Generalized Linear Models

Shuo Han

Introduction

GLM stands for Generalized Linear Models, which is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. The GLM generalizes linear regression by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

Data analysis

Logistic Model and Probit Model

For this question, we will use the `Chile.txt` dataset, which has a polytomous outcome: voting intention (yes, no, abstain, undecided). For this problem, focus only on the subset of the data with outcomes of either 'yes' or 'no'.

(a) Formulate a model that makes substantive sense in the context of the data set - for example,constructing dummy regressors to represent factors and including interaction regressors where these are appropriate - and fit a linear logistic regression of the response variable on the explanatory variables, reporting the estimated regression coefficients and their asymptotic standard errors.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
data <- read.table("data/Chile.txt")
chi <- data %>% filter(vote == "Y" | vote =="N")
chi <- na.omit(chi)
chi$outcome <- ifelse(chi$vote=="Y",1,0)
logit_1 <- glm(outcome ~ age + income + statusquo +
                education + population + region +
                sex + income * sex + education * sex +
                statusquo * sex, family = binomial(link = "logit"), data = chi)
summary(logit_1)
```

```
##
## Call:
## glm(formula = outcome ~ age + income + statusquo + education +
##     population + region + sex + income * sex + education * sex +
##     statusquo * sex, family = binomial(link = "logit"), data = chi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3257  -0.2694  -0.1403   0.1944   2.9707
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)       1.337e+00  4.869e-01   2.746  0.00604 **
## age               1.433e-03  7.509e-03   0.191  0.84860
## income           -3.720e-07  4.465e-06  -0.083  0.93361
## statusquo         3.330e+00  2.296e-01  14.504  < 2e-16 ***
## educationPS      -1.507e+00  5.368e-01  -2.808  0.00499 **
## educationS       -1.263e+00  3.567e-01  -3.540  0.00040 ***
## population        1.048e-06  1.420e-06   0.738  0.46029
## regionM           6.243e-01  6.155e-01   1.014  0.31044
## regionN          -1.043e-01  3.624e-01  -0.288  0.77354
## regionS          -3.582e-01  2.958e-01  -1.211  0.22591
## regionSA         -3.184e-01  3.436e-01  -0.927  0.35414
## sexM             -1.091e+00  3.545e-01  -3.078  0.00209 **
## income:sexM      -4.090e-06  5.721e-06  -0.715  0.47460
## educationPS:sexM  1.010e+00  6.865e-01   1.472  0.14110
## educationS:sexM   1.145e+00  4.729e-01   2.421  0.01547 *
## statusquo:sexM   -1.185e-01  3.034e-01  -0.390  0.69623
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2360.29  on 1702  degrees of freedom
## Residual deviance:  697.25  on 1687  degrees of freedom
## AIC: 729.25
##
## Number of Fisher Scoring iterations: 6
```

The coefficient Estimate and their standard errors are shown on the table above.

(b) Construct an analysis-of-deviance table for the model fit in part (a).

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
Anova(logit_1)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: outcome
##               LR Chisq Df Pr(>Chisq)
## age               0.04  1   0.848642
## income            0.97  1   0.325811
## statusquo      1458.15  1  < 2.2e-16 ***
## education        10.18  2   0.006170 **
## population        0.55  1   0.460263
## region            3.78  4   0.437105
## sex               7.36  1   0.006662 **
## income:sex        0.51  1   0.475874
## education:sex     6.17  2   0.045633 *
## statusquo:sex     0.15  1   0.695708
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the result, we can notice that Sex, education, intercept"education:sex" and statusquo has the p-values that are less than 0.05, we reject the null hypothesis that they are linearly independent with vote intention, which means that these variables are statistically significant.

(c) Fit a final model to the data that includes the statistically significant effects. Construct an effect display for each high-order term in the model. If the model is additive, (i) suggest two interpretations of each estimated coefficient; and (ii) construct likelihood-ratio-based 95- percent confidence intervals for the regression coefficients, comparing these with confidence intervals based on the Wald statistic.

```
fit2 <- glm(outcome ~ statusquo + education + sex + education * sex,
                      family = binomial(link = "logit"), data = chi)

fit3 <- glm(outcome ~ statusquo + education + sex,
                      family = binomial(link = "logit"), data = chi)


anova(fit3, fit2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: outcome ~ statusquo + education + sex
## Model 2: outcome ~ statusquo + education + sex + education * sex
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1698     708.24
## 2      1696     702.46  2    5.779  0.05561 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

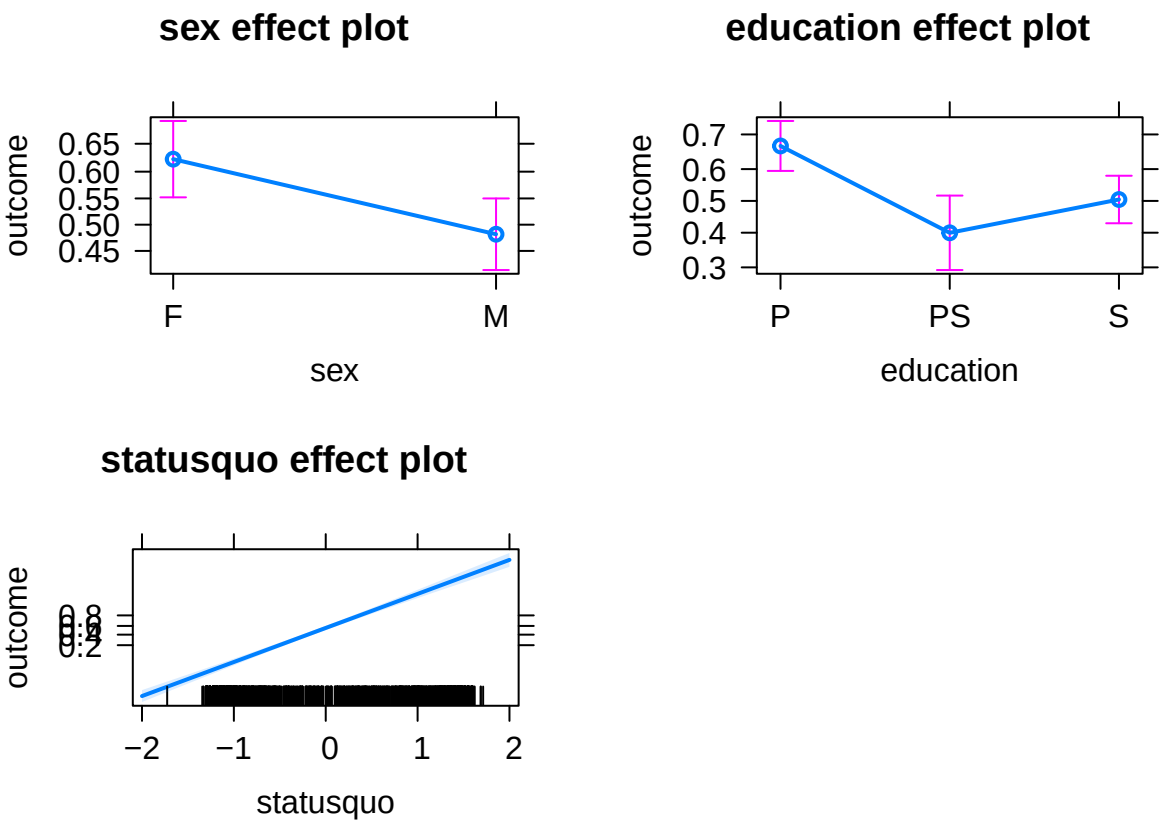
Since we have p-value = 0.05561, we can conclude that the interaction term is not significant, thus we will take the model without interaction education * sex.

Hence our final_model will be fit_3. With sex ,education and statusquo.

```
library(effects)
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
logit_final = glm(outcome~sex+education+statusquo,data=chi,family=binomial(link=logit))
plot(allEffects(logit_final))
```



From the effects plot, Since Y = 1 and N = 0, we can use the number to identify how different groups of people behave here. Here we can see: female has a higher expected probability of vote(Y) than male. People who are primary education have a larger expected probability of vote(Y) than people who are secondary that have a larger expected probability of vote(Y) than people who are post-secondary. As scale of support for the status-quo increases, the probability of vote(Y) increases, which is consistent with results of our model.

```
S(logit_final)
```

```
## Call: glm(formula = outcome ~ sex + education + statusquo, family =
##           binomial(link = logit), data = chi)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.0153     0.1890   5.373 7.75e-08 ***
## sexM         -0.5742     0.2022  -2.840 0.004518 **
## educationPS  -1.1074     0.2914  -3.800 0.000145 ***
```

```
## educationS   -0.6828     0.2217  -3.079 0.002077 **
## statusquo     3.1689     0.1448  21.886  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2360.29  on 1702  degrees of freedom
## Residual deviance:  708.24  on 1698  degrees of freedom
##
##  logLik      df     AIC     BIC
## -354.12       5  718.24  745.44
##
## Number of Fisher Scoring iterations: 6
##
## Exponentiated Coefficients and Confidence Bounds
##                Estimate      2.5 %     97.5 %
## (Intercept)  2.7601091  1.9143823  4.0192553
## sexM         0.5631303  0.3777179  0.8357350
## educationPS  0.3304143  0.1849771  0.5806931
## educationS   0.5052234  0.3259814  0.7786200
## statusquo   23.7820368 18.1196886 31.9915510
```

From the summary of this final model, the follows are the first interpretion:

The coeffient Intercept: $exp(\hat{\beta}_0)$=2.7601091 means that the odds of voting yes are 2.760109, when sex is female, education is primary and scale of support for the status-quo is 0.

The coeffient sexM: $exp(\hat{\beta}_1)$=0.5631303 means that the odds of voting yes are 0.5631303 smaller, when sex is changing to male, education is primary and statusquo remains constant.

The coeffient educationPS: $exp(\hat{\beta}_2)$=0.3304143 means that the odds of voting yes are 0.3304143 smaller, when sex is female, education is changing to post-secondary and statusquo remains constant.

The coeffient educationS: $exp(\hat{\beta}_3)$=0.5052234 means that the odds of voting yes are 0.5052234 smaller, when sex is female, education is changing to secondary and statusquo remains constant.

The coeffient statusquo$exp(\hat{\beta}_4)$=23.7820368 means that the odds of voting yes are 23.7820368 larger, when sex is female, education is primary and the scale of support for the status-quo increases by one unit.

We can also interpret this result by $\beta/4$ rule:

$\hat{\beta}_0/4 = 1.0153/4 = 0.253825$ means that when sex is female, education is primary and scale of support for the status-quo is 0, this corresponds to a positive difference in probability of voting yes of about 25.3825%.

$\hat{\beta}_1/4 = -0.5742/4 = -0.14355$ means that when sex is male, education is primary and statusquo remains constant, this corresponds to a negative difference in probability of voting yes of about 14.355%.

$\hat{\beta}_2/4 = -1.1074/4 = -0.27685$ means that when sex is female, education is post-secondary and statusquo remains constant, this corresponds to a negative difference in probability of voting yes of about 27.685%.

$\hat{\beta}_3/4 = -0.6828/4 = -0.1707$ means that when sex is female, education is secondary and statusquo remains constant, this corresponds to a negative difference in probability of voting yes of about 17.07%.

$\hat{\beta}_4/4 = 3.1689/4 = -0.792225$ means that when sex is female, education is primary and the scale of support for the status-quo increases by one unit, this corresponds to a positive difference in probability of voting yes of about 79.2225%.

```
Confint(logit_final) # LRT CI
```

```
##                Estimate      2.5 %     97.5 %
## (Intercept)  1.0152702  0.6493950  1.3910966
## sexM        -0.5742442 -0.9736077 -0.1794437
## educationPS -1.1074079 -1.6875230 -0.5435329
## educationS  -0.6827546 -1.1209148 -0.2502321
## statusquo    3.1689305  2.8969991  3.4654718
```

```
confint.default(logit_final) # ward CI
```

```
##                   2.5 %     97.5 %
## (Intercept)  0.6449063  1.3856341
## sexM        -0.9706092 -0.1778791
## educationPS -1.6785910 -0.5362247
## educationS  -1.1173685 -0.2481407
## statusquo    2.8851452  3.4527159
```

Then we get two CI of the coefficients. We see that the confidence intervals of LRT is similar with that of ward statistics in confidence bands.

(d) Fit a probit model to the data, comparing the results to those obtained with the logit model. Which do you think is better? Why?

```
probit <- glm(outcome~sex+education+statusquo, data=chi, family = binomial(link = "probit"))
summary(probit)
```

```
##
## Call:
## glm(formula = outcome ~ sex + education + statusquo, family = binomial(link = "probit"),
##     data = chi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4890  -0.2838  -0.0933   0.1887   3.0892
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.50536    0.09845   5.133 2.85e-07 ***
## sexM        -0.31859    0.10457  -3.047 0.002314 **
## educationPS -0.57146    0.14911  -3.833 0.000127 ***
## educationS  -0.32303    0.11497  -2.810 0.004958 **
## statusquo    1.72754    0.06679  25.865  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2360.3  on 1702  degrees of freedom
## Residual deviance:  708.9  on 1698  degrees of freedom
## AIC: 718.9
##
## Number of Fisher Scoring iterations: 6
```

Although both are fine here, I think the logit model is better than the probit model. Because the AIC score of logit model is smaller than that of in the probit model. Also, the coefficients of logit model are easier to interprate.

Multinomial Logistic Regression and Ordinal logistic regression

Proceed as in Exercise D14.1, but now include all of the data and the four possible outcome values.

Use, as appropriate, one or more of the following: a multinomial logit model; a proportional odds logit model; logit models fit to a set of nested dichotomies; or similar probit models. If you fit the proportional-odds model, test the assumption of parallel regressions. If you fit more than one kind of model, which model do you prefer? Why?

```
# multinomial logit model
library(nnet)
chile <- data[!(is.na(data$vote)),]
multi1 = multinom(as.factor(vote) ~ sex+education+statusquo, data=chile)
```

```
## # weights:  24 (15 variable)
## initial  value 3478.212552
## iter  10 value 2142.525403
## iter  20 value 2113.582658
## final  value 2113.581318
## converged
```

```
summary(multi1)
```

```
## Call:
## multinom(formula = as.factor(vote) ~ sex + education + statusquo,
##     data = chile)
##
## Coefficients:
##   (Intercept)        sexM educationPS educationS  statusquo
## N   0.1400749  0.70085893   0.3279190 -0.2190347 -1.7772884
## U   1.8458151 -0.25525004  -1.2132477 -0.8763069  0.3305808
## Y   1.2650561 -0.09240733  -0.5976214 -0.8568123  1.8778761
##
## Std. Errors:
##   (Intercept)      sexM educationPS educationS statusquo
## N   0.1979377 0.1736348   0.2593133  0.1993062 0.1300729
## U   0.1677076 0.1732759   0.2751163  0.1924466 0.1051224
## Y   0.1816155 0.1808594   0.2760477  0.2022784 0.1190310
##
## Residual Deviance: 4227.163
## AIC: 4257.163
```

```
# proportional odds logit model
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
m <- polr(factor(vote) ~ sex + education + statusquo , data = chile)
summary(m)
```

```
##
## Re-fitting to get Hessian

## Call:
## polr(formula = factor(vote) ~ sex + education + statusquo, data = chile)
##
## Coefficients:
##                Value Std. Error t value
## sexM        -0.2653    0.08153  -3.254
## educationPS -0.4054    0.11631  -3.486
## educationS  -0.4095    0.09006  -4.547
## statusquo    1.7733    0.05418  32.731
##
## Intercepts:
##     Value    Std. Error t value
## A|N  -4.0782   0.1175   -34.6971
## N|U  -1.1332   0.0849   -13.3502
## U|Y   0.6890   0.0820     8.4033
##
## Residual Deviance: 4757.931
## AIC: 4771.931
## (23 observations deleted due to missingness)
```

```
poTest(m)
```

```
##
## Tests for Proportional Odds
## polr(formula = factor(vote) ~ sex + education + statusquo, data = chile)
##
##               b[polr]     b[>A]     b[>N]     b[>U] Chisquare df Pr(>Chisq)
## Overall                                                1194.4  8    < 2e-16 ***
## sexM          -0.2653    0.1993   -0.6780   -0.0618      43.4  2    3.8e-10 ***
## educationPS   -0.4054   -0.3259   -1.1900    0.0313      50.7  2    9.7e-12 ***
## educationS    -0.4095   -0.6762   -0.7267   -0.2140      12.1  2     0.0023 **
## statusquo      1.7733    0.1839    2.0360    2.0783     993.5  2    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

I used a multinomial logit model and a proportional odds logit model. I think the multinomial logit model would be more appropriate.

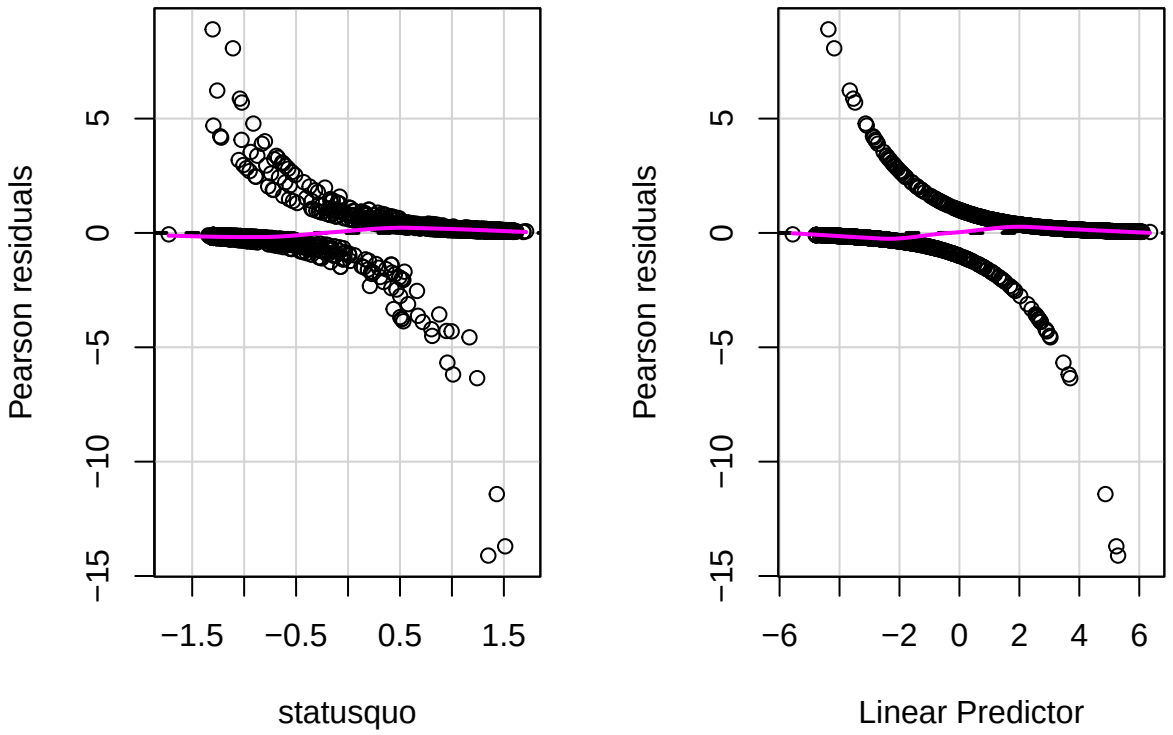
As we can see from the above summaries, first we can find that both AIC and deviances for the multinomial logit model is lower than that for the proportional odds logit model. Besides, to do the poTest, we assume that H0 There's a parallel regression assumption, Since the p-value for the overall test is less than 2e-16, we reject H0, which indicates that the proportional odds logit model is not appropriate. Also, since in this case our data is not nested, thus the logit models fit to a set of nested dichotomies doesn't work well.

GLM Diagnostics

Return to the logit (and probit) model that you fit.

(a) Use the diagnostic methods for generalized linear models described in this chapter to check the adequacy of the final model that you fit to the data.

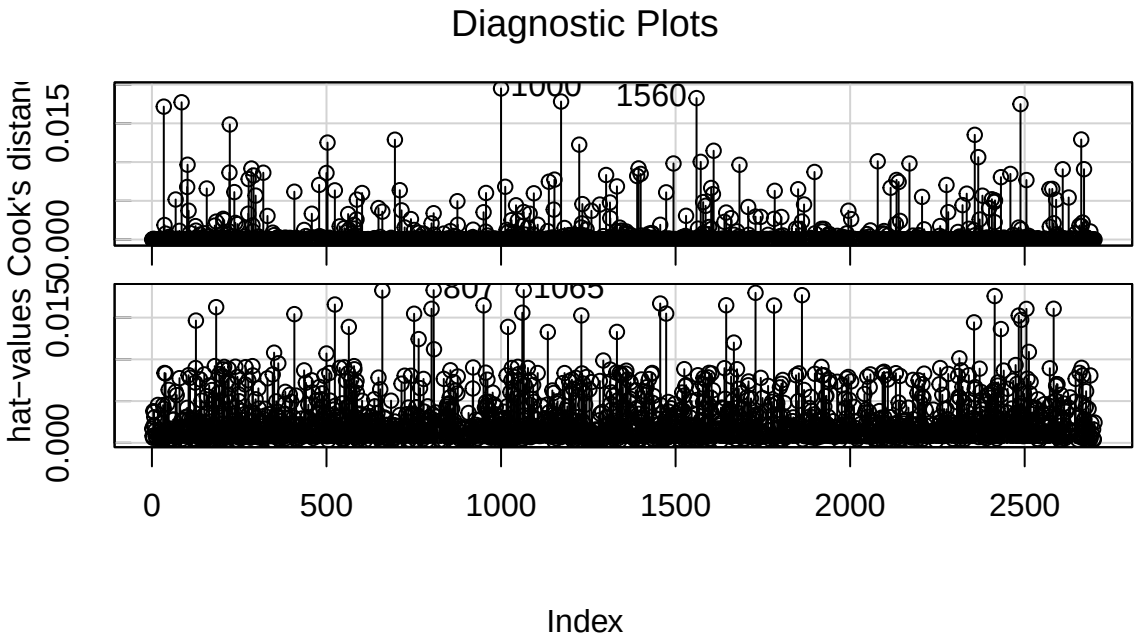
```
residualPlots(logit_final, layout = c(1,2))
```




```
##           Test stat Pr(>|Test stat|)
## statusquo     1.509           0.2193
```

From the plot we can clearly see two curves of deviance residual about linear predictor with asymptotic line to be an horizontal line with y-axis value equals 0 . The reason that the residual plot has such shape is that the response variable have only value 0 or 1 in logistic regression. So for a fixed fitted value, the residual can only take two values. So the residual plot aligns with the nature of logistic regression, but we can't get much insight into the fit of model from such plot.

```
influenceIndexPlot(logit_final, vars = c("Cook","hat"))
```



Firstly, we should check for the influence for the model. By looking at the influence plot of Cook's distance & hat values and added variable plots, 1000th and 1560th may be influential points.

```
compareCoefs(logit_final, update(logit_final, subset=-c(1000,1560)))
```

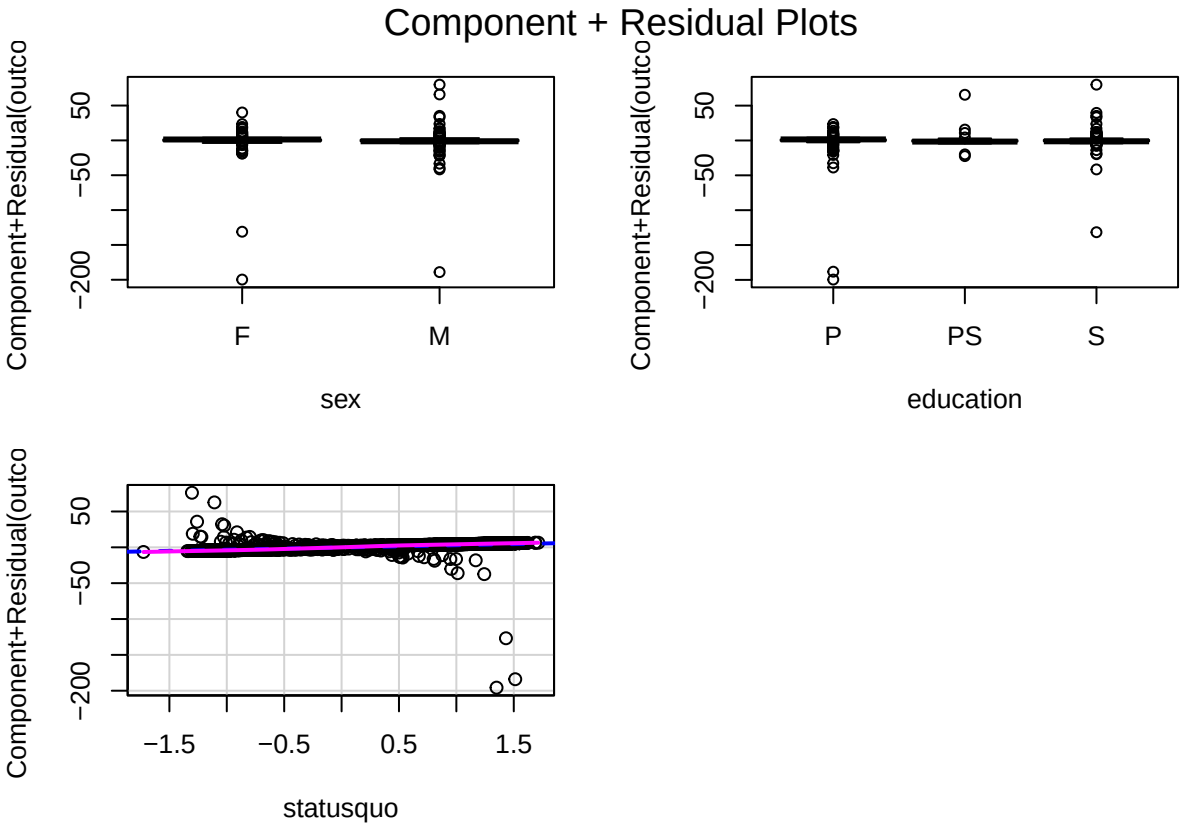
```
## Calls:
## 1: glm(formula = outcome ~ sex + education + statusquo, family =
##   binomial(link = logit), data = chi)
## 2: glm(formula = outcome ~ sex + education + statusquo, family =
##   binomial(link = logit), data = chi, subset = -c(1000, 1560))
##
##             Model 1 Model 2
## (Intercept)   1.015   1.055
## SE            0.189   0.191
##
## sexM         -0.574  -0.596
## SE            0.202   0.203
##
## educationPS  -1.107  -1.134
## SE            0.291   0.293
##
## educationS   -0.683  -0.709
## SE            0.222   0.223
##
## statusquo     3.169   3.184
## SE            0.145   0.146
##
```

```r
# remove influential points
compareCoefs(logit_final, update(logit_final, subset=-c(1000,1560)))
```

```
## Calls:
## 1: glm(formula = outcome ~ sex + education + statusquo, family =
##   binomial(link = logit), data = chi)
## 2: glm(formula = outcome ~ sex + education + statusquo, family =
##   binomial(link = logit), data = chi, subset = -c(1000, 1560))
##
##             Model 1 Model 2
## (Intercept)   1.015   1.055
## SE            0.189   0.191
##
## sexM         -0.574  -0.596
## SE            0.202   0.203
##
## educationPS  -1.107  -1.134
## SE            0.291   0.293
##
## educationS   -0.683  -0.709
## SE            0.222   0.223
##
## statusquo     3.169   3.184
## SE            0.145   0.146
##
```

By comparing their coefficients, coefficients of two models are close. Hence, removing the influential points does not change the results of the analysis that much. Therefore, there's no influential point in the model.

```r
# non-linearity
crPlots(logit_final)
```



From the component-plus-residual plots, since the component line are close to residual lines, there may not be nonlinearity problem.

Therefore, there's no diagnostic alert to our model.

(b) If the model contains a discrete quantitative explanatory variable, test for nonlinearity by specifying a model that treats this variable as a factor (e.g., using dummy regressors), and comparing that model via a likelihood-ratio test to the model that specifies that the variable has a linear effect. (If there is more than one discrete quantitative explanatory variable, then begin with a model that treats all of them as factors, contrasting this with a sequence of models that specifies a linear effect for each such variable in turn.) Note that this is analogous to the approach for testing for nonlinearity in a linear model with discrete explanatory variables described in Section 12.4.1.

Because population can be both treated as an number and a factor. So I change the use of population here:

```
test1 = glm(outcome~sex+statusquo+education+population, family = binomial, data = chi)
test2 = glm(outcome~sex+statusquo+education+as.factor(population), family = binomial, data = chi)

anova(test1, test2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: outcome ~ sex + statusquo + education + population
## Model 2: outcome ~ sex + statusquo + education + as.factor(population)
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1697     708.08
## 2      1689     698.89  8   9.1815   0.3272
```

```
test3 <- glm(outcome ~ statusquo + as.factor(education) + population,
                     family = binomial, data = chi)
anova(test1, test3, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: outcome ~ sex + statusquo + education + population
## Model 2: outcome ~ statusquo + as.factor(education) + population
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1697     708.08
## 2      1698     715.93 -1  -7.8554 0.005067 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As shown above, I think both options work here. Because both options give a significant results for the coefficient of education and population.

(c) Explore the use of the log-log and complementary-log-log links as alternatives to the logit link for this regression. Comparing deviances under the different links, which link appears to best represent the data?

```
# Ignoring the log-log link
model1 = glm(outcome~sex+education+statusquo, family = binomial(link = cauchit), data = chi)
model2 <- glm(outcome~sex+education+statusquo, data=chi, family = binomial(link = "cloglog"))
model_logit = glm(outcome~sex+education+statusquo, family = binomial(link = logit), data = chi)
model_probit = glm(outcome~sex+education+statusquo, family = binomial(link = probit), data = chi)
summary(model1)
```

```
##
## Call:
## glm(formula = outcome ~ sex + education + statusquo, family = binomial(link = cauchit),
##     data = chi)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6501  -0.3305  -0.2748   0.2903   2.5524
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.0511     0.3800   5.398 6.74e-08 ***
## sexM         -0.6024     0.3782  -1.593 0.111216
## educationPS  -2.1326     0.5938  -3.591 0.000329 ***
## educationS   -1.7659     0.4411  -4.003 6.25e-05 ***
## statusquo     6.0697     0.6116   9.925  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2360.29  on 1702  degrees of freedom
## Residual deviance:  744.01  on 1698  degrees of freedom
## AIC: 754.01
```

```
##
## Number of Fisher Scoring iterations: 9
```

```
# compare
deviance(model1)
```

```
## [1] 744.0115
```

```
deviance(model2)
```

```
## [1] 770.2568
```

```
deviance(model_logit)
```

```
## [1] 708.2366
```

```
deviance(model_probit)
```

```
## [1] 708.8987
```

It seems that the cauchit link and cloglog link does really bad job in fitting the data. And after comparing, I think the logit still works best here.

poisson regression, quasi-Poisson and negative-binomial model

Long (1990, 1997) investigates factors affecting the research productivity of doctoral students in biochemistry. Long's data (on 915 biochemists) are in the file `Long.txt`. The response variable in this investigation, `art`, is the number of articles published by the student during the last three years of his or her PhD programme.

The explanatory variables are as follows:

Table 1: *Explanatory variables in 'long.txt' data*

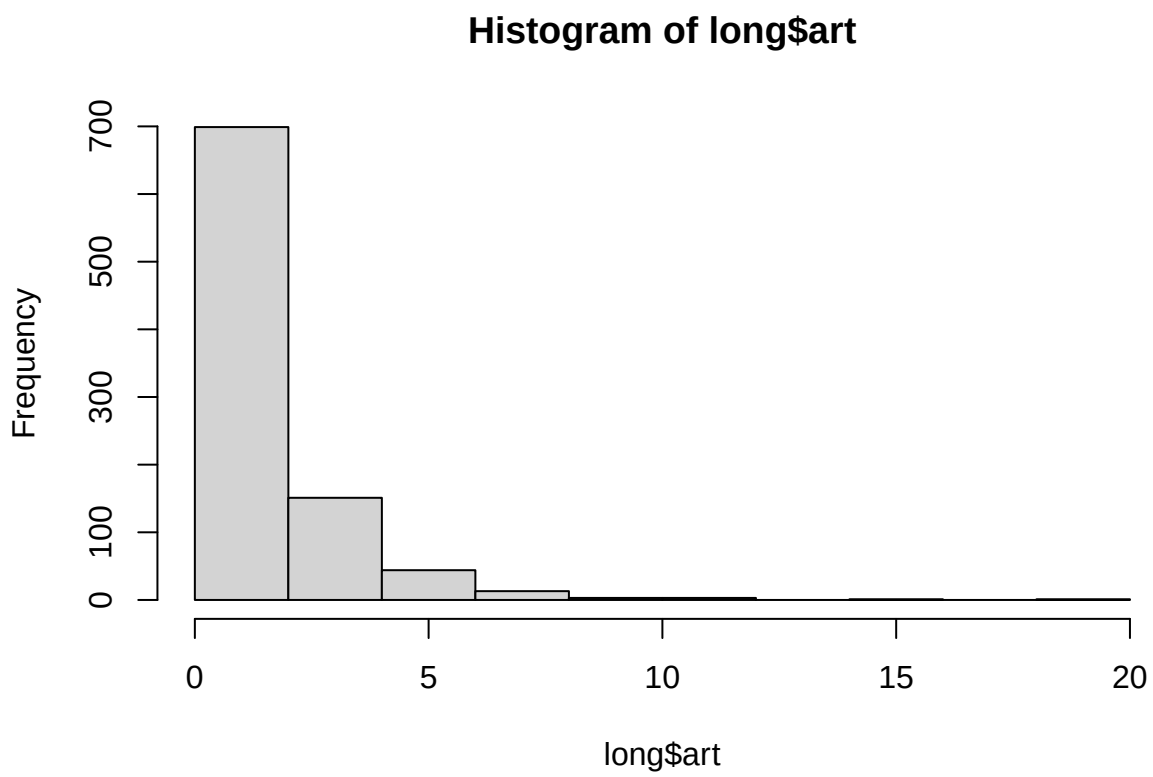
Variable name	Definition
fem	Gender: dummy variable - 1 if female, 0 if male
mar	Maritial status: dummy variable - 1 if married, 0 if not
kid5	Number of children five years old or younger
phd P	restige rating of PhD department
ment	Number of articles published by mentor during last three years

(a) Examine the distribution of the response variable. Based on this distribution, does it appear promising to model these data by linear least-squares regression, perhaps after transforming the response? Explain your answer.

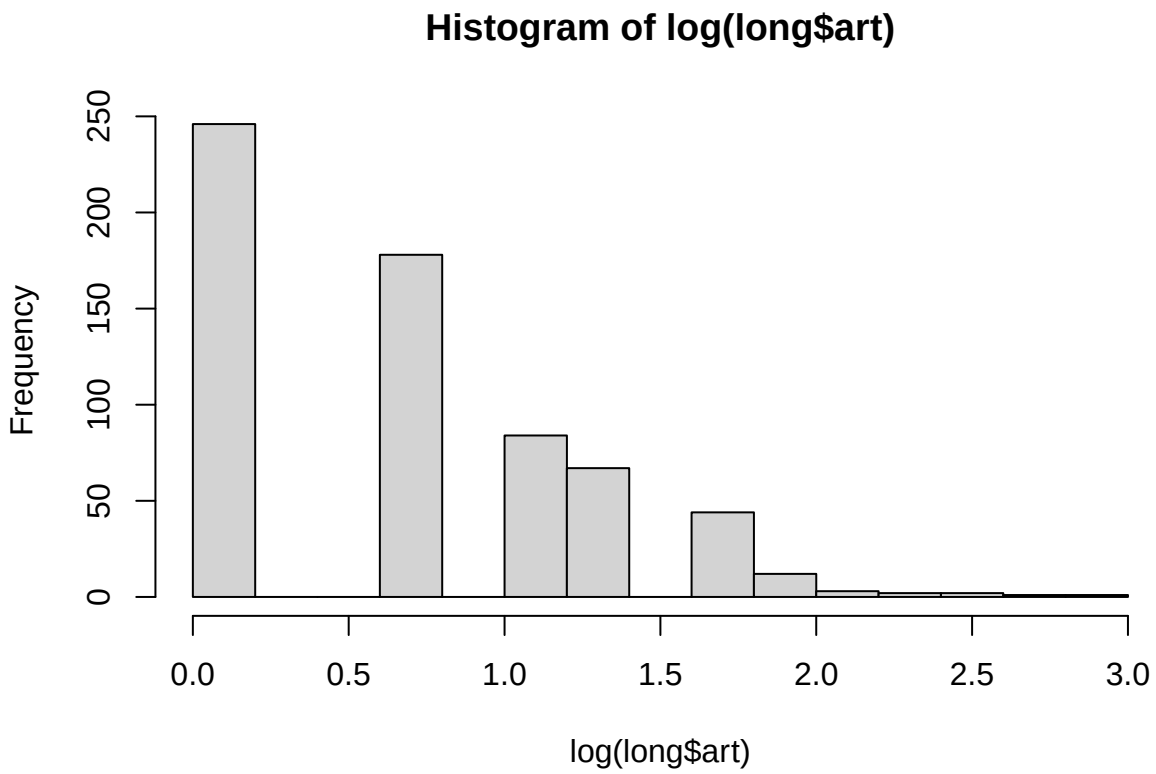
```
long = read.table('data/long.txt',header=TRUE)
lm <- lm(art ~ fem + mar + kid5 + phd + ment, data = long)
summary(lm)
```

```
##
## Call:
## lm(formula = art ~ fem + mar + kid5 + phd + ment, data = long)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0209 -1.2358 -0.4125  0.7517 14.8409
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.334256   0.240906   5.538 3.99e-08 ***
## fem         -0.380886   0.128139  -2.972  0.00303 **
## mar          0.263198   0.145423   1.810  0.07064 .
## kid5        -0.291442   0.090919  -3.206  0.00140 **
## phd         -0.011368   0.063664  -0.179  0.85833
## ment         0.061495   0.006605   9.310  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.822 on 909 degrees of freedom
## Multiple R-squared:  0.1104, Adjusted R-squared:  0.1055
## F-statistic: 22.55 on 5 and 909 DF,  p-value: < 2.2e-16
```

```
hist(long$art)
```



```
hist(log(long$art))
```



After log-transform, we can easily find that the distribution is still discrete and skewed. Therefore, the linear least-squares regression may not be appropriate.

(b) Following Long, perform a Poisson regression of art on the explanatory variables. What do you conclude from the results of this regression?

```
long_poisson = glm(art~fem+mar+kid5+phd+ment, data = long, family = "poisson")
S(long_poisson)
```

```
## Call: glm(formula = art ~ fem + mar + kid5 + phd + ment, family = "poisson",
##           data = long)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.304619   0.102980   2.958   0.0031 **
## fem         -0.224594   0.054613  -4.112 3.92e-05 ***
## mar          0.155243   0.061374   2.529   0.0114 *
## kid5        -0.184883   0.040127  -4.607 4.08e-06 ***
## phd          0.012822   0.026397   0.486   0.6271
## ment         0.025543   0.002006  12.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
##
##   logLik       df      AIC      BIC
## -1651.06        6  3314.11  3343.03
##
## Number of Fisher Scoring iterations: 5
##
## Exponentiated Coefficients and Confidence Bounds
##              Estimate     2.5 %    97.5 %
## (Intercept) 1.3561083 1.1068964 1.6574247
## fem         0.7988403 0.7175304 0.8888577
## mar         1.1679420 1.0358291 1.3176478
## kid5        0.8312018 0.7677989 0.8986167
## phd         1.0129045 0.9619326 1.0668111
## ment        1.0258718 1.0217753 1.0298438
```

From the summary of this model, the follows are the interpretion:

We can interpret the Poisson regression coefficient as follows: for a one unit change in the predictor variable, the difference in the logs of expected counts is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant.

The coeffient Intercept: $exp(\hat{\beta}_0)$=1.3561083 means that the odds of articles published by the student are 1.3561083, when the scale of fem, mar, kid5, phd and ment are all 0 unit.

The coeffient fem: $exp(\hat{\beta}_1)$=0.7988403 means that the Number of articles published by the student will be 0.7988403 times smaller, when Variable fem decrease by 1 unit, and when mar, kid5, phd and ment are fixed.

The coeffient mar: $exp(\hat{\beta}_2)$=1.1679420 means that the Number of articles published by the student are 1.1679420 times bigger, when Variable mar increase by 1 unit, and when fem, kid5, phd and ment are fixed.

The coeffient kid5: $exp(\hat{\beta}_3)$=0.8312018 means that of Number of articles published by the student are 0.8312018 times smaller, when Variable kid5 decrease by 1 unit, and when fem, mar, phd and ment are fixed.

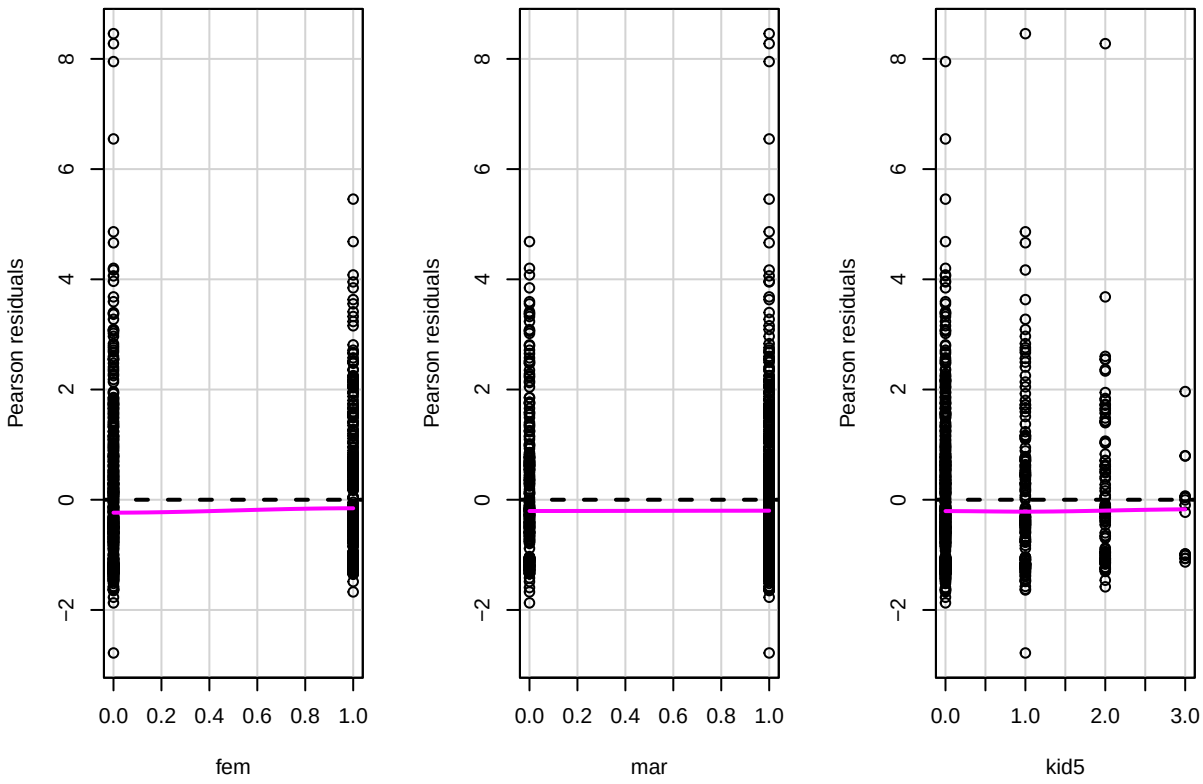
The coeffient phd: $exp(\hat{\beta}_4)$=1.0129045 means that the Number of articles published by the student are 1.0129045 units bigger, when Variable phd increase by 1 unit, and when fem, mar, kid5 and ment are fixed.

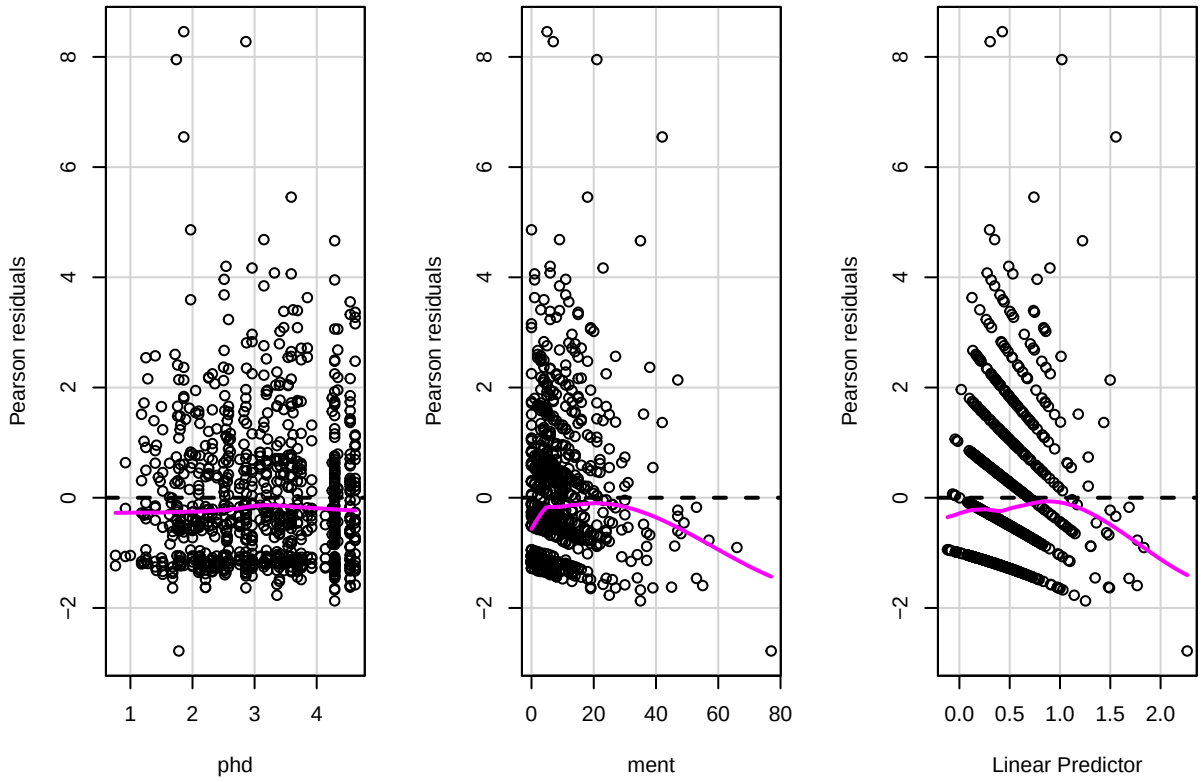
The coeffient ment: $exp(\hat{\beta}_5)$=1.0258718 means that the Number of articles published by the student are 1.0258718 units bigger, when Variable ment increase by 1 unit, and when fem, mar, kid5 and phd are fixed.

From the summary, we could also find that the coefficients for phd variable isn't significant with p = 0.6271.

(c) Perform regression diagnostics on the model fit in the previous question. If you identify any problems, try to deal with them. Are the conclusions of the research altered?

```
residualPlots(long_poisson, layout = c(1,3))
```

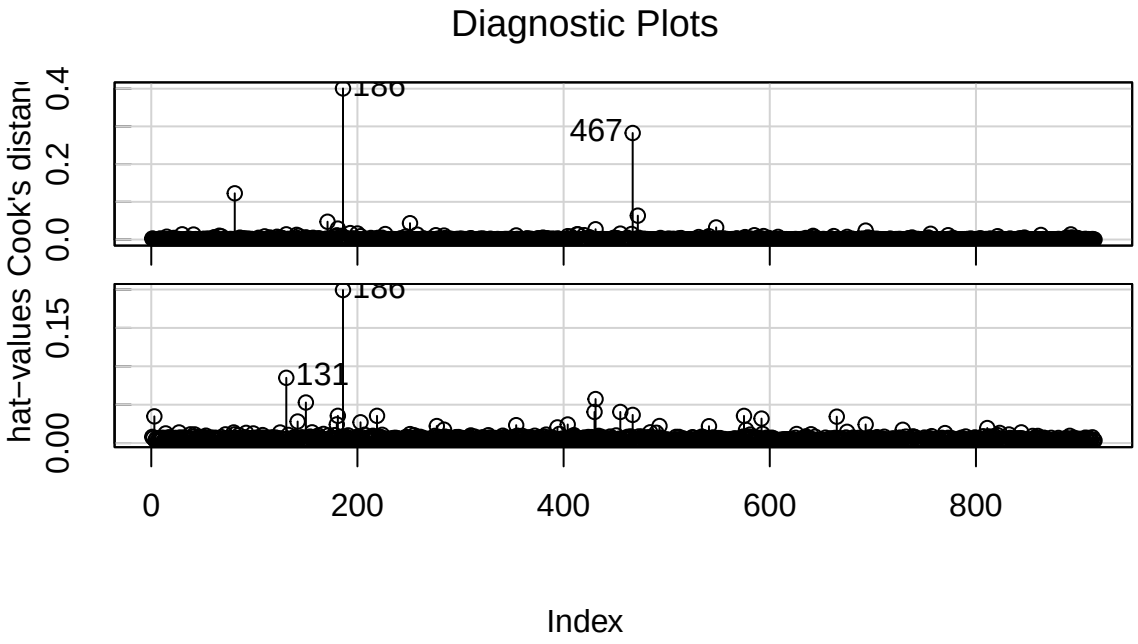

Pearson residuals
−2
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
fem
Pearson residuals
−2
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
mar
Pearson residuals
−2
0
2
4
6
8
0.0
1.0
2.0
3.0
kid5




```
##      Test stat Pr(>|Test stat|)
## fem     0.0000           1.0000
## mar     0.0000           1.0000
## kid5    0.3888           0.5329
## phd     2.3358           0.1264
## ment   36.3491         1.65e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the residual plots for all predictor variables except 'ment' have a red line around 0, the mean of those variables in the model are close to 0.

```
infIndexPlot(long_poisson, var = c("cook","hat"))
```

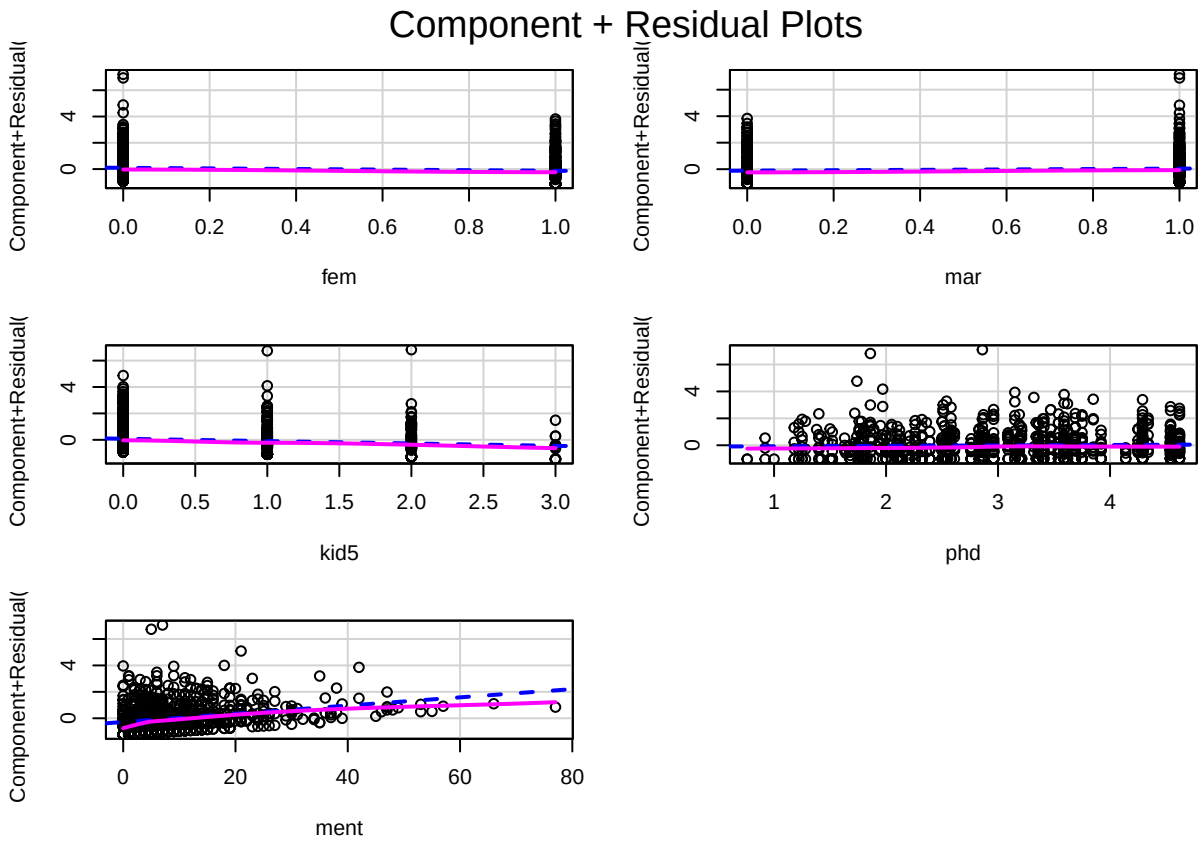



```
compareCoefs(long_poisson, update(long_poisson, subset = -c(186,467)))
```

```
## Calls:
## 1: glm(formula = art ~ fem + mar + kid5 + phd + ment, family = "poisson",
##   data = long)
## 2: glm(formula = art ~ fem + mar + kid5 + phd + ment, family = "poisson",
##   data = long, subset = -c(186, 467))
##
##             Model 1 Model 2
## (Intercept)   0.305   0.284
## SE            0.103   0.104
##
## fem         -0.2246 -0.2124
## SE           0.0546  0.0548
##
## mar          0.1552  0.1396
## SE           0.0614  0.0616
##
## kid5        -0.1849 -0.1650
## SE           0.0401  0.0401
##
## phd          0.0128  0.0155
## SE           0.0264  0.0273
##
## ment        0.02554 0.02600
```

```
## SE          0.00201 0.00224
##
```

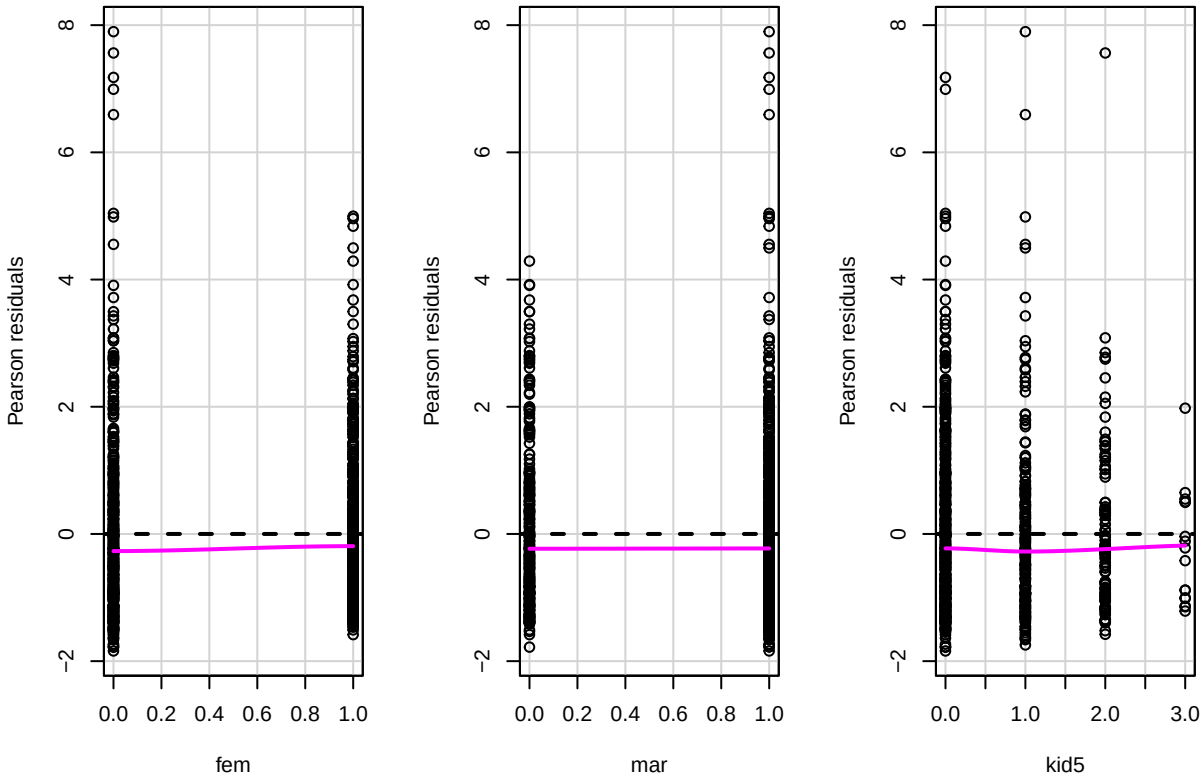
```
crPlots(long_poisson)
```

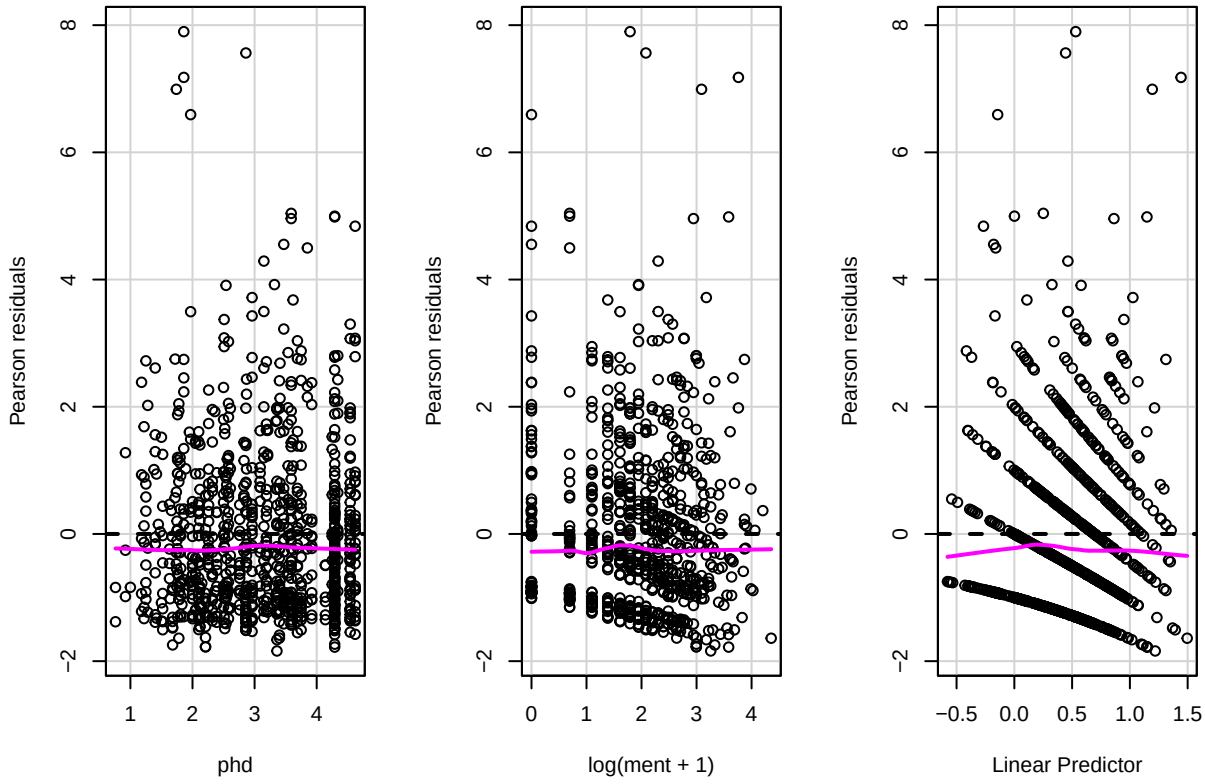


Even though in the influential index plot, we see two possible influential points, but compared to the removed dataset, the coefficients don't change a lot. Therefore, there's no influential point in the model. And from the CR Plots, we find that for each predictor variable in the model, their component+residual centered at line 0. Therefore, we can conclude that there's no non-linearity result in our model.

Here, since ment variable is strange, we will take a log-transformation to it to try to deal with it.

```
poisson <- glm(art ~ fem + mar + kid5 + phd + log(ment+1), family = "poisson", data = long)
residualPlots(poisson, layout = c(1,3))
```

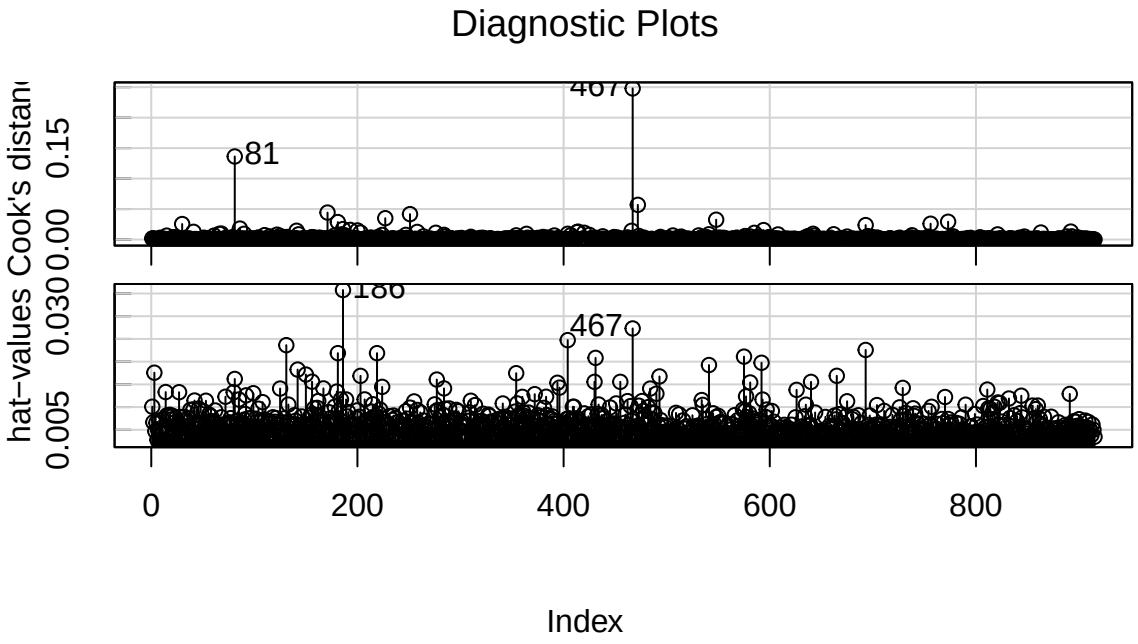

Pearson residuals
fem
mar
kid5




```
##                Test stat Pr(>|Test stat|)
## fem               0.0000          1.00000
## mar               0.0000          1.00000
## kid5              0.6710          0.41271
## phd               0.4401          0.50707
## log(ment + 1)     2.9392          0.08645 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from the graph that the strange situation with ment has greatly improved.

```
infIndexPlot(poisson, var = c("cook","hat"))
```

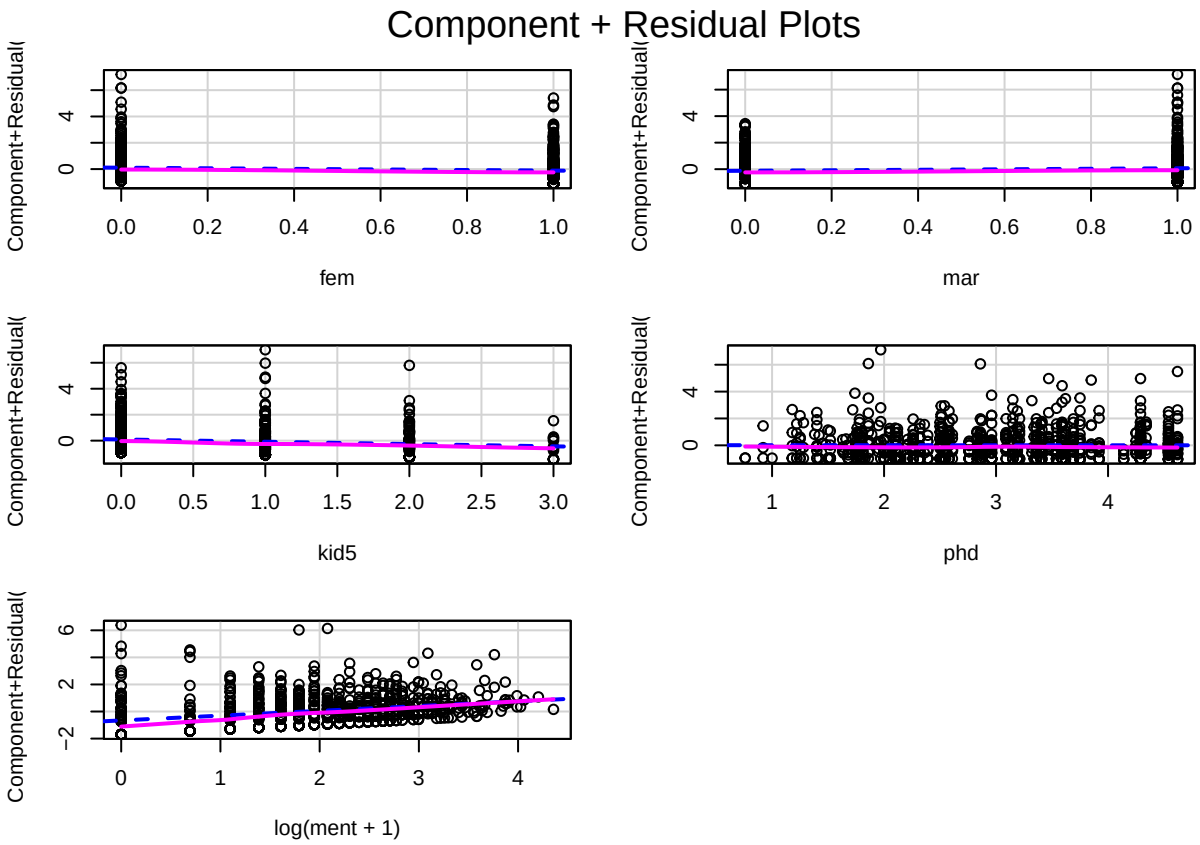



```
compareCoefs(poisson, update(poisson, subset = -c(186,467)))
```

```
## Calls:
## 1: glm(formula = art ~ fem + mar + kid5 + phd + log(ment + 1), family =
##   "poisson", data = long)
## 2: glm(formula = art ~ fem + mar + kid5 + phd + log(ment + 1), family =
##   "poisson", data = long, subset = -c(186, 467))
##
##                Model 1  Model 2
## (Intercept)    -0.0943  -0.1134
## SE              0.1110   0.1115
##
## fem            -0.2339  -0.2189
## SE              0.0545   0.0547
##
## mar             0.1677   0.1488
## SE              0.0613   0.0616
##
## kid5           -0.1722  -0.1521
## SE              0.0399   0.0400
##
## phd           -0.02302 -0.00767
## SE             0.02717  0.02765
##
## log(ment + 1)   0.3755   0.3571
```

```
## SE              0.0296    0.0302
##
```

```
crPlots(poisson)
```



So, the model works quite well after we do the log transfer to ment.

(d) Refit Long's model allowing for overdispersion (using a quasi-Poisson or negative-binomial model). Does this make a difference to the results?

```
long_quasipoisson = glm(art~fem+mar+kid5+phd+ment, data = long, family = "quasipoisson")
long_nb <- glm.nb(art ~ fem + mar + kid5 + phd + ment, data = long)

summary(long_quasipoisson)
```

```
##
## Call:
## glm(formula = art ~ fem + mar + kid5 + phd + ment, family = "quasipoisson",
##     data = long)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5672  -1.5398  -0.3660   0.5722   5.4467
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.304619   0.139271   2.187 0.028979 *
## fem         -0.224594   0.073860  -3.041 0.002427 **
## mar          0.155243   0.083003   1.870 0.061759 .
## kid5        -0.184883   0.054268  -3.407 0.000686 ***
## phd          0.012822   0.035699   0.359 0.719552
## ment         0.025543   0.002713   9.415  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.829006)
##
##     Null deviance: 1817.4  on 914  degrees of freedom
## Residual deviance: 1634.4  on 909  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
summary(long_nb)
```

```
##
## Call:
## glm.nb(formula = art ~ fem + mar + kid5 + phd + ment, data = long,
##     init.theta = 2.264387444, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1678  -1.3617  -0.2806   0.4476   3.4524
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.256143   0.137345   1.865 0.062188 .
## fem         -0.216418   0.072636  -2.979 0.002887 **
## mar          0.150490   0.082097   1.833 0.066790 .
## kid5        -0.176415   0.052813  -3.340 0.000837 ***
## phd          0.015272   0.035872   0.426 0.670313
## ment         0.029082   0.003214   9.048  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(2.2644) family taken to be 1)
##
##     Null deviance: 1109.0  on 914  degrees of freedom
## Residual deviance: 1004.3  on 909  degrees of freedom
## AIC: 3135.9
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  2.264
##           Std. Err.:  0.271
##
##  2 x log-likelihood:  -3121.917
```

```
deviance(long_poisson)
```

```
## [1] 1634.371
```

```
deviance(long_quasipoisson)
```

```
## [1] 1634.371
```

```
deviance(long_nb)
```

```
## [1] 1004.281
```

From the result above, we can find that the coefficients for each predictor variables are closer for all models, but the deviance for the negative-binomial model is much lower than that for the quasi-Poisson model and that of possion model.

Therefore, the negative-binomial model is the better than quasi-Poisson model and that of possion model.