MAST30027_Assignment2

Zi Ng

23/09/2020

Question 1

Read the data

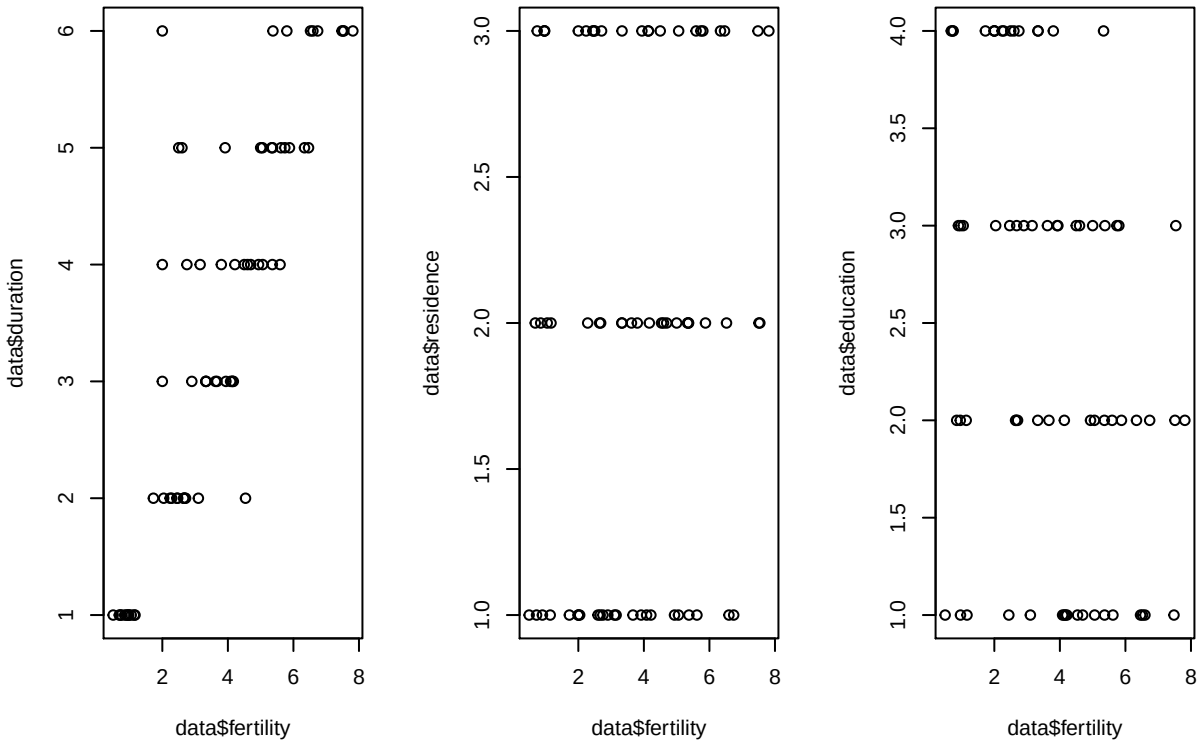
```
data <- read.table(file ="assignment2_prob1.txt", header=TRUE)
data$duration <- factor(data$duration,
                        levels=c("0-4","5-9","10-14","15-19","20-24","25-29"),
                        ordered=TRUE)
data$residence <- factor(data$residence, levels=c("Suva", "urban", "rural"))
data$education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

```
##                                  nChildren      nMother    fertility
## duration residence education
## 0-4      Suva      none          4.0000000    8.0000000    0.5000000
##                    lower        24.0000000   21.0000000    1.1428571
##                    upper        38.0000000   42.0000000    0.9047619
##                    sec+         37.0000000   51.0000000    0.7254902
##          urban     none         14.0000000   12.0000000    1.1666667
##                    lower        23.0000000   27.0000000    0.8518519
##                    upper        41.0000000   39.0000000    1.0512821
##                    sec+         35.0000000   51.0000000    0.6862745
##          rural     none         60.0000000   62.0000000    0.9677419
##                    lower        98.0000000  102.0000000    0.9607843
##                    upper       104.0000000  107.0000000    0.9719626
##                    sec+         35.0000000   47.0000000    0.7446809
## 5-9      Suva      none         31.0000000   10.0000000    3.1000000
##                    lower        80.0000000   30.0000000    2.6666667
##                    upper        49.0000000   24.0000000    2.0416667
##                    sec+         38.0000000   22.0000000    1.7272727
##          urban     none         59.0000000   13.0000000    4.5384615
##                    lower        98.0000000   37.0000000    2.6486486
##                    upper       118.0000000   44.0000000    2.6818182
##                    sec+         48.0000000   21.0000000    2.2857143
##          rural     none        171.0000000   70.0000000    2.4428571
##                    lower       317.0000000  117.0000000    2.7094017
##                    upper       200.0000000   81.0000000    2.4691358
##                    sec+         47.0000000   21.0000000    2.2380952
## 10-14    Suva      none         49.0000000   12.0000000    4.0833333
##                    lower        99.0000000   27.0000000    3.6666667
##                    upper        58.0000000   20.0000000    2.9000000
##                    sec+         24.0000000   12.0000000    2.0000000
```

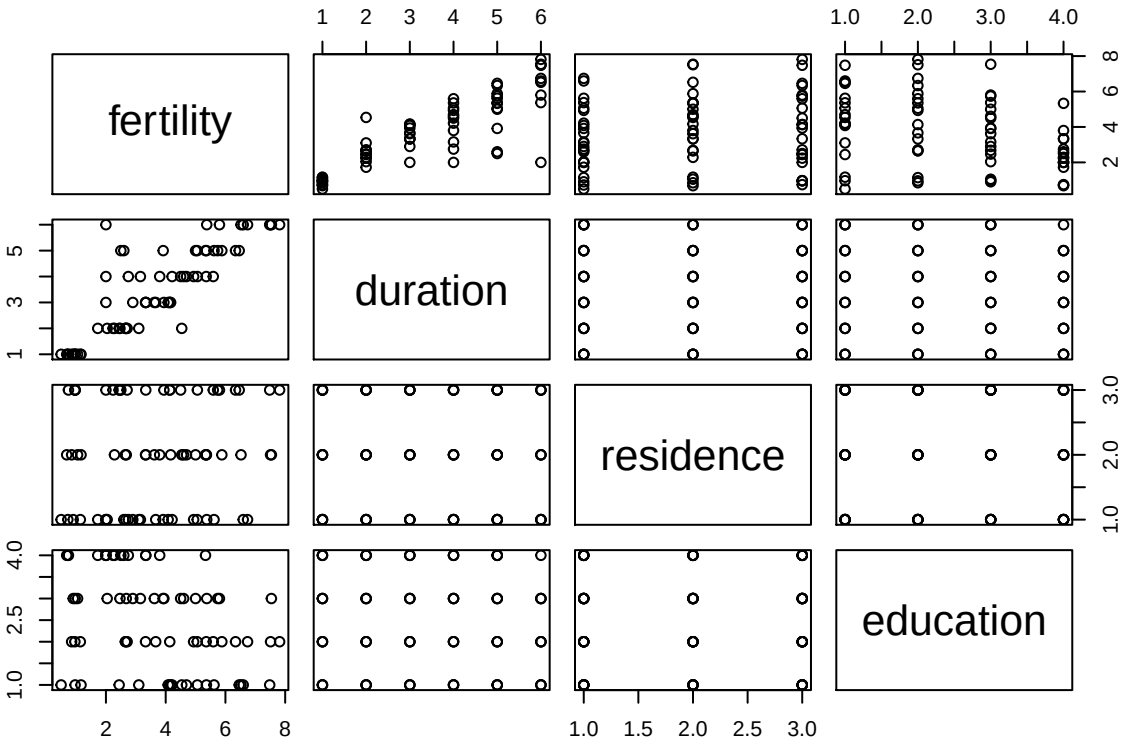
```
##          urban     none         75.0000000   18.0000000    4.1666667
##                    lower       143.0000000   43.0000000    3.3255814
##                    upper       105.0000000   29.0000000    3.6206897
##                    sec+         50.0000000   15.0000000    3.3333333
##          rural     none        364.0000000   88.0000000    4.1363636
##                    lower       546.0000000  132.0000000    4.1363636
##                    upper       197.0000000   50.0000000    3.9400000
##                    sec+         30.0000000    9.0000000    3.3333333
## 15-19    Suva      none         59.0000000   14.0000000    4.2142857
##                    lower       153.0000000   31.0000000    4.9354839
##                    upper        41.0000000   13.0000000    3.1538462
##                    sec+         11.0000000    4.0000000    2.7500000
##          urban     none        108.0000000   23.0000000    4.6956522
##                    lower       225.0000000   42.0000000    5.3571429
##                    upper        92.0000000   20.0000000    4.6000000
##                    sec+         19.0000000    5.0000000    3.8000000
##          rural     none        577.0000000  114.0000000    5.0614035
##                    lower       481.0000000   86.0000000    5.5930233
##                    upper       135.0000000   30.0000000    4.5000000
##                    sec+          2.0000000    1.0000000    2.0000000
## 20-24    Suva      none        118.0000000   21.0000000    5.6190476
##                    lower        91.0000000   18.0000000    5.0555556
##                    upper        47.0000000   12.0000000    3.9166667
##                    sec+         13.0000000    5.0000000    2.6000000
##          urban     none        118.0000000   22.0000000    5.3636364
##                    lower       147.0000000   25.0000000    5.8800000
##                    upper        65.0000000   13.0000000    5.0000000
##                    sec+         16.0000000    3.0000000    5.3333333
##          rural     none        756.0000000  117.0000000    6.4615385
##                    lower       431.0000000   68.0000000    6.3382353
##                    upper       132.0000000   23.0000000    5.7391304
##                    sec+          5.0000000    2.0000000    2.5000000
## 25-29    Suva      none        310.0000000   47.0000000    6.5957447
##                    lower       182.0000000   27.0000000    6.7407407
##                    upper        43.0000000    8.0000000    5.3750000
##                    sec+          2.0000000    1.0000000    2.0000000
##          urban     none        300.0000000   46.0000000    6.5217391
##                    lower       338.0000000   45.0000000    7.5111111
##                    upper        98.0000000   13.0000000    7.5384615
##                    sec+          0.0000000    0.0000000    0.0000000
##          rural     none       1459.0000000  195.0000000    7.4820513
##                    lower       461.0000000   59.0000000    7.8135593
##                    upper        58.0000000   10.0000000    5.8000000
##                    sec+          0.0000000    0.0000000    0.0000000
```

Data visualization

```
par(mfrow=c(1,3))
plot(data$fertility, data$duration)
plot(data$fertility, data$residence)
plot(data$fertility, data$education)
```




```
with(data, pairs(fertility ~ duration + residence + education))
```

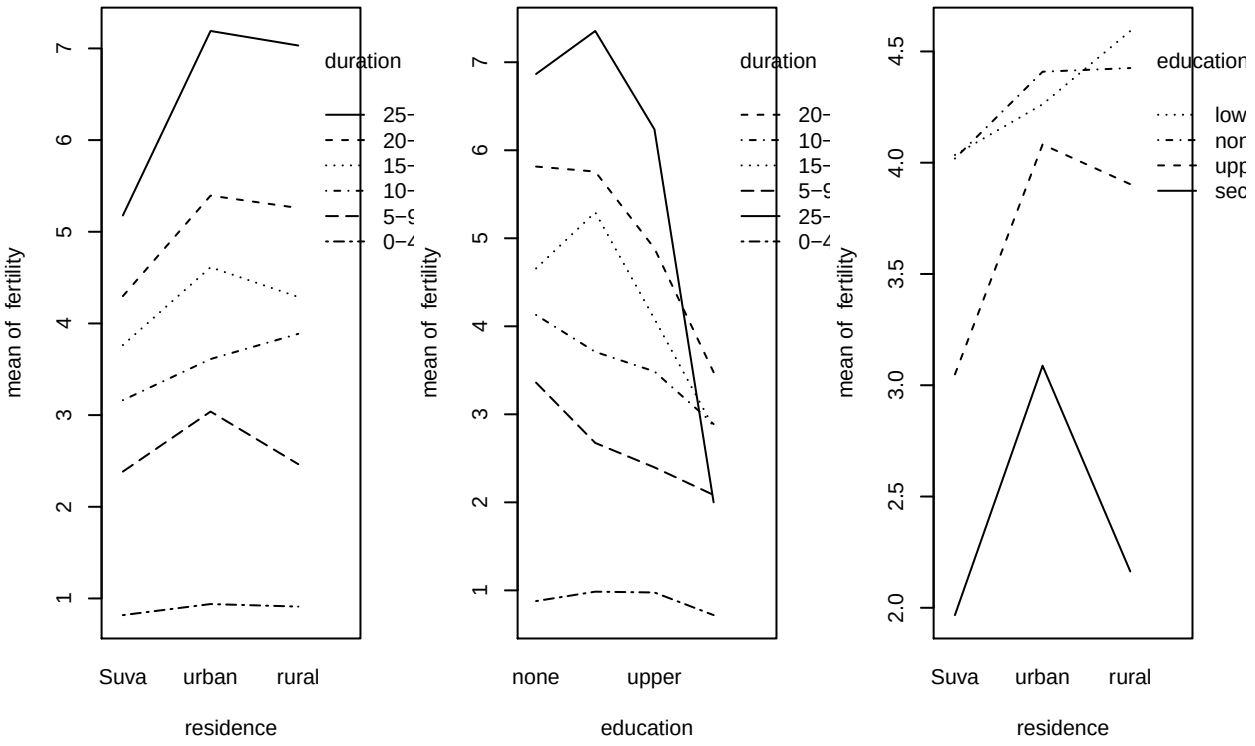


There are 70 women with 3 variables in the dataset. The relationship between the number of children and the duration, residence and education level or the women is of interest. Since the number of children a woman has is count data, it makes sense to fit a Poisson model.

Check pairwise relationships There seems to be interaction between the variables since the slope of the lines

vary.

```
par(mfrow=c(1,3))
with(data, interaction.plot(residence, duration, fertility))
with(data, interaction.plot(education, duration, fertility))
with(data, interaction.plot(residence, education, fertility))
```



Fit a Poisson model

```
model = glm(nChildren ~ offset(log(nMother)) + duration + residence + education + duration*residence + d
summary(model)
```

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

Stepwise model selection based on the AIC

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

```
## Start:  AIC=544.33
## nChildren ~ offset(log(nMother)) + duration + residence + education +
##     duration * residence + duration * education + education *
##     residence
##
##                       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
```

```
summary(model.step)
```

```
##
## 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
```

Significance of interaction

```
anova(model, test="Chi")
```

```
## 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
```

```
anova(model.step, test="Chi")
```

```
## 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
```

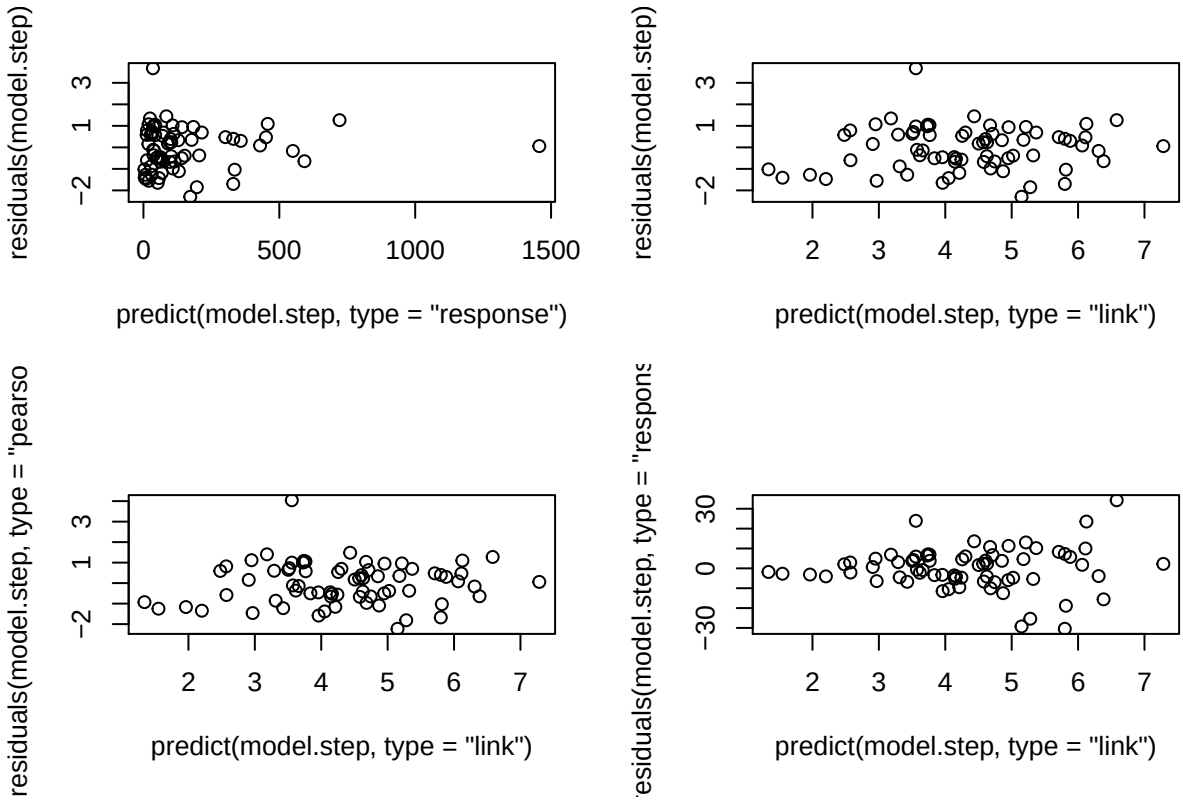
Interaction between the variables is not significant.

Checking linearity

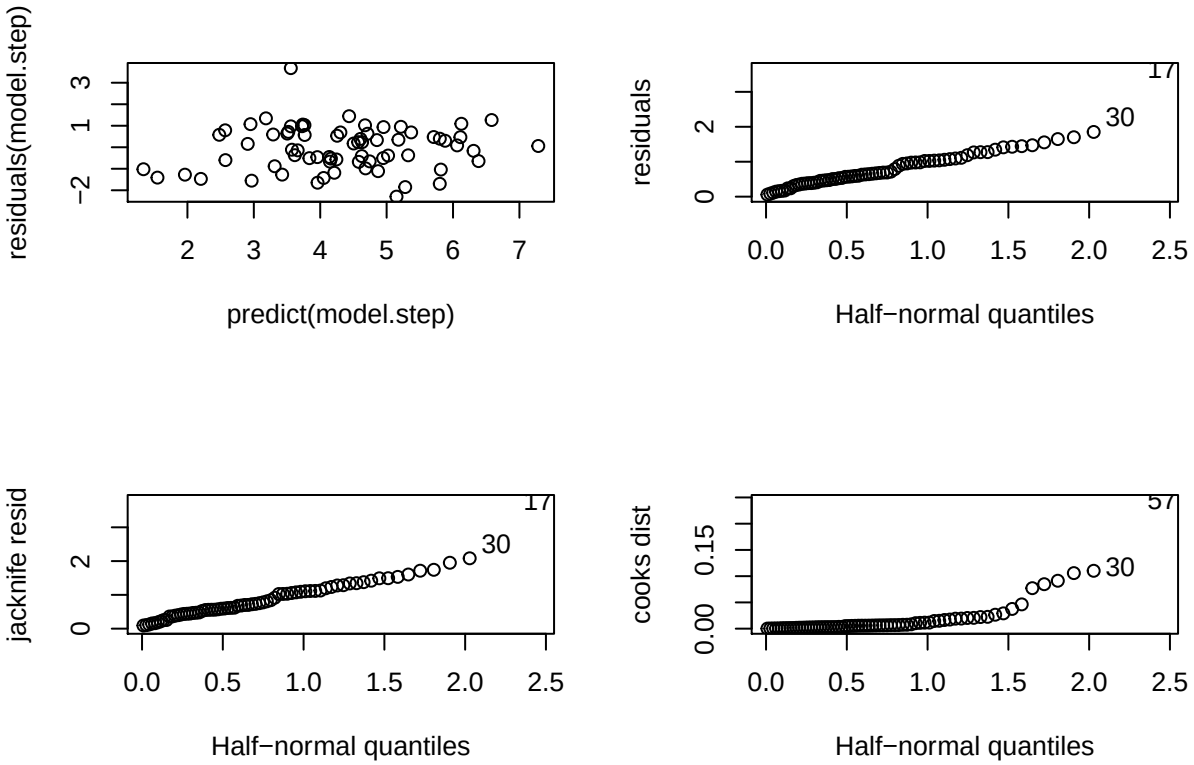
```
par(mfrow=c(1,2))
```

Checking for outliers or influential points

```
par(mfrow=c(2,2))
plot(residuals(model.step) ~ predict(model.step,type="response"))
plot(residuals(model.step) ~ predict(model.step, type="link"))
plot(residuals(model.step, type="pearson") ~ predict(model.step, type="link"))
plot(residuals(model.step, type="response") ~ predict(model.step, type="link"))
```




```
par(mfrow=c(2,2))
plot(predict(model.step), residuals(model.step))
halfnorm(residuals(model.step), ylab="residuals")
halfnorm(rstudent(model.step), ylab="jacknife resid")
halfnorm(cooks.distance(model.step), ylab="cooks dist")
```



From the response residuals vs linear fitted values plot, looks heteroskedastic. Observation 17 and 57 look quite influential.

Remove obs 57 (17 looks like it could possibly still belong on the smooth curve, however 57 is definitely out).

```
model.subset = glm(nChildren ~ offset(log(nMother)) + duration + residence + education + duration*reside
summary(model.subset)
```

```
##
## Call:
## glm(formula = nChildren ~ offset(log(nMother)) + duration + residence +
##     education + duration * residence + duration * education +
##     education * residence, family = poisson, data = data, subset = c(-57))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7787  -0.4211   0.0000   0.3396   2.7691
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.265067   0.054324  23.287  < 2e-16 ***
## duration.L                     1.310557   0.112276  11.673  < 2e-16 ***
## duration.Q                    -0.474022   0.099840  -4.748 2.06e-06 ***
## duration.C                     0.321782   0.092726   3.470 0.000520 ***
## duration^4                    -0.110931   0.086279  -1.286 0.198541
## duration^5                     0.010020   0.078557   0.128 0.898502
## residenceurban                 0.003124   0.066872   0.047 0.962744
## residencerural                 0.036243   0.060537   0.599 0.549383
## educationlower                -0.016838   0.064993  -0.259 0.795581
## educationupper                -0.286663   0.081194  -3.531 0.000415 ***
## educationsec+                 -0.667472   0.152932  -4.365 1.27e-05 ***
```

```
## duration.L:residenceurban      0.147354   0.109352   1.348 0.177814
## duration.Q:residenceurban     -0.100793   0.096920  -1.040 0.298358
## duration.C:residenceurban      0.051447   0.090949   0.566 0.571620
## duration^4:residenceurban     -0.058683   0.086256  -0.680 0.496293
## duration^5:residenceurban      0.085279   0.082498   1.034 0.301270
## duration.L:residencerural      0.227917   0.094924   2.401 0.016348 *
## duration.Q:residencerural     -0.108431   0.084662  -1.281 0.200281
## duration.C:residencerural     -0.024135   0.083912  -0.288 0.773640
## duration^4:residencerural      0.034775   0.080828   0.430 0.667022
## duration^5:residencerural     -0.029742   0.074305  -0.400 0.688957
## duration.L:educationlower      0.079037   0.098961   0.799 0.424483
## duration.Q:educationlower      0.017222   0.087444   0.197 0.843867
## duration.C:educationlower     -0.068801   0.086347  -0.797 0.425569
## duration^4:educationlower      0.052142   0.079816   0.653 0.513577
## duration^5:educationlower      0.079968   0.062317   1.283 0.199410
## duration.L:educationupper     -0.051786   0.106849  -0.485 0.627910
## duration.Q:educationupper      0.100347   0.096793   1.037 0.299870
## duration.C:educationupper     -0.051905   0.094664  -0.548 0.583477
## duration^4:educationupper      0.059688   0.089031   0.670 0.502594
## duration^5:educationupper     -0.036356   0.076619  -0.475 0.635140
## duration.L:educationsec+      -0.466965   0.445782  -1.048 0.294860
## duration.Q:educationsec+      -0.311453   0.410111  -0.759 0.447592
## duration.C:educationsec+      -0.175578   0.300405  -0.584 0.558903
## duration^4:educationsec+      -0.058018   0.201078  -0.289 0.772938
## duration^5:educationsec+      -0.051879   0.158332  -0.328 0.743168
## residenceurban:educationlower  0.015483   0.078831   0.196 0.844295
## residencerural:educationlower  0.048103   0.071000   0.678 0.498084
## residenceurban:educationupper  0.260219   0.099827   2.607 0.009142 **
## residencerural:educationupper  0.214231   0.092895   2.306 0.021102 *
## residenceurban:educationsec+   0.320239   0.144495   2.216 0.026674 *
## residencerural:educationsec+   0.256557   0.149336   1.718 0.085799 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3568.432  on 68  degrees of freedom
## Residual deviance:   30.616  on 27  degrees of freedom
## AIC: 535.63
##
## Number of Fisher Scoring iterations: 4
```

```
model.subset.step = step(model.subset, scope=~.)
```

```
## Start:  AIC=535.63
## nChildren ~ offset(log(nMother)) + duration + residence + education +
##     duration * residence + duration * education + education *
##     residence
##
##                       Df Deviance    AIC
## - duration:education  15   43.025 518.04
## - duration:residence  10   43.200 528.21
## <none>                     30.616 535.63
## - residence:education  6   42.651 535.66
##
```

```
## Step:  AIC=518.04
## nChildren ~ duration + residence + education + duration:residence +
##     residence:education + offset(log(nMother))
##
##                       Df Deviance    AIC
## - duration:residence  10   56.647 511.66
## <none>                     43.025 518.04
## - residence:education  6   56.335 519.35
## + duration:education  15   30.616 535.63
##
## Step:  AIC=511.66
## nChildren ~ duration + residence + education + residence:education +
##     offset(log(nMother))
##
##                       Df Deviance     AIC
## - residence:education  6    67.62  510.63
## <none>                     56.65  511.66
## + duration:residence  10    43.02  518.04
## + duration:education  15    43.20  528.21
## - duration             5  2587.60 3032.61
##
## Step:  AIC=510.63
## nChildren ~ duration + residence + education + offset(log(nMother))
##
##                       Df Deviance     AIC
## <none>                     67.62  510.63
## + residence:education  6    56.65  511.66
## + duration:residence  10    56.34  519.35
## + duration:education  15    53.86  526.87
## - residence            2    93.75  532.76
## - education            3   117.38  554.39
## - duration             5  2605.55 3038.56
```

```
summary(model.subset.step)
```

```
##
## Call:
## glm(formula = nChildren ~ duration + residence + education +
##     offset(log(nMother)), family = poisson, data = data, subset = c(-57))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3884  -0.6241   0.0929   0.6219   3.7210
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.16426    0.03101  37.540  < 2e-16 ***
## duration.L      1.48577    0.03415  43.502  < 2e-16 ***
## duration.Q     -0.52279    0.03037 -17.213  < 2e-16 ***
## duration.C      0.26981    0.02953   9.138  < 2e-16 ***
## duration^4     -0.05706    0.02800  -2.037  0.04160 *
## duration^5      0.04095    0.02481   1.650  0.09885 .
## residenceurban  0.11111    0.03250   3.418  0.00063 ***
## residencerural  0.14451    0.02864   5.046 4.52e-07 ***
## educationlower  0.03591    0.02388   1.504  0.13255
```

```
## educationupper -0.08855    0.03188  -2.778  0.00547 **
## educationsec+  -0.29975    0.05558  -5.394 6.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3568.432  on 68  degrees of freedom
## Residual deviance:   67.621  on 58  degrees of freedom
## AIC: 510.63
##
## Number of Fisher Scoring iterations: 4
```

Check scaled deviance

```
anova(model.subset.step)
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: nChildren
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev
## NULL                         68     3568.4
## duration   5   3412.2        63      156.3
## residence  2     38.9        61      117.4
## education  3     49.8        58       67.6
```

```
pchisq(deviance(model.subset.step), 58, lower.tail = FALSE)
```

```
## [1] 0.1816001
```

Size of scaled deviance makes sense.

Check if there's overdispersion

```
# estimate phi
(phihat <- sum(residuals(model.subset.step, type="pearson")^2) / 58)
```

```
## [1] 1.184437
```

There is no overdispersion.