

The Influence of Multiple Variables on High-Risk Neonatal Survivors Based on the Binomial Distribution GLM Model

Group 12

Congye Wang 35427962

Kieran Molloy 35762970

Nov 11, 2020

1 Introduction

The purpose of this report is to quantify the impact of CNS services on the family satisfaction. According to the dataset of Langley, D (2002), the baby is visited by family at least 5 times a week in the first month after discharge from the neonatal department. CNS may provide temporary advice to a family within 12 months of age. At the same time, it also provides designated nurses between primary and secondary health services.

Besides, 1488 complete data were collected by means of a questionnaire containing CNS information which was from 8 NNUs for 3 years. The data collected 10 covariates as well, and the dependent variable was whether the patient was hospitalized again within one year.

2 Methods

Firstly, the data was transformation and preliminary analysis. we begin by factorizing the qualitative variables in the data, and operate the exploratory analysis and visualization of the continuous variables. Violin plots are shown to depict the distribution of bwt and los variables.

Secondly, the preliminary model is created. We first model using a generalized linear model based on binomial distribution, which considers all of the interactions and transformations of the variables. Thus, the los variable was added in the initial model to calculate the effect, and use ANOVA with χ^2 testing to determine if the model is the same. The form of the two models are as follow

$$f(x, \pi) = \binom{n}{x} \pi^x (1 - \pi)^{(n-x)}, x = 0, 1, 2, \dots, n$$

and the quantitative variables are factored to some dummies matrix. Such as

$$X_{gest} = [x_1, x_2, x_3, x_4, x_5]$$

$$\Rightarrow X_{gest} = X_{1488 \times 5} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

After data preparation is complete, we first run a regression on all the independent variables except los and consider the interaction effects of all variables as follow:

```
mod.nolos.full <- glm(re.ad ~ . - los + cns:size + cns:gest + cns:bwt + cns:emp.f
+ cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt
+ size:emp.f + size:emp.m + size:edu + size:sex + gest:bwt
+ gest:emp.f + gest:emp.m + gest:edu + gest:sex + bwt:emp.f
+ bwt:emp.m + bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu
+ emp.f:sex + emp.m:edu + emp.m:sex + edu:sex,
data= neomod,
family=binomial)
```

We then added los variables to the independent variables and compared them using the ANOVA model. After determining that adding los variables had a significant effect on the model, we determined that the preliminary model with los and consider the effect of the interactions on these variables, namely cns:emp.f, size:emp.f, size:emp.m, size:edu, gest:edu, bwt:edu, emp.m:edu. The model diagnosis diagram is given in the result.

Thirdly, the model needs to be checked for overdispersion. The expected variance of the data sampled from the binomial distribution is $\sigma^2 = n\pi(1 - \pi)$, where n is the number of observations and π is the probability of belonging to the $Y = 1$ group. An overdispersion is defined as the observed variance of the response variable is greater than the expected variance of the binomial distribution. Overdispersion can lead to odd standard error tests and imprecise significance tests. When overdispersion occurs, the GLM function can still be used to fit the logistic regression, but then the binomial distribution needs to be changed for the Quasibinomial distribution.

This can test the overdispersion. If ϕ is greater than 1, it needs to be considered using Quasi-likelihood. One way to detect an overdispersion is to compare the residual deviation of the binomial distribution model with the residual degrees of freedom, if the ratio

$$\phi = \frac{\text{Residual Deviation}}{\text{Residual Degrees of Freedom}}$$

If it is much larger than 1, we can consider that there is an overdispersion. Fit the model twice, as well as use binomial family for the first time and use quasi binomial family for the second time. Suppose that the first glm() return object is marked as fit and the second return object is marked as fit.od.

Then:

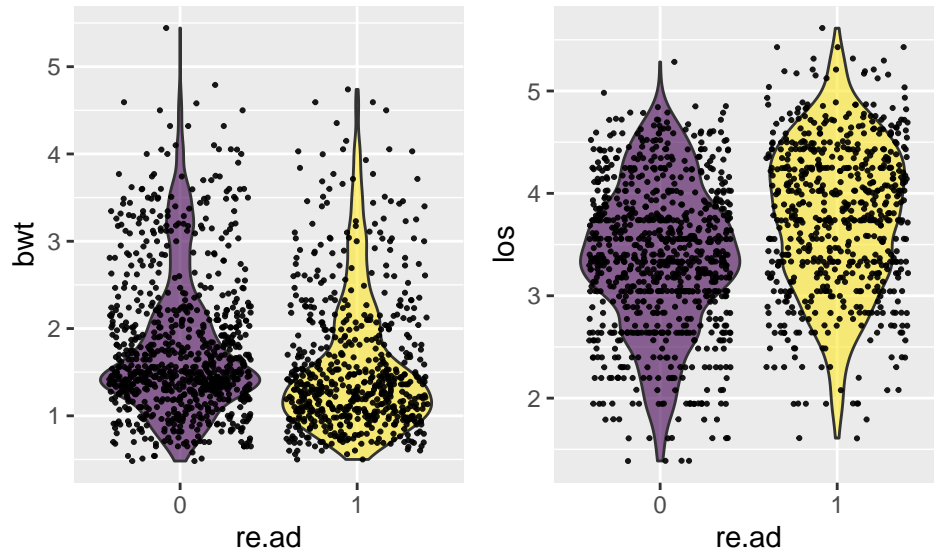
```
pchisq(summary(fit.od)$dispersion * fit$df.residual, fit$df.residual, lower = F)
```

The null hypothesis $H_0 : \phi = 1$ and alternative hypothesis $H_1 : \phi \neq 1$ can be tested by the provided p-value. If P is very small (less than 0.05), we can reject the null hypothesis.

Finally, the model was to be selected. Both the forward method and the backward method have obvious shortcomings. The forward method may have such a problem: it can not reflect the changes after the introduction of new independent variable values, since the AIC corresponding to the regression equation is the minimum when an independent variable is introduced. However, when other variables are introduced, the AIC value may be smaller if it is proposed from the regression equation, but there is no chance to put forward it by using the forward method, that is, once introduced, it will be lifelong. It is obviously not comprehensive to consider only the introduction but not the elimination. Similarly, once an independent variable is eliminated in the backward method, it has no chance to re-enter the regression equation. Stepwise regression combines forward stepwise regression and backward stepwise regression. Variables enter one at a time, but in each step, variables will be re-evaluated, variables that do not contribute to the model will be deleted, and prediction variables may be added or deleted several times until the optimal model is obtained. We use ANOVA to compare the full model and the step model. If the results show no significant difference, we tend to use the stepwise regression model.

3 Results

3.1 Exploratory Analysis



The violin graphs demonstrate that bwt's data are more concentrated when re.ad is used as a grouping, while los is more scattered and has outliers.

3.2 Model Established

```
mod.nolos.full <- glm(re.ad ~ . -los + cns:size + cns:gest + cns:bwt + cns:emp.f +  
  cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt +  
  size:emp.f + size:emp.m + size:edu + size:sex + gest:bwt +  
  gest:emp.f + gest:emp.m + gest:edu + gest:sex + bwt:emp.f +  
  bwt:emp.m + bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu +  
  emp.f:sex + emp.m:edu + emp.m:sex + edu:sex,  
  data= neomod,  
  family=binomial)  
  
summary.glm(mod.nolos.full)
```

Due to the lengthy results, it will be placed in the appendix. Based on the results shown, we were able to identify the significant interaction effects between cns and emp.f, size and emp.f, size and emp.m, size and edu, gest and edu, bwt and edu, as well as emp.m and edu.

```
mod.los.full <- glm(re.ad ~ . + cns:size + cns:gest + cns:bwt + cns:emp.f + cns:emp.m +  
  cns:edu + cns:sex + size:gest + size:bwt + size:emp.f + size:emp.m +  
  size:edu + size:sex + gest:bwt + gest:emp.f + gest:emp.m + gest:edu +  
  gest:sex + bwt:emp.f + bwt:emp.m + bwt:edu + bwt:sex + emp.f:emp.m +  
  emp.f:edu + emp.f:sex + emp.m:edu + emp.m:sex + edu:sex,  
  data= neomod,  
  family=binomial)  
  
summary.glm(mod.los.full)
```

This time the model added los variables and the results showed a significant p-value for los, so we need to consider regression using los as an independent variable. In addition, we may consider the effect of the interactions on these variables, namely cns:emp.f, size:emp.f, size:emp.m, gest:edu, bwt:edu, and emp.m:edu.

```
anova(mod.los.full, mod.nolos.full, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: re.ad ~ cns + size + gest + bwt + emp.f + emp.m + edu + los +
##      sex + accom + cns:size + cns:gest + cns:bwt + cns:emp.f +
##      cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt + size:emp.f +
##      size:emp.m + size:edu + size:sex + gest:bwt + gest:emp.f +
##      gest:emp.m + gest:edu + gest:sex + bwt:emp.f + bwt:emp.m +
##      bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu + emp.f:sex +
##      emp.m:edu + emp.m:sex + edu:sex
## Model 2: re.ad ~ (cns + size + gest + bwt + emp.f + emp.m + edu + los +
##      sex + accom) - los + cns:size + cns:gest + cns:bwt + cns:emp.f +
##      cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt + size:emp.f +
##      size:emp.m + size:edu + size:sex + gest:bwt + gest:emp.f +
##      gest:emp.m + gest:edu + gest:sex + bwt:emp.f + bwt:emp.m +
##      bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu + emp.f:sex +
##      emp.m:edu + emp.m:sex + edu:sex
##   Resid. Df Resid. Dev Df Deviance   Pr(>Chi)
## 1         1404      1722.7
## 2         1405      1768.7 -1   -46.068 1.142e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This result shows that we cannot ignore the los variable simply since these two models are different. Therefore, we may choose the second model.

3.3 Overdispersion Test

Firstly, we calculated the ϕ in this step.

```
mod.binomial <- glm(re.ad ~ . + cns:emp.f + size:emp.f + size:emp.m + gest:edu + bwt:edu + emp.m:edu,
                    data= neomod,
                    family=binomial)

deviance(mod.binomial) / df.residual(mod.binomial)

## [1] 1.254831
```

The results show that although X is greater than 1, but it is still a small value. Hence, in order to quantitatively determine the analysis, we chose to calculate the p-value.

```
mod.quasibinomial <- glm(re.ad ~ . + cns:emp.f + size:emp.f + size:emp.m + gest:edu + bwt:edu + emp.m:edu,
                         data= neomod,
                         family=quasibinomial)

pchisq(summary(mod.binomial)$dispersion * mod.quasibinomial$df.residual, mod.quasibinomial$df.residual,

## [1] 0.4950646
```

This p-value is 0.4950646 which is much bigger than 0.05. As a result, we can make sure we do not need to use Quasi-Binomial family model.

3.4 Model Selection

We used stepwise regression intending to reduce the AIC value to select the model.

```
library(MASS)
mod.step <- stepAIC(mod.binomial, direction = "both", trace = FALSE)
mod.step
```

After obtaining the stepwise regression equation, we performed an ANOVA test on the original model and it as follow.

```
anova(mod.step, mod.binomial, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: re.ad ~ cns + size + emp.f + emp.m + edu + los + sex + accom +
##      cns:emp.f + size:emp.f + emp.m:edu
## Model 2: re.ad ~ cns + size + gest + bwt + emp.f + emp.m + edu + los +
##      sex + accom + cns:emp.f + size:emp.f + size:emp.m + gest:edu +
##      bwt:edu + emp.m:edu
##      Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1          1472      1850.8
## 2          1452      1822.0 20   28.838  0.09101 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that the p-value is 0.114, which is greater than 0.05. Therefore, we can judge that the two models are not significantly different at a confidence interval of 95%. The stepwise regression results are chosen as the final model.

3.5 Model Diagnostics

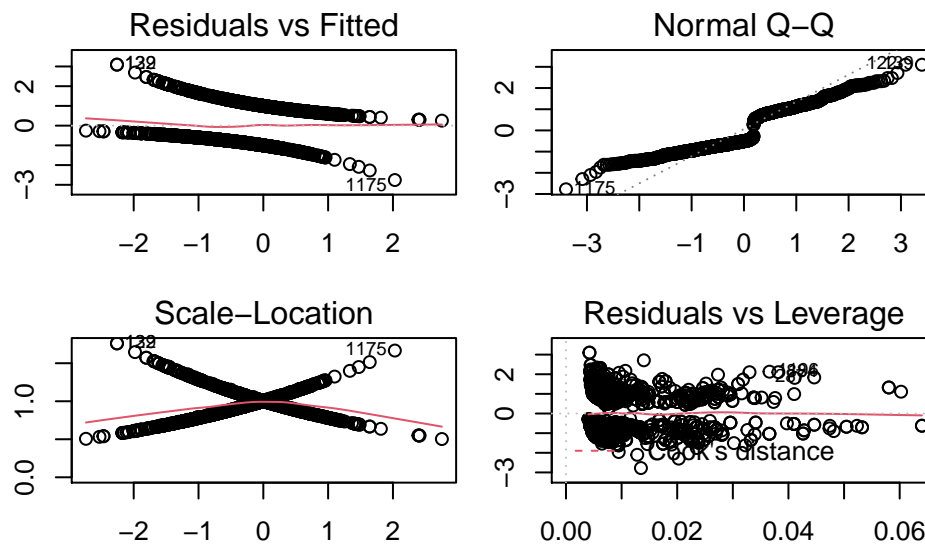
```
summary.glm(mod.step)

##
## Call:
## glm(formula = re.ad ~ cns + size + emp.f + emp.m + edu + los +
##      sex + accom + cns:emp.f + size:emp.f + emp.m:edu, family = binomial,
##      data = neomod)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0750  -0.9914  -0.6761   1.1112   2.1715
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.92135    0.53208  -5.490 4.01e-08 ***
## cns1          -0.77454    0.36077  -2.147  0.03180 *
## size1         1.23064    0.37916   3.246  0.00117 **
## emp.f1        0.24307    0.35590   0.683  0.49463
## emp.m1       -1.30891    0.56873  -2.301  0.02137 *
## edu2         -0.44006    0.31121  -1.414  0.15735
## edu3          0.12542    0.33830   0.371  0.71083
## edu4         -0.21846    0.40323  -0.542  0.58798
## los           0.76432    0.08051   9.493 < 2e-16 ***
## sex1          0.47432    0.11353   4.178 2.94e-05 ***
## accom1       -0.44403    0.14160  -3.136  0.00171 **
## cns1:emp.f1   0.96974    0.38087   2.546  0.01089 *
## size1:emp.f1 -1.24017    0.40183  -3.086  0.00203 **
```

```
## emp.m1:edu2  1.39247    0.58795    2.368  0.01787 *
## emp.m1:edu3  0.36251    0.61228    0.592  0.55381
## emp.m1:edu4  1.04384    0.66234    1.576  0.11503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2031.3  on 1487  degrees of freedom
## Residual deviance: 1850.9  on 1472  degrees of freedom
## AIC: 1882.9
##
## Number of Fisher Scoring iterations: 4
```

From the summary results, `emp.fl` and `edu` have p-values greater than 0.05, which is not significant. Apart from those, the coefficients of the variables all pass the significance test. Meanwhile, the interaction of `emp.m1` and `edu2` has the greatest impact on the positive correlation of the model, and `emp.m1` has the greatest impact on the negative correlation of the model.

```
par(mfrow=c(2,2), mar=c(2.5,2,2,1))
plot(mod.step)
```



As can be seen from the residual versus fitted graph plots, there is no systematic correlation between the residual and the fitted values, which means that we do not need to perform a responsible transformation of the independent variables. The normal Q-Q plot shows the probability plot of the standardized residuals for the values corresponding to the normal distribution, which can be seen that the model largely satisfies the assumption of normality. If the assumption of constant variance is met, then in the Scale-Location Graph, the points around the horizontal line should be randomly distributed. The graph does not seem to meet this assumption. The last “Residuals vs Leverage” graph provides information on individual observations that may be of interest. Outliers, high leverage value points, and strong influence points can be identified from the graph. The pictures show that the Cook distances for most of the data are concentrated around 0.00 to 0.01.

4 Conclusion

In Chapter 3 we discussed in detail the results of the model, which showed that a mother’s employment versus leaving full-time education largely drove her child’s admission to hospital within 12 months, whereas a mother’s care at home would have prevented the child from going to the hospital. As well as, the larger of

the size of NUU, the higher probability of entering the hospital. By contraries, living alone and receiving help from CNS can save a child from going to the hospital.

The model is good overall, but there are two shortcomings. First, the model has large AIC values, which means that the fit is not very good. Second, there are insignificant coefficients in the model, which reduces the model's goodness-of-fit and the adjusted R^2 .

References

Langley, D, S Hollis, T Friede, D MacGregor, and A Gatrell. 2002. "Impact of Community Neonatal Services: A Multicentre Survey." *Archives of Disease in Childhood-Fetal and Neonatal Edition* 87 (3): F204–F208.

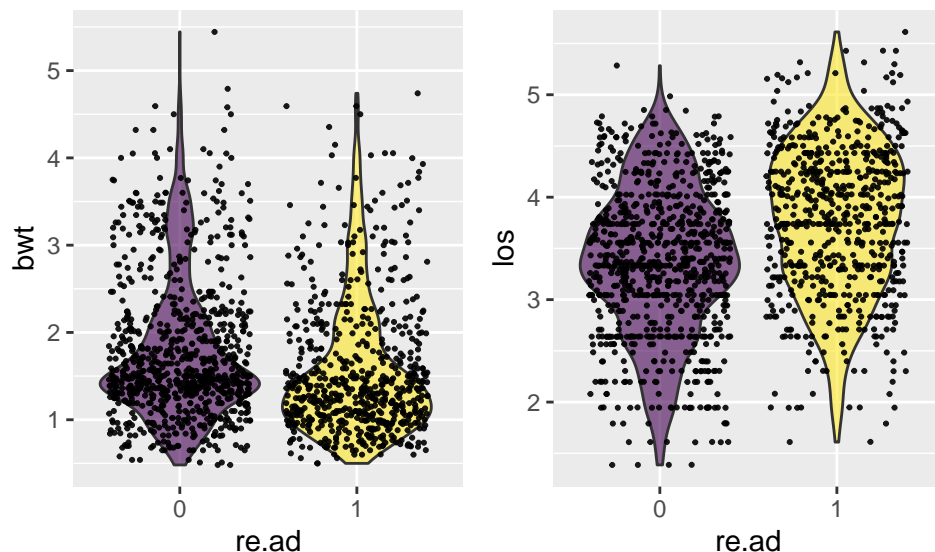
Appendix of R code

```
library(tidyverse)
library(viridis)
library(car)
library(ggpubr)
library(MASS)

neomod <- read.table("../Data/neomod_isam12.dat", header = TRUE)

neomod$cns      = factor(neomod$cns)
neomod$size     = factor(neomod$size)
neomod$gest     = factor(neomod$gest)
neomod$emp.f    = factor(neomod$emp.f)
neomod$emp.m    = factor(neomod$emp.m)
neomod$edu      = factor(neomod$edu)
neomod$re.ad    = factor(neomod$re.ad)
neomod$sex      = factor(neomod$sex)
neomod$accom    = factor(neomod$accom)

p.bwt <- neomod %>%
  ggplot(aes(x = re.ad, y = bwt, fill = re.ad)) +
    geom_violin() +
    scale_fill_viridis(discrete = TRUE, alpha = 0.6) +
    geom_jitter(color = "black", size = 0.4, alpha = 0.9) +
    theme(
      legend.position = "none",
      plot.title = element_text(size = 11)
    )
p.los <- neomod %>%
  ggplot(aes(x = re.ad, y = los, fill = re.ad)) +
    geom_violin() +
    scale_fill_viridis(discrete = TRUE, alpha=0.6) +
    geom_jitter(color="black", size=0.4, alpha=0.9) +
    theme(
      legend.position = "none",
      plot.title = element_text(size = 11)
    )
ggarrange(p.bwt, p.los, ncol = 2, nrow = 1)
```

```
mod.nolos.full <- glm(re.ad ~ . - los + cns:size + cns:gest + cns:bwt + cns:emp.f +
  cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt +
  size:emp.f + size:emp.m + size:edu + size:sex + gest:bwt +
  gest:emp.f + gest:emp.m + gest:edu + gest:sex + bwt:emp.f +
  bwt:emp.m + bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu +
  emp.f:sex + emp.m:edu + emp.m:sex + edu:sex,
  data = neomod,
  family = binomial)
```

```
summary.glm(mod.nolos.full)
```

```
##
## Call:
## glm(formula = re.ad ~ . - los + cns:size + cns:gest + cns:bwt +
##   cns:emp.f + cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt +
##   size:emp.f + size:emp.m + size:edu + size:sex + gest:bwt +
##   gest:emp.f + gest:emp.m + gest:edu + gest:sex + bwt:emp.f +
##   bwt:emp.m + bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu +
##   emp.f:sex + emp.m:edu + emp.m:sex + edu:sex, family = binomial,
##   data = neomod)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4757  -0.9438  -0.6251   1.0749   2.2946
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.194e+01  9.849e+02  -0.012  0.99032
## cns1         -2.389e-02  1.117e+00  -0.021  0.98294
## size1        2.214e+00  1.268e+00   1.746  0.08082 .
## gest2        9.052e+00  9.849e+02   0.009  0.99267
## gest3        9.759e+00  9.849e+02   0.010  0.99209
## gest4        6.192e+00  9.849e+02   0.006  0.99498
## gest5        5.224e+00  9.849e+02   0.005  0.99577
## bwt          7.545e-01  3.311e+00   0.228  0.81973
## emp.f1       -1.308e+01  4.367e+02  -0.030  0.97611
```

## emp.m1	-1.211e+00	1.368e+00	-0.885	0.37591
## edu2	2.661e+01	8.827e+02	0.030	0.97595
## edu3	2.541e+01	8.828e+02	0.029	0.97703
## edu4	8.156e+00	3.135e+00	2.602	0.00927 **
## sex1	2.692e+00	1.495e+00	1.801	0.07171 .
## accom1	-4.847e-01	1.523e-01	-3.183	0.00146 **
## cns1:size1	1.772e-04	2.768e-01	0.001	0.99949
## cns1:gest2	2.440e-01	8.468e-01	0.288	0.77326
## cns1:gest3	-6.134e-01	8.555e-01	-0.717	0.47337
## cns1:gest4	4.390e-01	9.090e-01	0.483	0.62916
## cns1:gest5	-3.590e-01	1.095e+00	-0.328	0.74291
## cns1:bwt	-1.983e-01	2.756e-01	-0.719	0.47184
## cns1:emp.f1	1.172e+00	4.383e-01	2.674	0.00750 **
## cns1:emp.m1	8.537e-02	2.530e-01	0.337	0.73580
## cns1:edu2	-7.798e-01	6.582e-01	-1.185	0.23608
## cns1:edu3	2.145e-01	6.899e-01	0.311	0.75584
## cns1:edu4	-6.244e-01	7.746e-01	-0.806	0.42020
## cns1:sex1	-3.513e-01	2.524e-01	-1.392	0.16404
## size1:gest2	7.474e-01	9.596e-01	0.779	0.43605
## size1:gest3	2.840e-01	9.773e-01	0.291	0.77140
## size1:gest4	-8.958e-03	1.029e+00	-0.009	0.99306
## size1:gest5	6.156e-01	1.243e+00	0.495	0.62047
## size1:bwt	-3.029e-01	3.033e-01	-0.999	0.31798
## size1:emp.f1	-1.018e+00	4.397e-01	-2.315	0.02061 *
## size1:emp.m1	2.252e-01	2.810e-01	0.801	0.42290
## size1:edu2	-1.671e+00	7.891e-01	-2.117	0.03426 *
## size1:edu3	-9.310e-01	8.258e-01	-1.127	0.25955
## size1:edu4	-1.955e+00	8.717e-01	-2.243	0.02489 *
## size1:sex1	3.617e-01	2.769e-01	1.306	0.19145
## gest2:bwt	9.175e-01	3.220e+00	0.285	0.77569
## gest3:bwt	4.824e-01	3.213e+00	0.150	0.88067
## gest4:bwt	1.522e+00	3.209e+00	0.474	0.63533
## gest5:bwt	1.163e+00	3.208e+00	0.363	0.71693
## gest2:emp.f1	1.320e+01	4.367e+02	0.030	0.97589
## gest3:emp.f1	1.224e+01	4.367e+02	0.028	0.97763
## gest4:emp.f1	1.267e+01	4.367e+02	0.029	0.97686
## gest5:emp.f1	1.303e+01	4.367e+02	0.030	0.97620
## gest2:emp.m1	-7.453e-01	9.710e-01	-0.768	0.44273
## gest3:emp.m1	-6.992e-01	9.832e-01	-0.711	0.47701
## gest4:emp.m1	-5.281e-01	1.033e+00	-0.511	0.60935
## gest5:emp.m1	4.709e-01	1.203e+00	0.392	0.69535
## gest2:edu2	-2.420e+01	8.827e+02	-0.027	0.97813
## gest3:edu2	-2.280e+01	8.827e+02	-0.026	0.97940
## gest4:edu2	-2.139e+01	8.827e+02	-0.024	0.98067
## gest5:edu2	-2.112e+01	8.827e+02	-0.024	0.98091
## gest2:edu3	-2.369e+01	8.827e+02	-0.027	0.97859
## gest3:edu3	-2.281e+01	8.827e+02	-0.026	0.97939
## gest4:edu3	-2.195e+01	8.827e+02	-0.025	0.98017
## gest5:edu3	-2.259e+01	8.827e+02	-0.026	0.97959
## gest2:edu4	-4.059e+00	2.047e+00	-1.983	0.04741 *
## gest3:edu4	-3.363e+00	1.677e+00	-2.006	0.04490 *
## gest4:edu4	-9.134e-01	1.510e+00	-0.605	0.54532
## gest5:edu4	NA	NA	NA	NA
## gest2:sex1	-1.305e+00	1.243e+00	-1.049	0.29400

```

## gest3:sex1      -1.790e+00  1.246e+00  -1.436  0.15091
## gest4:sex1      -2.184e+00  1.284e+00  -1.701  0.08899 .
## gest5:sex1      -9.334e-01  1.415e+00  -0.660  0.50943
## bwt:emp.f1       1.474e-01  4.190e-01   0.352  0.72491
## bwt:emp.m1      -5.124e-01  2.793e-01  -1.834  0.06658 .
## bwt:edu2        -1.513e+00  7.423e-01  -2.038  0.04152 *
## bwt:edu3        -1.009e+00  7.940e-01  -1.271  0.20383
## bwt:edu4        -2.405e+00  8.641e-01  -2.783  0.00539 **
## bwt:sex1        -4.052e-01  2.787e-01  -1.454  0.14589
## emp.f1:emp.m1    2.023e-01  5.109e-01   0.396  0.69210
## emp.f1:edu2     -8.722e-02  7.791e-01  -0.112  0.91086
## emp.f1:edu3     -4.367e-02  8.999e-01  -0.049  0.96130
## emp.f1:edu4     -4.810e-01  1.075e+00  -0.447  0.65464
## emp.f1:sex1      3.108e-01  4.262e-01   0.729  0.46592
## emp.m1:edu2      1.772e+00  7.625e-01  2.323  0.02016 *
## emp.m1:edu3      6.860e-01  7.884e-01   0.870  0.38421
## emp.m1:edu4      1.659e+00  8.386e-01  1.979  0.04785 *
## emp.m1:sex1      8.277e-01  2.589e-01  3.197  0.00139 **
## edu2:sex1       -6.823e-01  6.718e-01  -1.016  0.30979
## edu3:sex1       -4.017e-01  6.990e-01  -0.575  0.56545
## edu4:sex1        3.078e-01  7.855e-01   0.392  0.69521
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2031.3  on 1487  degrees of freedom
## Residual deviance: 1768.7  on 1405  degrees of freedom
## AIC: 1934.7
##
## Number of Fisher Scoring iterations: 13
mod.los.full <- glm(re.ad ~ . + cns:size + cns:gest + cns:bwt + cns:emp.f + cns:emp.m +
  cns:edu + cns:sex + size:gest + size:bwt + size:emp.f + size:emp.m +
  size:edu + size:sex + gest:bwt + gest:emp.f + gest:emp.m + gest:edu +
  gest:sex + bwt:emp.f + bwt:emp.m + bwt:edu + bwt:sex + emp.f:emp.m +
  emp.f:edu + emp.f:sex + emp.m:edu + emp.m:sex + edu:sex,
  data = neomod,
  family = binomial)
summary.glm(mod.los.full)

##
## Call:
## glm(formula = re.ad ~ . + cns:size + cns:gest + cns:bwt + cns:emp.f +
##      cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt + size:emp.f +
##      size:emp.m + size:edu + size:sex + gest:bwt + gest:emp.f +
##      gest:emp.m + gest:edu + gest:sex + bwt:emp.f + bwt:emp.m +
##      bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu + emp.f:sex +
##      emp.m:edu + emp.m:sex + edu:sex, family = binomial, data = neomod)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.3691  -0.9233  -0.5668   1.0104   2.1828
##

```

```

## Coefficients: (1 not defined because of singularities)
##           Estimate Std. Error z value Pr(>|z|)
## (Intercept) -15.30554  985.68627  -0.016 0.987611
## cns1        -0.04972   1.12440  -0.044 0.964728
## size1        2.05727   1.28920   1.596 0.110541
## gest2        8.05689  985.68420   0.008 0.993478
## gest3        9.33197  985.68372   0.009 0.992446
## gest4        6.30369  985.68373   0.006 0.994897
## gest5        6.21208  985.68341   0.006 0.994972
## bwt          1.15775   3.33528   0.347 0.728498
## emp.f1      -12.91927  438.55002  -0.029 0.976498
## emp.m1       -1.09892   1.37232  -0.801 0.423264
## edu2        25.29236  882.74988   0.029 0.977142
## edu3        24.23824  882.75049   0.027 0.978095
## edu4         7.32128   3.17416   2.307 0.021081 *
## los          0.88172   0.13432   6.564 5.23e-11 ***
## sex1         2.69855   1.50925   1.788 0.073776 .
## accom1      -0.55990   0.15503  -3.612 0.000304 ***
## cns1:size1    0.03417   0.28143   0.121 0.903350
## cns1:gest2    0.30078   0.85085   0.354 0.723708
## cns1:gest3   -0.43990   0.86038  -0.511 0.609152
## cns1:gest4    0.58664   0.91475   0.641 0.521319
## cns1:gest5   -0.28978   1.10116  -0.263 0.792427
## cns1:bwt     -0.19197   0.27676  -0.694 0.487901
## cns1:emp.f1   1.07618   0.44910   2.396 0.016561 *
## cns1:emp.m1   0.10633   0.25663   0.414 0.678633
## cns1:edu2    -0.62743   0.65733  -0.955 0.339829
## cns1:edu3     0.32094   0.68930   0.466 0.641496
## cns1:edu4    -0.57554   0.77292  -0.745 0.456494
## cns1:sex1    -0.39627   0.25712  -1.541 0.123266
## size1:gest2   0.67426   0.96478   0.699 0.484630
## size1:gest3   0.20021   0.98215   0.204 0.838476
## size1:gest4   0.03114   1.03474   0.030 0.975992
## size1:gest5   0.76597   1.25304   0.611 0.541012
## size1:bwt    -0.33924   0.30516  -1.112 0.266276
## size1:emp.f1  -1.03773   0.45328  -2.289 0.022056 *
## size1:emp.m1   0.32094   0.28585   1.123 0.261533
## size1:edu2   -1.45609   0.80799  -1.802 0.071526 .
## size1:edu3   -0.73177   0.84519  -0.866 0.386596
## size1:edu4   -1.57826   0.88936  -1.775 0.075964 .
## size1:sex1    0.35580   0.28086   1.267 0.205222
## gest2:bwt     1.12372   3.24129   0.347 0.728824
## gest3:bwt     0.44161   3.23454   0.137 0.891402
## gest4:bwt     1.37654   3.23040   0.426 0.670020
## gest5:bwt     0.73488   3.22891   0.228 0.819963
## gest2:emp.f1  13.34340  438.54924   0.030 0.975727
## gest3:emp.f1  12.47037  438.54936   0.028 0.977315
## gest4:emp.f1  12.77743  438.54980   0.029 0.976756
## gest5:emp.f1  13.44624  438.55071   0.031 0.975540
## gest2:emp.m1  -0.75164   0.97774  -0.769 0.442041
## gest3:emp.m1  -0.75263   0.99030  -0.760 0.447252
## gest4:emp.m1  -0.50354   1.04184  -0.483 0.628867
## gest5:emp.m1   0.52206   1.21238   0.431 0.666753
## gest2:edu2   -23.29475  882.74745  -0.026 0.978947

```

```

## gest3:edu2      -21.87563  882.74686  -0.025  0.980229
## gest4:edu2      -20.64104  882.74669  -0.023  0.981345
## gest5:edu2      -20.89347  882.74579  -0.024  0.981117
## gest2:edu3      -22.83799  882.74782  -0.026  0.979360
## gest3:edu3      -21.84202  882.74721  -0.025  0.980260
## gest4:edu3      -21.11132  882.74705  -0.024  0.980920
## gest5:edu3      -22.14078  882.74631  -0.025  0.979990
## gest2:edu4       -3.34145    2.07125  -1.613  0.106690
## gest3:edu4       -2.75542    1.69636  -1.624  0.104308
## gest4:edu4       -0.42570    1.55180  -0.274  0.783834
## gest5:edu4            NA         NA      NA      NA
## gest2:sex1       -1.20150    1.25274  -0.959  0.337511
## gest3:sex1       -1.74589    1.25675  -1.389  0.164769
## gest4:sex1       -2.10355    1.29473  -1.625  0.104224
## gest5:sex1       -0.74942    1.42814  -0.525  0.599756
## bwt:emp.f1        0.02875    0.43414    0.066  0.947208
## bwt:emp.m1       -0.51826    0.28120  -1.843  0.065328 .
## bwt:edu2         -1.19610    0.76016  -1.573  0.115606
## bwt:edu3         -0.81190    0.81067  -1.002  0.316580
## bwt:edu4         -2.09621    0.87970  -2.383  0.017179 *
## bwt:sex1         -0.50630    0.28220  -1.794  0.072790 .
## emp.f1:emp.m1     0.17142    0.51984    0.330  0.741591
## emp.f1:edu2      -0.10241    0.80560  -0.127  0.898844
## emp.f1:edu3      -0.15611    0.92611  -0.169  0.866138
## emp.f1:edu4      -0.82165    1.09175  -0.753  0.451691
## emp.f1:sex1       0.42070    0.43598    0.965  0.334568
## emp.m1:edu2       1.64910    0.75893    2.173  0.029786 *
## emp.m1:edu3       0.57090    0.78516    0.727  0.467155
## emp.m1:edu4       1.48965    0.83492    1.784  0.074395 .
## emp.m1:sex1       0.85437    0.26294    3.249  0.001157 **
## edu2:sex1        -0.85332    0.67927  -1.256  0.209031
## edu3:sex1        -0.43644    0.70659  -0.618  0.536798
## edu4:sex1         0.07395    0.79262    0.093  0.925669
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2031.3  on 1487  degrees of freedom
## Residual deviance: 1722.7  on 1404  degrees of freedom
## AIC: 1890.7
##
## Number of Fisher Scoring iterations: 13
anova(mod.los.full, mod.nolos.full, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: re.ad ~ cns + size + gest + bwt + emp.f + emp.m + edu + los +
##      sex + accom + cns:size + cns:gest + cns:bwt + cns:emp.f +
##      cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt + size:emp.f +
##      size:emp.m + size:edu + size:sex + gest:bwt + gest:emp.f +
##      gest:emp.m + gest:edu + gest:sex + bwt:emp.f + bwt:emp.m +
##      bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu + emp.f:sex +
##      emp.m:edu + emp.m:sex + edu:sex

```

```

## Model 2: re.ad ~ (cns + size + gest + bwt + emp.f + emp.m + edu + los +
##      sex + accom) - los + cns:size + cns:gest + cns:bwt + cns:emp.f +
##      cns:emp.m + cns:edu + cns:sex + size:gest + size:bwt + size:emp.f +
##      size:emp.m + size:edu + size:sex + gest:bwt + gest:emp.f +
##      gest:emp.m + gest:edu + gest:sex + bwt:emp.f + bwt:emp.m +
##      bwt:edu + bwt:sex + emp.f:emp.m + emp.f:edu + emp.f:sex +
##      emp.m:edu + emp.m:sex + edu:sex
##   Resid. Df Resid. Dev Df Deviance   Pr(>Chi)
## 1      1404      1722.7
## 2      1405      1768.7 -1   -46.068 1.142e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

mod.binomial <- glm(re.ad ~ . + cns:emp.f + size:emp.f + size:emp.m + gest:edu + bwt:edu + emp.m:edu,
  data = neomod,
  family = binomial)

deviance(mod.binomial) / df.residual(mod.binomial)

## [1] 1.254831

mod.quasibinomial <- glm(re.ad ~ . + cns:emp.f + size:emp.f + size:emp.m + gest:edu + bwt:edu + emp.m:edu,
  data = neomod,
  family = quasibinomial)

pchisq(summary(mod.binomial)$dispersion * mod.quasibinomial$df.residual, mod.quasibinomial$df.residual,

## [1] 0.4950646

mod.step <- stepAIC(mod.binomial, direction = "both", trace = FALSE)
mod.step

##
## Call:  glm(formula = re.ad ~ cns + size + emp.f + emp.m + edu + los +
##      sex + accom + cns:emp.f + size:emp.f + emp.m:edu, family = binomial,
##      data = neomod)
##
## Coefficients:
##   (Intercept)          cns1          size1          emp.f1          emp.m1
##      -2.9214      -0.7745       1.2306       0.2431      -1.3089
##      edu2          edu3          edu4          los          sex1
##      -0.4401       0.1254      -0.2185       0.7643       0.4743
##      accom1  cns1:emp.f1  size1:emp.f1  emp.m1:edu2  emp.m1:edu3
##      -0.4440       0.9697      -1.2402       1.3925       0.3625
##      emp.m1:edu4
##       1.0438
##
## Degrees of Freedom: 1487 Total (i.e. Null);  1472 Residual
## Null Deviance:      2031
## Residual Deviance: 1851  AIC: 1883

anova(mod.step, mod.binomial, test = "Chisq")

## Analysis of Deviance Table
##
## Model 1: re.ad ~ cns + size + emp.f + emp.m + edu + los + sex + accom +
##      cns:emp.f + size:emp.f + emp.m:edu

```

```

## Model 2: re.ad ~ cns + size + gest + bwt + emp.f + emp.m + edu + los +
##      sex + accom + cns:emp.f + size:emp.f + size:emp.m + gest:edu +
##      bwt:edu + emp.m:edu
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1472      1850.8
## 2      1452      1822.0 20   28.838  0.09101 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary.glm(mod.step)

##
## Call:
## glm(formula = re.ad ~ cns + size + emp.f + emp.m + edu + los +
##      sex + accom + cns:emp.f + size:emp.f + emp.m:edu, family = binomial,
##      data = neomod)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.0750  -0.9914  -0.6761   1.1112   2.1715
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.92135     0.53208  -5.490 4.01e-08 ***
## cns1          -0.77454     0.36077  -2.147  0.03180 *
## size1         1.23064     0.37916   3.246  0.00117 **
## emp.f1        0.24307     0.35590   0.683  0.49463
## emp.m1       -1.30891     0.56873  -2.301  0.02137 *
## edu2         -0.44006     0.31121  -1.414  0.15735
## edu3          0.12542     0.33830   0.371  0.71083
## edu4         -0.21846     0.40323  -0.542  0.58798
## los           0.76432     0.08051   9.493 < 2e-16 ***
## sex1          0.47432     0.11353   4.178 2.94e-05 ***
## accom1       -0.44403     0.14160  -3.136  0.00171 **
## cns1:emp.f1   0.96974     0.38087   2.546  0.01089 *
## size1:emp.f1 -1.24017     0.40183  -3.086  0.00203 **
## emp.m1:edu2   1.39247     0.58795   2.368  0.01787 *
## emp.m1:edu3   0.36251     0.61228   0.592  0.55381
## emp.m1:edu4   1.04384     0.66234   1.576  0.11503
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2031.3  on 1487  degrees of freedom
## Residual deviance: 1850.9  on 1472  degrees of freedom
## AIC: 1882.9
##
## Number of Fisher Scoring iterations: 4

par(mfrow = c(2,2), mar = c(2.5,2,2,1))
plot(mod.step)

```

