Homework 3

Sarah Adilijiang

Problem 2

(a)

Since the observed mean number of children born per woman in each combination of factors depends on the sample sizes, namely the number of woman in each category, and the sample sizes in each category are different in this problem. Therefore, we should use a rate Poisson model.

To model the rate, i.e. the mean number of children born per woman - "*childrate*" , while still maintaining the count response, i.e. the total number of children of all women in each given combination of factors - "*numchild*", in the Poisson model, we will use the log link connection:

$$\log(childrate) = \log(\frac{numchild}{numwomen}) = \eta \sim years + location + education$$

$$\Rightarrow \quad \log(numchild_i) = 1 \times \log(numwomen_i) + \beta_0 + (\beta_1 to \beta_5)\, years_i + \beta_6\, location_i + (\beta_7 to \beta_9)\, education_i$$

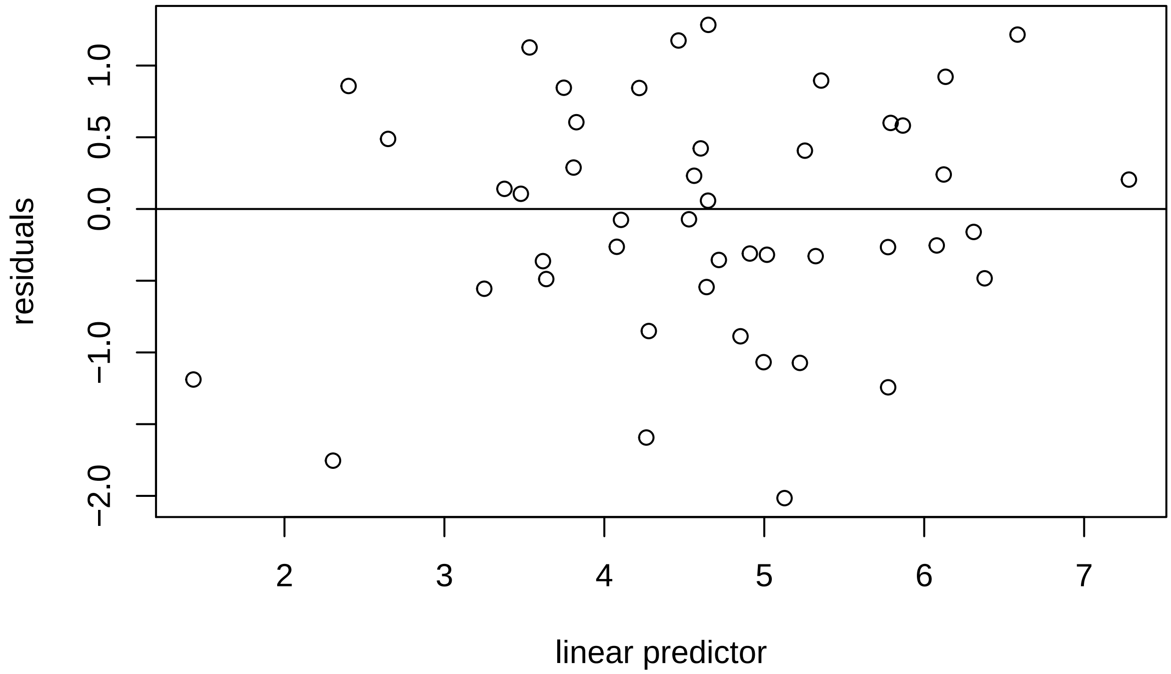
where the coefficients of each levels of factor variables "years", "location", and "education" will be estimated by the model.

```r
# create the dataset
childrate = c(1.17,0.85,1.05,0.69,0.97,0.96,0.97,0.74,
              2.54,2.65,2.68,2.29,2.44,2.71,2.47,2.24,
              4.17,3.33,3.62,3.33,4.14,4.14,3.94,3.33,
              4.70,5.36,4.60,3.80,5.06,5.59,4.50,2.00,
              5.36,5.88,5.00,5.33,6.46,6.34,5.74,2.50,
              6.52,7.51,7.54, NA ,7.48,7.81,5.80, NA )
numwomen = c(12,27,39,51,62,102,107,47,
             13,37,44,21,70,117,81,21,
             18,43,29,15,88,132,50,9,
             23,42,20,15,114,86,30,1,
             22,25,13,3,117,68,23,2,
             46,45,13,0,195,59,10,0)
numchild = round(childrate * numwomen)  # get integer count numbers

years = gl(6,8,48,labels=c("<5","5-9","10-14","15-19","20-24","25+"))
location = gl(2,4,48,labels=c("Urban","Rural"))
education = gl(4,1,48,labels=c("None","Lower","Upper","Secondary"))

child_data = data.frame(childrate, numwomen, years, location, education, numchild)

# fit the rate Poisson model
m.glm_rate = glm(numchild ~ offset(log(numwomen))+years+location+education,
                 child_data, family=poisson)
summary(m.glm_rate)
```

```
##
## Call:
## glm(formula = numchild ~ offset(log(numwomen)) + years + location +
##     education, family = poisson, data = child_data)
##
## Deviance Residuals:
```

```
##      Min        1Q    Median        3Q       Max
## -2.01562  -0.48720  -0.07375   0.55851   1.28367
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -0.08410    0.05876  -1.431   0.1524
## years5-9            0.99715    0.05869  16.990   <2e-16 ***
## years10-14          1.41220    0.05648  25.003   <2e-16 ***
## years15-19          1.66473    0.05628  29.580   <2e-16 ***
## years20-24          1.84435    0.05667  32.546   <2e-16 ***
## years25+            2.03017    0.05546  36.606   <2e-16 ***
## locationRural       0.06107    0.02484   2.458   0.0140 *
## educationLower      0.03750    0.02450   1.531   0.1259
## educationUpper     -0.04710    0.03377  -1.395   0.1631
## educationSecondary -0.21093    0.06147  -3.431   0.0006 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3035.429  on 45  degrees of freedom
## Residual deviance:   29.831  on 36  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 346.3
##
## Number of Fisher Scoring iterations: 4
```

```r
# plot residuals ~ fitted linear predictor
plot(residuals(m.glm_rate)~predict(m.glm_rate),
     xlab="linear predictor", ylab="residuals")
abline(h=0)
```



```
# significance of the rate Poisson model (chi-square LRT)
p_val = 1 - pchisq(deviance(m.glm_rate), df.residual(m.glm_rate)); p_val
```

```
## [1] 0.755961
```

```
# significance of each factor variables
drop1(m.glm_rate, test="Chi")
```

```
## Single term deletions
##
## Model:
## numchild ~ offset(log(numwomen)) + years + location + education
##           Df Deviance     AIC     LRT  Pr(>Chi)
## <none>          29.83  346.30
## years      5  2340.11 2646.58 2310.28 < 2.2e-16 ***
## location   1    35.93  350.39    6.09   0.01356 *
## education  3    51.03  361.50   21.20 9.584e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# dispersion parameter of quasi-Poisson model
m.glm_rate_q = glm(numchild ~ offset(log(numwomen))+years+location+education,
                 child_data, family=quasipoisson)
summary(m.glm_rate_q)$dispersion
```

```
## [1] 0.8014253
```

Model Adequacy:

The residuals against linear predictor plot shows no significant abnormalities in the distribution of residuals. And the likelihood ratio test (chi-square) has a p-value = 0.755961 (against the saturated full model), so our rate Poisson model fits very well. Also, all the three factor variables are shown to be significant predictors.

Further, in the model summary, the residual deviance is 29.831 on 36 degrees of freedom, indicating a dispersion parameter smaller than one. The actual dispersion parameter (0.8014253) is consistent with this diagnostics. Therefore, a quasi-Poisson model is not needed in this case.

(b)

In question (a), we have shown that all the three factor variables are significant predictors. Let's look at the exponentials of their estimated parameters as well.

```r
# exponentials of estimated parameters
exp(coef(m.glm_rate))
```

```
##       (Intercept)          years5-9         years10-14
##         0.9193406         2.7105452          4.1049594
##         years15-19         years20-24           years25+
##         5.2842554         6.3239824          7.6153651
##      locationRural    educationLower     educationUpper
##         1.0629701         1.0382088          0.9539902
## educationSecondary
##         0.8098291
```

Answer:

As shown in question (a), the model is:

$$\log(numchild_i) = 1 \times \log(numwomen_i) + \beta_0 + (\beta_1 to \beta_5)\, years_i + \beta_6\, location_i + (\beta_7 to \beta_9)\, education_i$$

i.e.

$$\log(childrate_i) = \log(\frac{numchild_i}{numwomen_i}) = \beta_0 + (\beta_1 to \beta_5)\, years_i + \beta_6\, location_i + (\beta_7 to \beta_9)\, education_i$$

Take the predictor "location" as an example, which has two levels: "Urban"(reference level) and "Rural". When controlling the other two predictors, we have equation:

$$\log(\frac{childrate_{Rural}}{childrate_{Urban}}) = \log(childrate_{Rural}) - \log(childrate_{Urban}) = \beta_6$$

so

$$\frac{childrate_{Rural}}{childrate_{Urban}} = e^{\beta_6}$$

Hence when $\beta_6 > 0$, i.e. $e^{\beta_6} > 1$, we get $childrate_{Rural} > childrate_{Urban}$, and $childrate_{Rural}$ is e^{β_6} times of the value of $childrate_{Urban}$.

Therefore, as a summary, there are three factor predictors in the model:

"years" has 6 levels with the first level "5-9" being the reference level, so it has 5 parameters from β_1 to β_5, each showing the difference of fertility rate between these levels and the reference level "5-9" when controlling the other two predictors the same. Results of parameters show that as the years since first marriage increases, the fertility rate of women increases.

"location" has 2 levels with the first level "Urban" being the reference level, so it has 1 parameters β_6, showing the difference of fertility rate between level "Rural" and the reference level "Urban" when controlling the other two predictors the same. Result of the parameter shows that the fertility rate of women in rural area is higher than that of women in urban area.

“education” has 4 levels with the first level “None” being the reference level, so it has 3 parameters from β_7 to β_9, each showing the difference of fertility rate between these levels and the reference level “None” when controlling the other two predictors the same. Results of parameters show that the fertility rate of women with lower elementary education is higher than that of women with none education. However, the fertility rate of women with upper elementary education and sencondary or higher education are both lower than that of women with none education. And the the fertility rate of women with sencondary or higher education is even lower than that of women with upper elementary education.

(c)

```r
# years since first marriage: "10-14"
# get the fitted values and standard errors
newdata = data.frame(location="Urban", education="Upper", years="10-14", numwomen=1)
preds = predict(m.glm_rate, newdata, se.fit=TRUE)
fit = preds$fit
se = preds$se.fit

# compute the 95% CI
z = qnorm(0.975,0,1)
log_CI = c(fit-z*se, fit+z*se)
CI = exp(log_CI); CI
```

```
##        1        1
## 3.335226 3.886271
```

Answer:

The 95% confidence interval for the mean number of children born to an urban woman with upper elmentary education and 10-14 years since first marriage is (3.335226,3.886271).

```r
# years since first marriage: "15-19"
# get the fitted values and standard errors
newdata = data.frame(location="Urban", education="Upper", years="15-19", numwomen=1)
preds = predict(m.glm_rate, newdata, se.fit=TRUE)
fit = preds$fit
se = preds$se.fit

# compute the 95% CI
z = qnorm(0.975,0,1)
log_CI = c(fit-z*se, fit+z*se)
CI = exp(log_CI); CI
```

```
##        1        1
## 4.296754 4.998823
```

Answer:

The 95% confidence interval for the mean number of children born to an urban woman with upper elmentary education and 15-19 years since first marriage is (4.296754,4.998823).

```r
# years since first marriage: "20-24"
# get the fitted values and standard errors
newdata = data.frame(location="Urban", education="Upper", years="20-24", numwomen=1)
preds = predict(m.glm_rate, newdata, se.fit=TRUE)
fit = preds$fit
se = preds$se.fit

# compute the 95% CI
```

```r
z = qnorm(0.975,0,1)
log_CI = c(fit-z*se, fit+z*se)
CI = exp(log_CI); CI
```

```
##        1        1
## 5.132516 5.993654
```

Answer:

The 95% confidence interval for the mean number of children born to an urban woman with upper elmentary education and 20-24 years since first marriage is (5.132516,5.993654).

```r
# years since first marriage: "25+"
# get the fitted values and standard errors
newdata = data.frame(location="Urban", education="Upper", years="25+", numwomen=1)
preds = predict(m.glm_rate, newdata, se.fit=TRUE)
fit = preds$fit
se = preds$se.fit

# compute the 95% CI
z = qnorm(0.975,0,1)
log_CI = c(fit-z*se, fit+z*se)
CI = exp(log_CI); CI
```

```
##        1        1
## 6.204893 7.189320
```

Answer:

The 95% confidence interval for the mean number of children born to an urban woman with upper elmentary education and 25+ years since first marriage is (6.204893,7.189320).

(d)

```r
# years since first marriage: "25+"
# get the fitted values and standard errors
newdata = data.frame(location="Rural", education="Secondary", years="25+", numwomen=1)
preds = predict(m.glm_rate, newdata, se.fit=TRUE)
fit = preds$fit;   exp(fit)
```

```
##        1
## 6.026728
```

```r
se = preds$se.fit

# compute the 90% CI
z = qnorm(0.95,0,1)
log_CI = c(fit-z*se, fit+z*se)
CI = exp(log_CI); CI
```

```
##        1        1
## 5.422105 6.698772
```

Answer:

The estimation for the lifetime (25+ years since first marriage) average number of children born to an rural woman with secondary or higher education is about 6 children, and its 90% confidence interval is (5.422105,6.698772).