6     11.8        28       30.9   0.06669 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the summary table, it can be seen that not many levels' coefficients are statistically significant at 5% significance level; thus suggesting this model is too complex. The ANOVA table using chi-square test show that none of the three interaction terms are significant, and should be removed. Although we could just build another additive model, an easy way to simplify the model is to do step-wise selection of variables using AIC, which should produce the same result.

```
# we store results as amodel because that was what was found after stepwise w
as performed
amodel = step(imodel)
```

```
## Start:  AIC=544.33
## nChildren ~ offset(log(nMother)) + (duration + residence + education)^2
##
##                       Df Deviance    AIC
## - duration:education  15   44.311 527.79
## - duration:residence  10   44.523 538.00
## - residence:education  6   42.652 544.13
## <none>                    30.856 544.33
##
```

```
## Step:  AIC=527.79
## nChildren ~ duration + residence + education + duration:residence +
##     residence:education + offset(log(nMother))
##
##                       Df Deviance    AIC
## - duration:residence  10   59.921 523.40
## <none>                     44.311 527.79
## - residence:education  6   57.135 528.61
##
## Step:  AIC=523.4
## nChildren ~ duration + residence + education + residence:education +
##     offset(log(nMother))
##
##                       Df Deviance     AIC
## - residence:education  6    70.67  522.14
## <none>                     59.92  523.40
## - duration             5  2625.89 3079.36
##
## Step:  AIC=522.14
## nChildren ~ duration + residence + education + offset(log(nMother))
##
##             Df Deviance     AIC
## <none>           70.67  522.14
## - residence  2   100.19  547.67
## - education  3   120.68  566.16
## - duration   5  2646.49 3087.97
```

We find that the 2-way interaction terms have not been selected as significant features/predictive variables according to the AIC criteria, and hence will continue our analysis based on the additive model.

```
summary(amodel)
```

```
##
## Call:
## glm(formula = nChildren ~ duration + residence + education +
##     offset(log(nMother)), family = poisson(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2960  -0.6641   0.0725   0.6336   3.6782
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.17314    0.03054  38.415  < 2e-16 ***
## duration.L      1.49288    0.03387  44.082  < 2e-16 ***
## duration.Q     -0.52726    0.03026 -17.424  < 2e-16 ***
## duration.C      0.25258    0.02776   9.098  < 2e-16 ***
## duration^4     -0.07613    0.02570  -2.962 0.003059 **
## duration^5      0.03025    0.02402   1.259 0.207880
```

```
## residenceurban  0.11242    0.03250   3.459 0.000541 ***
## residencerural  0.15166    0.02833   5.353 8.63e-08 ***
## educationlower  0.02297    0.02266   1.014 0.310597
## educationupper -0.10127    0.03099  -3.268 0.001082 **
## educationsec+  -0.31015    0.05521  -5.618 1.94e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3731.852  on 69  degrees of freedom
## Residual deviance:   70.665  on 59  degrees of freedom
## AIC: 522.14
##
## Number of Fisher Scoring iterations: 4
```

```
anova(amodel, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: nChildren
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         69     3731.9
## duration   5   3565.8        64      166.1 < 2.2e-16 ***
## residence  2     45.4        62      120.7 1.391e-10 ***
## education  3     50.0        59       70.7 7.930e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the anova table, all three variables are significant, with marriage duration being the strongest.

We now want to check deviance of the poisson variable to make sure it is not significantly different from 1, or else a quasipoisson model will be required.

```
# additive model - check deviance
n = dim(data)[1]
(phihat = sum(residuals(amodel, type = "pearson")^2)/amodel$df.residual)
```

```
## [1] 1.212432
```

It seems like 1.21 is close enough to 1, but we will still fit a quasipoisson model to ensure our inference and testing of significance of regression coefficients are not overoptimistic.

```r
# additive model - quasipoisson
amodel_quasi = glm(nChildren ~ offset(log(nMother)) + duration + residence +
education, data = data, family=quasipoisson())
summary(amodel_quasi)
```

```
##
## Call:
## glm(formula = nChildren ~ offset(log(nMother)) + duration + residence +
##     education, family = quasipoisson(), data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2960  -0.6641   0.0725   0.6336   3.6782
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.17314    0.03363  34.888  < 2e-16 ***
## duration.L      1.49288    0.03729  40.035  < 2e-16 ***
## duration.Q     -0.52726    0.03332 -15.824  < 2e-16 ***
## duration.C      0.25258    0.03057   8.263 1.97e-11 ***
## duration^4     -0.07613    0.02830  -2.690  0.00928 **
## duration^5      0.03025    0.02644   1.144  0.25734
## residenceurban  0.11242    0.03578   3.142  0.00263 **
## residencerural  0.15166    0.03119   4.862 8.98e-06 ***
## educationlower  0.02297    0.02495   0.921  0.36087
## educationupper -0.10127    0.03412  -2.968  0.00432 **
## educationsec+  -0.31015    0.06079  -5.102 3.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.212432)
##
##     Null deviance: 3731.852  on 69  degrees of freedom
## Residual deviance:   70.665  on 59  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

However, quasideviance (larger Standard Errors) does not really change the significance level of any of the factor levels according to Wald test; and there is still one level of duration and one level of education which is not statistically significant at 0.05.

```r
anova(amodel_quasi, test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model: quasipoisson, link: log
##
## Response: nChildren
##
## Terms added sequentially (first to last)
```

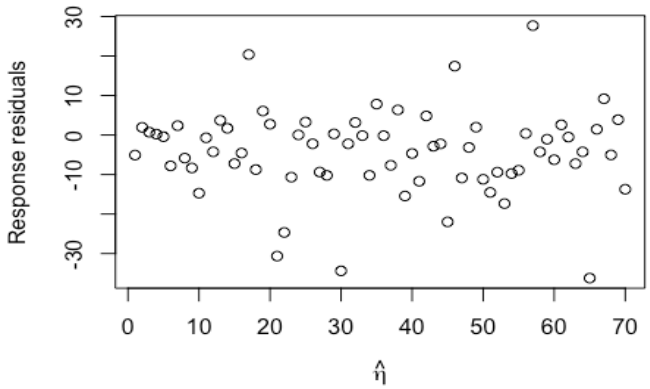
```
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         69     3731.9
## duration   5   3565.8        64      166.1 < 2.2e-16 ***
## residence  2     45.4        62      120.7 7.420e-09 ***
## education  3     50.0        59       70.7 5.782e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nonetheless, even with taking into account deviance, which gives larger standard errors to the coefficient estimates, all 3 variables were still significant.

The quasi deviance two-way interaction models will not need to be checked because in the regular poisson model which had smaller standard errors for regression coefficients, they were already insignificant; so for the quasi-poisson model they could not possibly be significant.

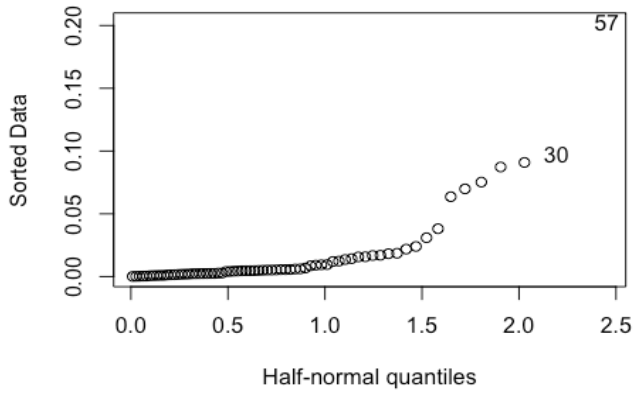
Now we check diagnostic plots:

```
plot(residuals(amodel_quasi, type='response') - predict(amodel_quasi, type='l
ink'), xlab=expression(hat(eta)), ylab="Response residuals")
```



The eta to residual plot show no significant trend, so it is fair to assume the residuals are independent and random, with no deviance present.

```
halfnorm(cooks.distance(amodel_quasi))
```



Observing the cooks distance plot, observation 57 looks to be a kink so has potential to be highly influential on all fitted means; thus, removing point 57 and re-fitting the poisson regression was tested.

First, we re-fit the interaction quasipoisson model and take anova

```
data2 = data[-c(57),]
imodel_quasi2 = glm(nChildren ~ offset(log(nMother)) + (duration + residence
+ education)^2, data = data2, family=quasipoisson())
anova(imodel_quasi2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: quasipoisson, link: log
##
## Response: nChildren
##
## Terms added sequentially (first to last)
##
##
##                     Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                    68     3568.4
## duration             5   3412.2        63      156.3 < 2.2e-16 ***
## residence            2     38.9        61      117.4 3.844e-08 ***
## education            3     49.8        58       67.6 1.746e-09 ***
## duration:residence  10     11.3        48       56.3    0.4484
## duration:education  15     13.7        33       42.7    0.6777
## residence:education  6     12.0        27       30.6    0.1027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see that the two-way interaction terms are still insignificant.

Now we re-fit the additive quasipoisson model and take anova.

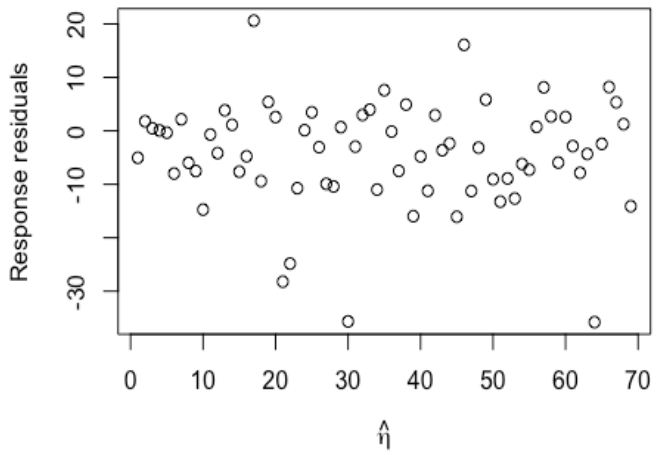
```
amodel_quasi2 = glm(nChildren ~ offset(log(nMother)) + duration + residence +
 education, data = data2, family=quasipoisson())
anova(amodel_quasi2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: quasipoisson, link: log
##
## Response: nChildren
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         68     3568.4
## duration   5   3412.2        63      156.3 < 2.2e-16 ***
## residence  2     38.9        61      117.4 7.434e-08 ***
## education  3     49.8        58       67.6 3.986e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

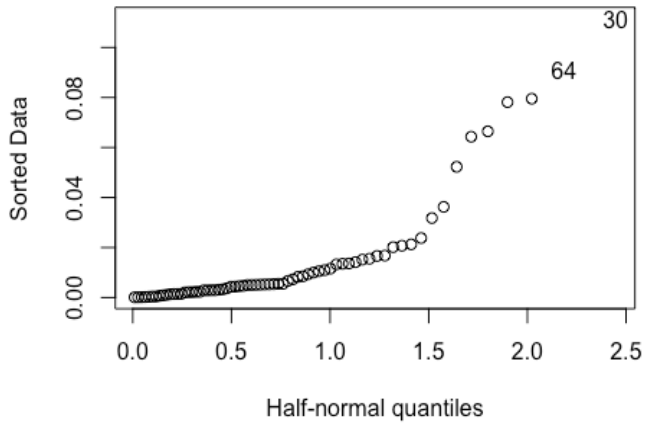
All three variables are still statistically significant at 5% significance level.

Now we re-evaluate the eta to residual and cook's distance plots.

```
plot(residuals(amodel_quasi2, type='response') - predict(amodel_quasi2, type=
'link'), xlab=expression(hat(eta)), ylab="Response residuals")
```




```
halfnorm(cooks.distance(amodel_quasi2))
```



From the plots, the cooks distance looks to have less kinks, and the residuals are still uncorrelated and demonstrates no deviance.

Thus, using a poisson regression model to fit the data, it is concluded that all three factors of **marriage duration**, **residence** and **education** levels are related to the fertility rate of Indian women in Fiji.