We will first load the data into working environment. The table below displays the number of children and mothers grouped by duration of mother's marriage, residence of families and education level of mothers.

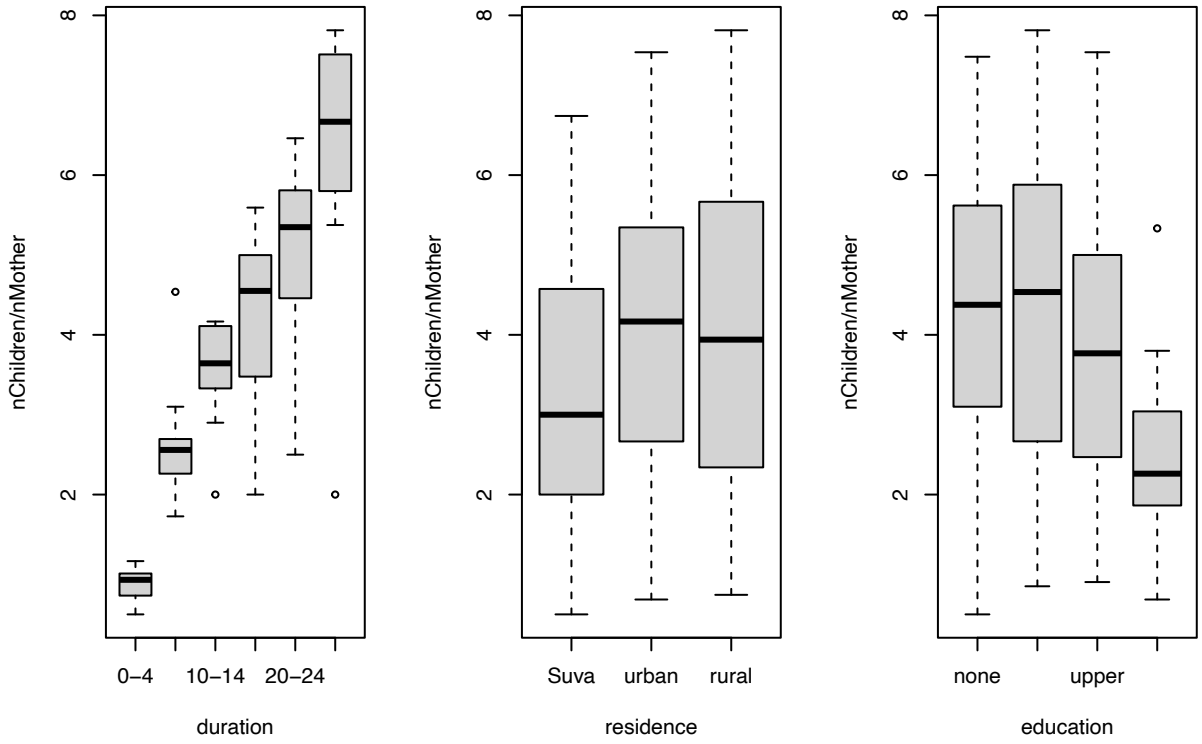
```
data <- read.table(file ="assignment2_prob1.txt", header=TRUE)
data$duration <- factor(data$duration,
                        levels=c("0-4","5-9","10-14","15-19","20-24","25-29"),
                        ordered=TRUE)
data$residence <- factor(data$residence,
                         levels=c("Suva", "urban", "rural"))
data$education <- factor(data$education,
                         levels=c("none", "lower", "upper", "sec+"))
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
```

Note that marriage *duration* is a blocking factor as more children are born over time. We can make the assumption by looking at the plots:

```
par(mfrow = c(1,3))
plot(nChildren/nMother ~ duration, data)
plot(nChildren/nMother ~ residence, data)
plot(nChildren/nMother ~ education, data)
```



The box plots illustrates that the fertility rates does vary across groups of marriage *duration*, hence, concretes the assumption of the blocking factor.

We consider the number of children born as the response with mother's marriage duration, level of education and residence of family as discrete predictors. From this assumption, the straightforward way is to fit an addictive model using Poisson regression with identity link, as we are analysing the rate of children given birth by each mother.

Instead of treating the fertility rate ($nChildren/nMother$) as the response, $nChildren$ is kept as the response, plus we have another term as offset for log of number of mother. This can be explained by:

$$log(fertility\_rate_i) = \beta_0 + \beta_1 \times duration + \beta_2 \times residence + \beta_3 \times education$$

$$log(\frac{nChildren}{nMother}) = \beta_0 + \beta_1 \times duration + \beta_2 \times residence + \beta_3 \times education$$

$$log(nChildren) = log(nMother) + \beta_0 + \beta_1 \times duration + \beta_2 \times residence + \beta_3 \times education$$

The model is fitted as below:

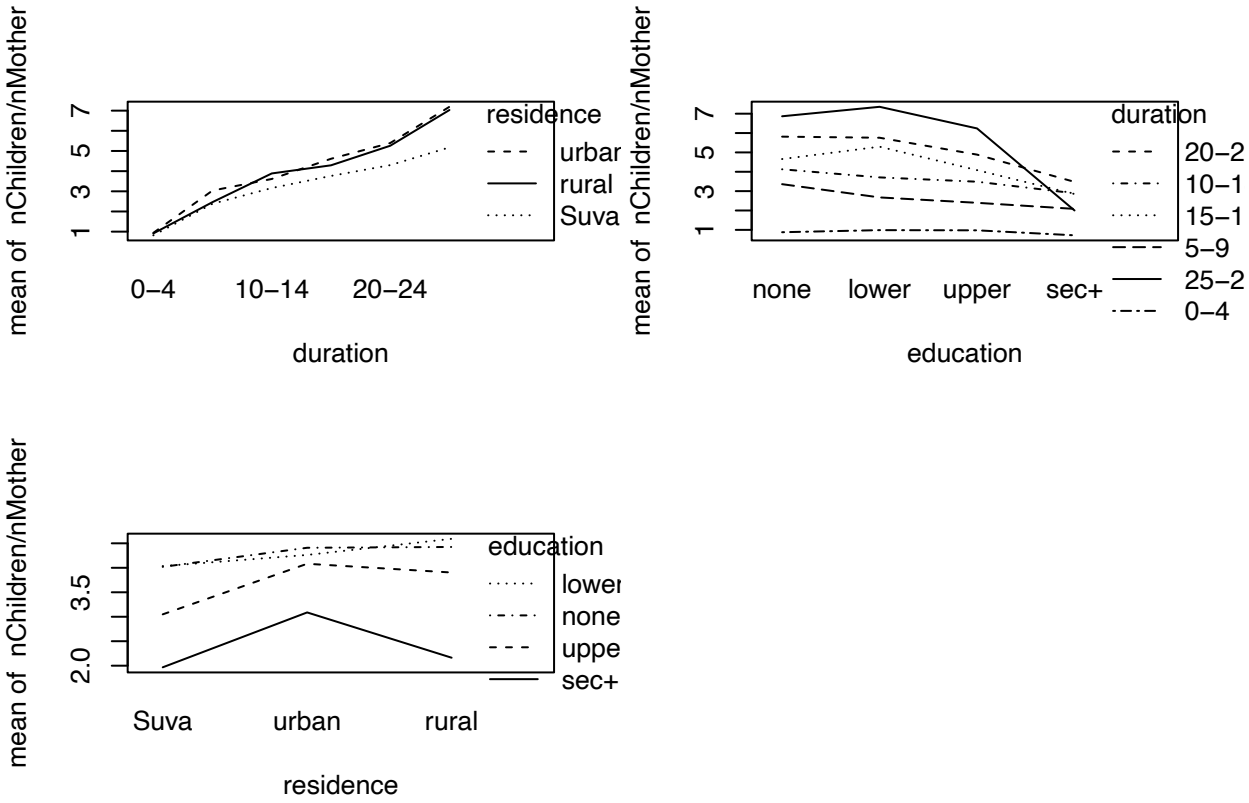
```
addictive <- glm(nChildren ~ offset(log(nMother)) + duration + residence + education,
                 family = poisson, data)
summary(addictive)
```

```
##
## Call:
## glm(formula = nChildren ~ offset(log(nMother)) + duration + residence +
##     education, 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
```

We now plot interaction between predictors to check whether our assumption is valid:

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



Although we have considered marriage *duration* as blocking factor, the plots show otherwise as there might be interactions between *education* and *duration* factors on the fertility rate. Apart from that, we can clearly observe there is another interaction between *education* and *residence*. We need to verify this by constructing a model with interaction between predictors.

```
full.model <- glm(nChildren ~ offset(log(nMother)) + (duration + residence + education)^2,
                  family = poisson, data)
summary(full.model)
```

```
##
## 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
```

Comparing to the *addictive* model with residual deviance of 70.665 on 59 degrees of freedom, the interactive model $full.model$ has residual deviance of 30.856 on 28 df, which shows that it is reasonable to take two-way interactions into account.

To test whether the interaction between *education* and *duration* affects the model, we fit a reduced model where the interaction between the two predictors is not included, and perform a likelihood ratio test between the two models. The result is shown as below:

```
reduced.interaction.edu.dur = glm(nChildren ~ offset(log(nMother)) + (residence + education)^2 + (reside
                family = poisson, data)
anova(reduced.interaction.edu.dur, full.model, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: nChildren ~ offset(log(nMother)) + (residence + education)^2 +
##     (residence + duration)^2
## Model 2: nChildren ~ offset(log(nMother)) + (duration + residence + education)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        43     44.311
## 2        28     30.856 15   13.455   0.5672
```

With p-value = 0.5672, it is clearly that the interaction between *education* and marriage *duration* has negligible effect on the model, which verifies our aforementioned assumption.

Now we perform model selection from the full model using AIC. We will set the direction as both ways so at every step, the algorithm will try to drop or re-add any term to the current model and compute AIC scores. The selection steps are done as follows:

```
final.model = step(full.model, scope=~.)
```

```
## 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
## + duration:education  15   30.856 544.33
##
## 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:residence  10    44.31  527.79
## + duration:education  15    44.52  538.00
## - 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:education    6    59.92  523.40
## + duration:residence    10    57.13  528.61
## + duration:education    15    54.80  536.28
## - residence              2   100.19  547.67
## - education              3   120.68  566.16
## - duration               5  2646.49 3087.97
```

From this model selection, we can observe that dropping all interaction from the full model will reduce AIC score to lowest, thus, conclude our final model as the addictive model without any two-way interaction.

Finally, we will check for overdispersion of the final model. We estimate the dispersion parameter by dividing the Pearson's χ^2-statistics by the degrees of freedom:

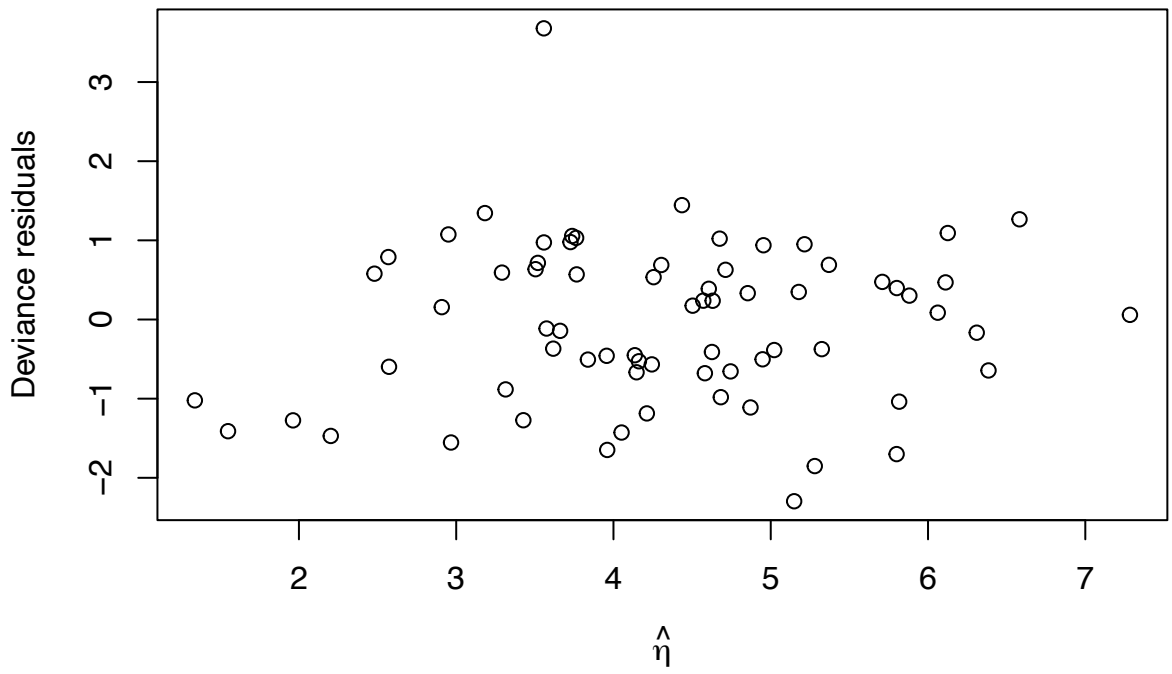
```
sum(residuals(final.model, type="pearson") ^ 2) / final.model$df.residual
```

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

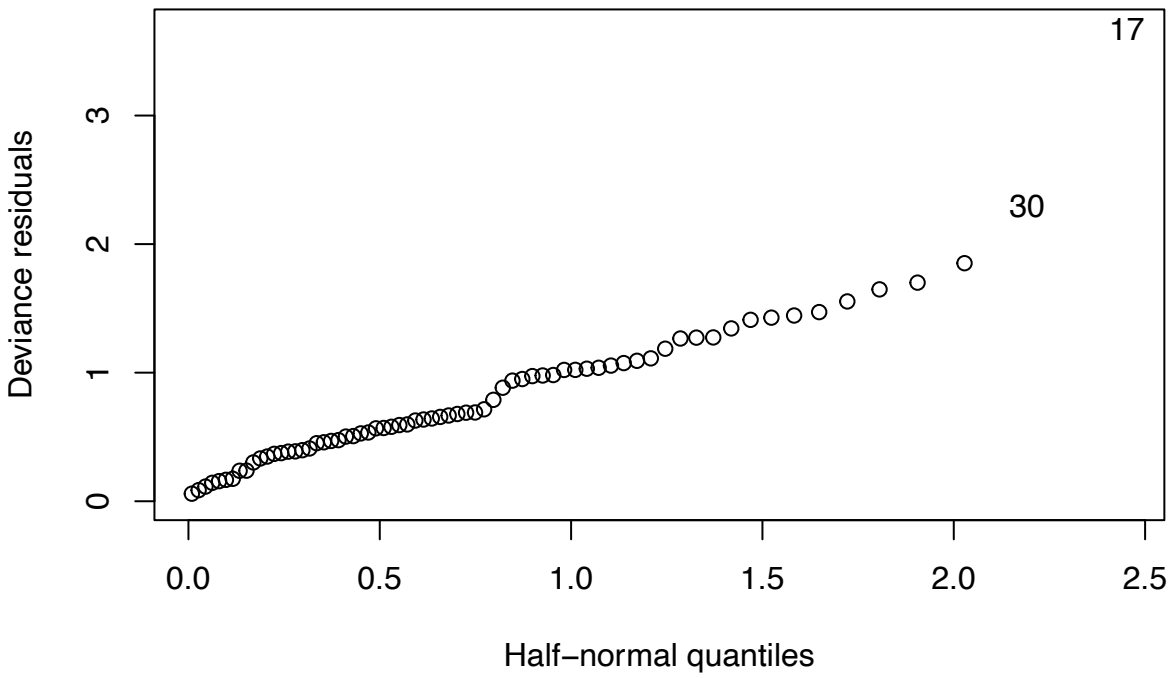
This result suggests that this model does not suffer from overdispersion.

Diagnostic plots:

```
res <- residuals(final.model)
eta.hat <- predict(final.model,type="link")
plot(res ~ eta.hat, xlab=expression(hat(eta)), ylab="Deviance residuals")
```




```
halfnorm(res, ylab="Deviance residuals")
```



We can see from diagnostic plots that there is no extreme fitted values nor abnormal case for the quantile plot except for the 17th and 30th data.

In conclusion, all factors in the data set affect the number of children given birth but there is no two-way interaction significance between any factors.