The Factors that Influnce the Fertility of Women

Arshia Azarhoush

In this report, we will be investigating and determining which factors (duration, residence, education) and two-way interactions are related to the fertility rate of women.

First, we will read the data and create a *fertility* variable from *nChildren* and *nMother*. Then, we can summaries the data as a table.

```
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
str(data)
```

```
## 'data.frame':    70 obs. of  6 variables:
##  $ duration : Ord.factor w/ 6 levels "0-4"<"5-9"<"10-14"<..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ residence: Factor w/ 3 levels "Suva","urban",..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ education: Factor w/ 4 levels "none","lower",..: 1 2 3 4 1 2 3 4 1 2 ...
##  $ nMother  : int  8 21 42 51 12 27 39 51 62 102 ...
##  $ nChildren: int  4 24 38 37 14 23 41 35 60 98 ...
##  $ fertility: num  0.5 1.143 0.905 0.725 1.167 ...
```

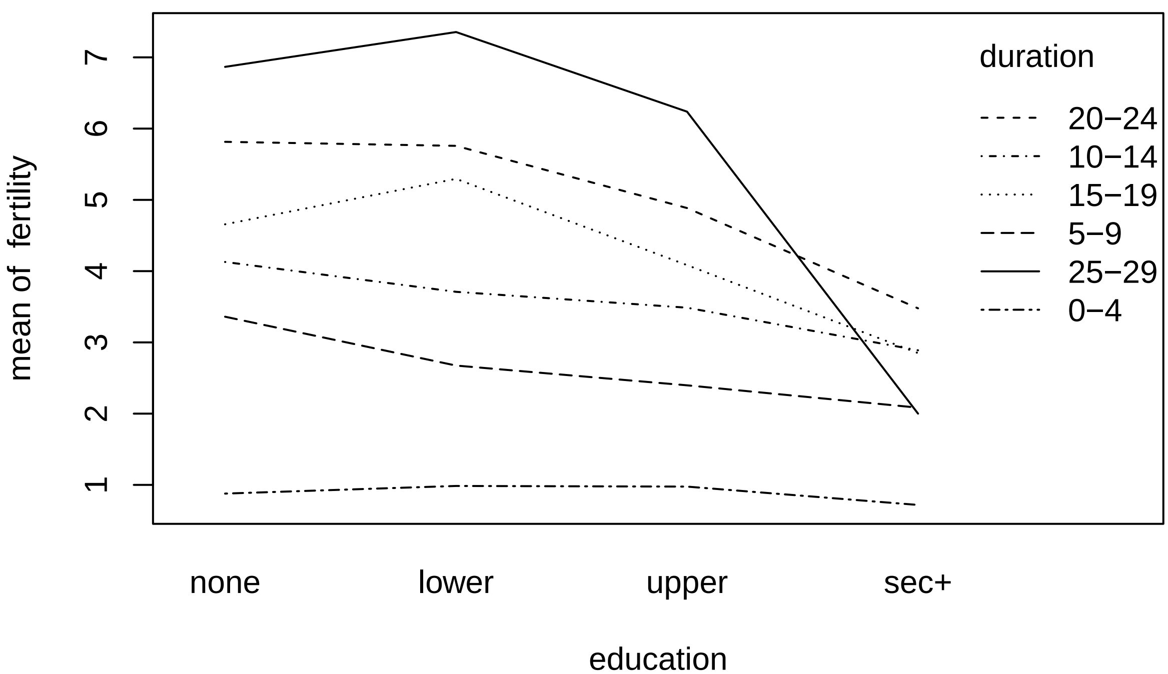
```
ftable(xtabs(cbind(fertility) ~ duration + residence + education, data))
```

```
##                     education      none     lower     upper      sec+
## duration residence
## 0-4      Suva                 0.5000000 1.1428571 0.9047619 0.7254902
##          urban                1.1666667 0.8518519 1.0512821 0.6862745
##          rural                0.9677419 0.9607843 0.9719626 0.7446809
## 5-9      Suva                 3.1000000 2.6666667 2.0416667 1.7272727
##          urban                4.5384615 2.6486486 2.6818182 2.2857143
##          rural                2.4428571 2.7094017 2.4691358 2.2380952
## 10-14    Suva                 4.0833333 3.6666667 2.9000000 2.0000000
##          urban                4.1666667 3.3255814 3.6206897 3.3333333
##          rural                4.1363636 4.1363636 3.9400000 3.3333333
## 15-19    Suva                 4.2142857 4.9354839 3.1538462 2.7500000
##          urban                4.6956522 5.3571429 4.6000000 3.8000000
##          rural                5.0614035 5.5930233 4.5000000 2.0000000
## 20-24    Suva                 5.6190476 5.0555556 3.9166667 2.6000000
##          urban                5.3636364 5.8800000 5.0000000 5.3333333
##          rural                6.4615385 6.3382353 5.7391304 2.5000000
## 25-29    Suva                 6.5957447 6.7407407 5.3750000 2.0000000
```

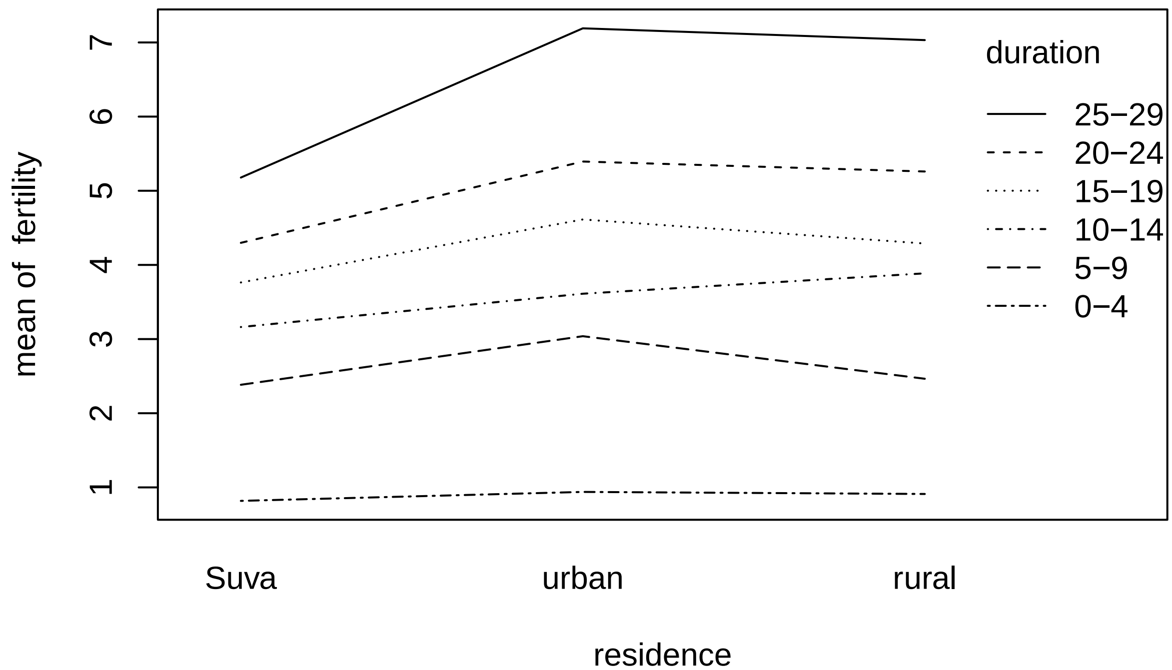
```
##          urban              6.5217391 7.5111111 7.5384615 0.0000000
##          rural              7.4820513 7.8135593 5.8000000 0.0000000
```

Next, we can check for interaction between the explanatory variables. As we can observe, the slopes of residence against duration are almost parallel. It is likely that the interaction is insignificant. We also observe that the slopes of residence depend on education, and the slopes of education depend on duration. This tells us that there there may be two way interaction between them, which may affect the response variable. We should include the interaction terms in our model.

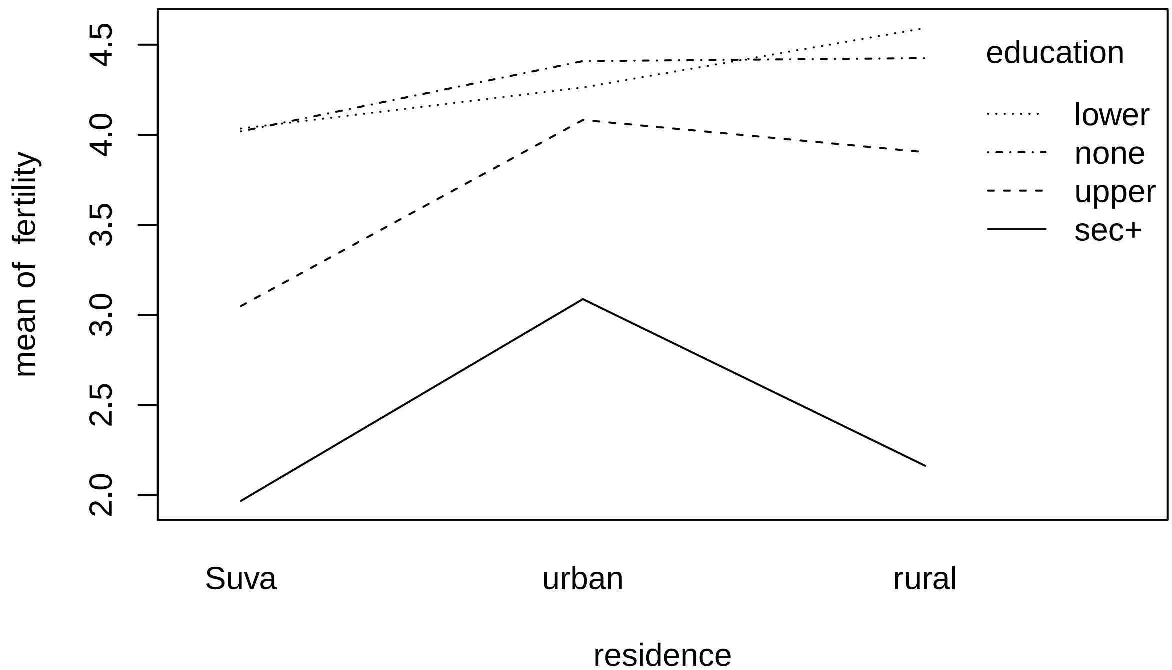
```
with(data, interaction.plot(education, duration, fertility))
```




```
with(data, interaction.plot(residence, duration, fertility))
```




```
with(data, interaction.plot(residence, education, fertility))
```



Poisson regression looks to be a viable model as the number of children in each group is count data. We must take into account the number of mothers in each group as we are looking to model fertility rate. We can model the rate per unit (fertility rate) using a log link via

$$\log(\lambda_i/t_i) = x_i^T\beta \tag{1}$$

$$\tag{2}$$

so we model *nchildren* by

$$\log(\lambda_i) = \log(t_i) + x_i^T\beta. \tag{3}$$

This is a form of Poisson glm with log-link, but the coefficient $\log(t_i)$ has been constrained to 1. This is called a rate model.

```
model = glm(nChildren ~ offset(log(nMother)) + duration + residence + education +
              duration*education + education*residence, family = poisson, data)
summary(model)
```

```
##
## Call:
## glm(formula = nChildren ~ offset(log(nMother)) + duration + residence +
##     education + duration * education + education * residence,
##     family = poisson, data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1367  -0.4502  -0.0095   0.4285   3.7144
```

```
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     1.210064   0.046967  25.764  < 2e-16 ***
## duration.L                      1.518901   0.073663  20.620  < 2e-16 ***
## duration.Q                     -0.575203   0.068303  -8.421  < 2e-16 ***
## duration.C                      0.289727   0.058884   4.920 8.64e-07 ***
## duration^4                     -0.117254   0.050160  -2.338  0.01941 *
## duration^5                     -0.012834   0.043354  -0.296  0.76722
## residenceurban                  0.044901   0.056947   0.788  0.43042
## residencerural                  0.110445   0.045402   2.433  0.01499 *
## educationlower                  0.004223   0.062909   0.067  0.94648
## educationupper                 -0.236249   0.076477  -3.089  0.00201 **
## educationsec+                  -0.597910   0.149151  -4.009 6.10e-05 ***
## duration.L:educationlower       0.047015   0.093635   0.502  0.61559
## duration.Q:educationlower       0.024440   0.086767   0.282  0.77819
## duration.C:educationlower      -0.041328   0.075775  -0.545  0.58547
## duration^4:educationlower       0.067491   0.065410   1.032  0.30216
## duration^5:educationlower       0.100810   0.056849   1.773  0.07618 .
## duration.L:educationupper      -0.104008   0.101112  -1.029  0.30365
## duration.Q:educationupper       0.112532   0.095479   1.179  0.23856
## duration.C:educationupper      -0.021788   0.085922  -0.254  0.79982
## duration^4:educationupper       0.067866   0.077576   0.875  0.38166
## duration^5:educationupper      -0.004022   0.072011  -0.056  0.95546
## duration.L:educationsec+       -0.596489   0.442144  -1.349  0.17731
## duration.Q:educationsec+       -0.314452   0.407594  -0.771  0.44042
## duration.C:educationsec+       -0.149450   0.297150  -0.503  0.61500
## duration^4:educationsec+       -0.072526   0.196438  -0.369  0.71198
## duration^5:educationsec+       -0.008128   0.154454  -0.053  0.95803
## residenceurban:educationlower   0.004952   0.076568   0.065  0.94843
## residencerural:educationlower   0.015318   0.064049   0.239  0.81099
## residenceurban:educationupper   0.230689   0.093915   2.456  0.01403 *
## residencerural:educationupper   0.140165   0.083322   1.682  0.09253 .
## residenceurban:educationsec+    0.248930   0.132061   1.885  0.05944 .
## residencerural:educationsec+    0.116804   0.137367   0.850  0.39516
## ---
## 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:   44.523  on 38  degrees of freedom
## AIC: 538
##
## Number of Fisher Scoring iterations: 4
```

We can test the significance of interaction in our model using a chi-squred test. As it turns out, the interaction terms are not statistically significant in our model.

```
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:education  15     15.9        44       54.8    0.3912
## residence:education  6     10.3        38       44.5    0.1134
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can remove the interaction terms and model the data again.

```r
model2 = glm(nChildren ~ offset(log(nMother)) + duration + residence + education,
             family = poisson, data)
summary(model2)
```

```
##
## 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 can utilize AIC in a Stepwise Algorithm to select the most statistically significant model. The full model has the lowest AIC so no further changes need to be made.

```
model3 = step(model2, scope = ~.)
```

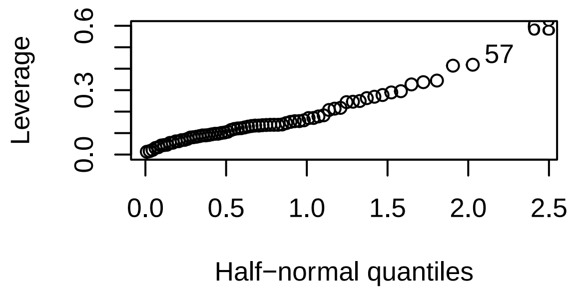
```
## Start:  AIC=522.14
## nChildren ~ offset(log(nMother)) + duration + residence + education
##
##             Df Deviance     AIC
## <none>              70.67  522.14
## - residence  2   100.19  547.67
## - education  3   120.68  566.16
## - duration   5  2646.49 3087.97
```

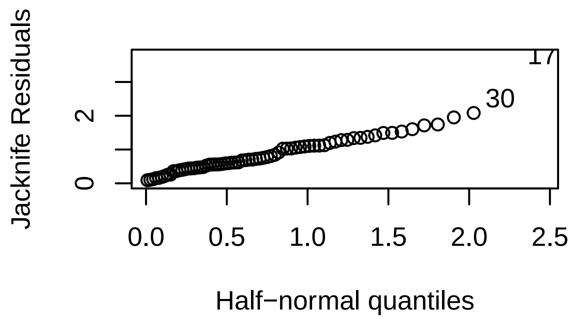
We can now check for outliers and points with significant impact by checking the leverage, jackknife residuals and Cook's distance of our data. Based on our tests, observations 17, 57 and 68 are influential points and may have high impact on our regression. These data points may be outliers or may have been subject to some errors (mis-recorded, etc.).

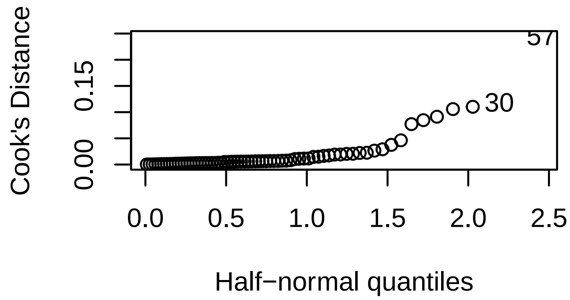
```
library("faraway")
```

```
## Warning: package 'faraway' was built under R version 4.1.3
```

```
par(mfrow=c(2,2))
# Observation 68 has moderately high leverage
halfnorm(influence(model2)$hat, ylab="Leverage")
# Observation 17 looks influential
halfnorm(rstudent(model2), ylab="Jacknife Residuals")
# Observation 57 looks influential
halfnorm(cooks.distance(model2), ylab="Cook's Distance")
```







We can remove observations 17, 57 and 68, and refit the model.

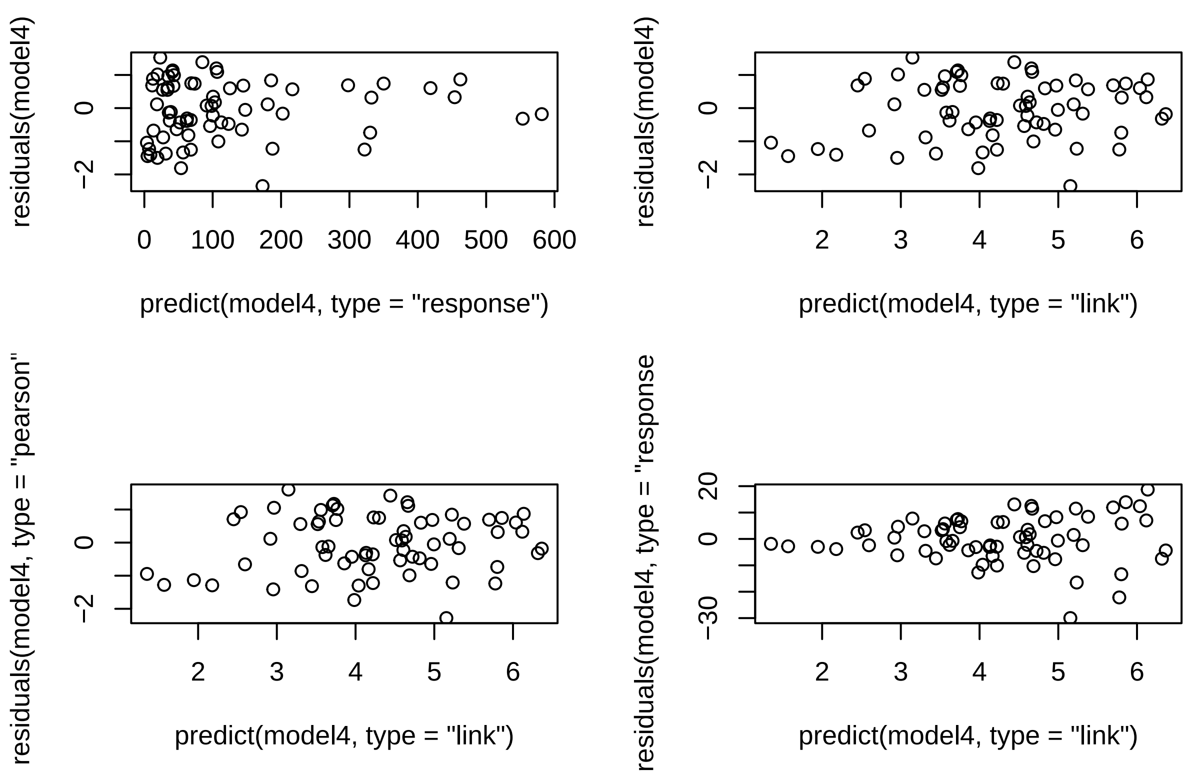
```
model4 = glm(nChildren ~ offset(log(nMother)) + duration + residence + education,
                    family = poisson, data, subset = c(-57, -17, -68))
summary(model4)
```

```
##
## Call:
## glm(formula = nChildren ~ offset(log(nMother)) + duration + residence +
##     education, family = poisson, data = data, subset = c(-57,
##     -17, -68))
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.35090  -0.66365  -0.05309   0.68099   1.52400
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      1.15176    0.03139  36.695  < 2e-16 ***
## duration.L       1.49280    0.03582  41.678  < 2e-16 ***
## duration.Q      -0.52530    0.03170 -16.573  < 2e-16 ***
## duration.C       0.25746    0.03012   8.547  < 2e-16 ***
## duration^4      -0.04445    0.02829  -1.571  0.11611
## duration^5       0.03577    0.02488   1.438  0.15045
## residenceurban   0.09932    0.03270   3.038  0.00239 **
## residencerural   0.14077    0.03038   4.633 3.60e-06 ***
```

```
## educationlower  0.05266    0.02641   1.994  0.04612 *
## educationupper -0.06973    0.03313  -2.105  0.03529 *
## educationsec+  -0.27915    0.05596  -4.989 6.07e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2984.582  on 66  degrees of freedom
## Residual deviance:   52.459  on 56  degrees of freedom
## AIC: 480.43
##
## Number of Fisher Scoring iterations: 4
```

Now we can perform diagnostics to check how well our model fits the data. All of our tests generally look OK.

```
par(mfrow=c(2,2))
plot(residuals(model4) ~ predict(model4, type="response"))
plot(residuals(model4) ~ predict(model4, type="link"))
plot(residuals(model4, type="pearson") ~ predict(model4, type="link"))
plot(residuals(model4, type="response") ~ predict(model4, type="link"))
```



Lastly, we can check for overdispersion. We can do this by estimating ϕ to see if it is close to 1. It is close enough to 1 to confirm that there is no overdispersion.

```r
(phihat <- sum(residuals(model4, type="pearson")^2) / 56)
```

```
## [1] 0.9070933
```

In conclusion, to estimate the number of children per woman, we can use a Poisson model modeled on attributes about the mothers, such as the marriage duration, residence of families and the education level. We found a lack of two-way interaction between any of these attributes.