Assignment 3 STAT 315-463: Multivariable Statistical Methods and Applications

Due date: Friday 28 April 2023

- Your assignment needs to show the R code you used, and your well discussed answers to the questions.
- Submit your assignments on Learn.

Background

In `Contraception315.csv`, you are provided with a dataset, modified from a study originally undertaken to ascertain associations concerning contraceptive use among Bangladeshi women. The data available for this assignment consists of data on 453 women in 5 districts. The predictor variables are

- *use*, an indicator for contraceptive use (coded N for no and Y for yes).
- Two geographical location covariates, *district* (5 levels), and *urban* (2 levels), which should be treated as factor variables.
- A continuous covariate for standardised age *age*

The response variable is the number of living children *livch* (0, 1, 2, 3, 4, 5, 6, 7).

Questions

1. Fit a Poisson Regression including all possible predictor variables. (1 mark)

```r
# Read in data
dataset <- read.csv("Contraception315.csv")
# Convert District variable into factor
dataset$district <- as.factor(dataset$district)
str(dataset)
```

```
## 'data.frame':    453 obs. of  6 variables:
##  $ woman   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ district: Factor w/ 5 levels "1","6","14","25",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ use     : chr  "N" "N" "N" "N" ...
##  $ livch   : int  3 0 2 4 0 0 7 3 1 3 ...
##  $ age     : num  18.44 -5.56 1.44 8.44 -13.56 ...
##  $ urban   : chr  "Y" "Y" "Y" "Y" ...
```

```r
# Fit a Poisson Regression
poisson_full <- glm(formula = livch ~ district + use + urban + age,
                    data = dataset, family = poisson())
summary(poisson_full)
```

```
##
## Call:
## glm(formula = livch ~ district + use + urban + age, family = poisson(),
##     data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6302  -1.2605  -0.2517   0.5293   3.5116
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.7054865  0.0742477   9.502  < 2e-16 ***
## district6   -0.1072500  0.1042029  -1.029  0.30337
## district14  -0.2638122  0.0967032  -2.728  0.00637 **
## district25   0.0006459  0.1038326   0.006  0.99504
## district46  -0.0930164  0.0971626  -0.957  0.33840
## useY         0.3474549  0.0676195   5.138 2.77e-07 ***
## urbanY      -0.1782521  0.0787776  -2.263  0.02365 *
## age          0.0655158  0.0036114  18.141  < 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: 991.36  on 452  degrees of freedom
## Residual deviance: 618.13  on 445  degrees of freedom
## AIC: 1592.7
##
## Number of Fisher Scoring iterations: 5
```

2. Apply either forward or backward selection to determine the most appropriate model for this data. For the chosen model, write down the model equation. (2 marks)

```r
library(MASS)
# Apply the forward selection
poisson_null <- glm(livch ~ 1, data = dataset, family = poisson())
poisson_model1 <- step(poisson_null, direction = "forward",
                      scope = list(lower = poisson_null, upper = poisson_full))
```

```
## Start:  AIC=1951.9
## livch ~ 1
##
##            Df Deviance    AIC
## + age       1   658.37 1620.9
## + urban     1   974.15 1936.7
## + use       1   980.08 1942.6
## + district  4   978.21 1946.8
## <none>          991.36 1951.9
##
## Step:  AIC=1620.91
## livch ~ age
##
##            Df Deviance    AIC
```

```
## + use       1   641.06 1605.6
## + urban     1   651.19 1615.7
## + district  4   646.31 1616.8
## <none>          658.37 1620.9
##
## Step:  AIC=1605.6
## livch ~ age + use
##
##            Df Deviance    AIC
## + urban     1   627.19 1593.7
## + district  4   623.30 1595.8
## <none>          641.06 1605.6
##
## Step:  AIC=1593.73
## livch ~ age + use + urban
##
##            Df Deviance    AIC
## + district  4   618.13 1592.7
## <none>          627.19 1593.7
##
## Step:  AIC=1592.67
## livch ~ age + use + urban + district
```

```
summary(poisson_model1)
```

```
##
## Call:
## glm(formula = livch ~ age + use + urban + district, family = poisson(),
##     data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6302  -1.2605  -0.2517   0.5293   3.5116
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.7054865  0.0742477   9.502  < 2e-16 ***
## age          0.0655158  0.0036114  18.141  < 2e-16 ***
## useY         0.3474549  0.0676195   5.138 2.77e-07 ***
## urbanY      -0.1782521  0.0787776  -2.263  0.02365 *
## district6   -0.1072500  0.1042029  -1.029  0.30337
## district14  -0.2638122  0.0967032  -2.728  0.00637 **
## district25   0.0006459  0.1038326   0.006  0.99504
## district46  -0.0930164  0.0971626  -0.957  0.33840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 991.36  on 452  degrees of freedom
## Residual deviance: 618.13  on 445  degrees of freedom
## AIC: 1592.7
##
## Number of Fisher Scoring iterations: 5
```

Here the forward model selection was used. The most significant variable here is the age of the women, followed by the contraceptive use. The other two less significant variables are living in district 14 and living in the urban.

$$log(livch) = 0.705 + 0.066 * age + 0.347 * useY - 0.264 * district14 - 0.178 * urbanY$$

3. For the model chosen in 2, provide an interpretation of the regression coefficients, on both the link scale and the response scale. Include 95 % confidence intervals (4 marks)

Here, three variables are discussed here, namely, **age**, **use**, and **urban**. Because there is not much information on the district, it is not much considered here.

As we can see from the coefficients of the variables in model chosen in Q2, age and use have positive correlation with the number of living children, whereas the negative sign of urban suggests that it has a negative correlation with the number rate of living children.

```
# Age
age_response <- (exp(poisson_model1$coefficients['age']) - 1) * 100; age_response
```

```
##      age
## 6.77096
```

This indicates that each increase of age is associated with 6.77% more rate of living children.

```
# useY
use_response <- (exp(poisson_model1$coefficients['useY']) - 1) * 100; use_response
```

```
##      useY
## 41.54605
```

```
# 95% confidence interval with SE of 0.0676
(exp(poisson_model1$coefficients['useY'] - 2 * 0.0676) - 1) * 100
```

```
##      useY
## 23.64631
```

```
(exp(poisson_model1$coefficients['useY'] + 2 * 0.0676) - 1) * 100
```

```
##      useY
## 62.03707
```

The number here suggests that the contraceptive use is associated with an increase of 41.55% in living children rate among Bangladeshi women. The 95% confidence interval shows that the use of contraception are expected an increase between 23.65 and 62.04% in the rate of numbers of living children.

```
# urbanY
urban_response <- (exp(-poisson_model1$coefficients['urbanY']) - 1) * 100; urban_response
```

```
##    urbanY
## 19.51266
```

```r
# 95% confidence interval with SE of 0.0788
(exp(-poisson_model1$coefficients['urbanY'] - 2 * 0.0788) - 1) * 100
```

```
##   urbanY
## 2.086684
```

```r
(exp(-poisson_model1$coefficients['urbanY'] + 2 * 0.0788) - 1) * 100
```

```
##   urbanY
## 39.91321
```

This can be interpreted that living in the urban area is associated with a reduction of 19.51% in the rate of the number of living children among Bangladeshi women. Also, from the 95% confidence interval, we can know that there is a decrease between 2.09 and 39.91% of the number rate of living children for the women living in the urban to those who are not.

4. Describe what the implication of unaccounted overdispersion would be for any inference made. Comment on whether you believe overdispersion is present in this dataset. Describe how you would change your analysis to deal with overdispersion. (4 marks)

```r
summary(poisson_model1)
```

```
##
## Call:
## glm(formula = livch ~ age + use + urban + district, family = poisson(),
##     data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6302  -1.2605  -0.2517   0.5293   3.5116
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.7054865  0.0742477   9.502  < 2e-16 ***
## age          0.0655158  0.0036114  18.141  < 2e-16 ***
## useY         0.3474549  0.0676195   5.138 2.77e-07 ***
## urbanY      -0.1782521  0.0787776  -2.263  0.02365 *
## district6   -0.1072500  0.1042029  -1.029  0.30337
## district14  -0.2638122  0.0967032  -2.728  0.00637 **
## district25   0.0006459  0.1038326   0.006  0.99504
## district46  -0.0930164  0.0971626  -0.957  0.33840
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 991.36  on 452  degrees of freedom
## Residual deviance: 618.13  on 445  degrees of freedom
## AIC: 1592.7
##
## Number of Fisher Scoring iterations: 5
```

Overdispersion can be detected by dividing the residual deviance by the degrees of freedoms. Here we can see that there is overdispersion present in this dataset. One potential approach to deal with overdispersion is to use negative binomial model, which explicitly models the overdispersion that happens in the dataset.

5. You are later told that the researchers in comparisons between women without children and women with children. Create a new variable `child` such that a `0` indicates a women with no living children, and `1` indicates a women with at least one living child. (1 mark)

```
library(dplyr)
dataset1 <- dataset %>%
  mutate(child = case_when(livch == 0 ~ 0,
                           livch > 0 ~ 1))
head(dataset1)
```

```
##   woman district use livch      age urban child
## 1     1        1   N     3  18.4400     Y     1
## 2     2        1   N     0  -5.5599     Y     0
## 3     3        1   N     2   1.4400     Y     1
## 4     4        1   N     4   8.4400     Y     1
## 5     5        1   N     0 -13.5590     Y     0
## 6     6        1   N     0 -11.5600     Y     0
```

6. Fit a Logistic Regression including all possible predictor variables and `child` as the response. (1 mark)

```
lr_full <- glm(child ~ district + use + age + urban, data = dataset1, family = binomial())
summary(lr_full)
```

```
##
## Call:
## glm(formula = child ~ district + use + age + urban, family = binomial(),
##     data = dataset1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.8275  -0.5029   0.2166   0.5870   2.0520
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.41356    0.40464   5.965 2.45e-09 ***
## district6   -0.58600    0.49453  -1.185 0.236033
## district14  -0.49151    0.39921  -1.231 0.218252
## district25  -0.22601    0.47423  -0.477 0.633661
## district46  -0.85674    0.47970  -1.786 0.074100 .
## useY         1.18373    0.30810   3.842 0.000122 ***
## age          0.26015    0.02926   8.892  < 2e-16 ***
## urbanY      -0.89039    0.36311  -2.452 0.014202 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.43  on 452  degrees of freedom
```

```
## Residual deviance: 330.90  on 445  degrees of freedom
## AIC: 346.9
##
## Number of Fisher Scoring iterations: 6
```

7. Apply either forward or backward selection to determine the most appropriate model for this data. For the chosen model, write down the model equation. Were the same predictor variables chosen as the Poisson regression (3 marks)

```r
lr_null<- glm(child ~ 1, family = binomial(), data = dataset1)
# Apply the backward selection
lr1 <- stepAIC(lr_full, direction = 'backward',
                scope = list(upper = lr_full,lower = lr_null))
```

```
## Start:  AIC=346.9
## child ~ district + use + age + urban
##
##            Df Deviance    AIC
## - district  4   334.93 342.93
## <none>          330.90 346.90
## - urban     1   337.06 351.06
## - use       1   346.67 360.67
## - age       1   487.79 501.79
##
## Step:  AIC=342.93
## child ~ use + age + urban
##
##         Df Deviance    AIC
## <none>       334.93 342.93
## - urban  1   340.99 346.99
## - use    1   347.89 353.89
## - age    1   495.00 501.00
```

```r
summary(lr1)
```

```
##
## Call:
## glm(formula = child ~ use + age + urban, family = binomial(),
##     data = dataset1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7498  -0.5039   0.2276   0.5950   1.9494
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.97780    0.26611   7.432 1.07e-13 ***
## useY         1.02452    0.29181   3.511 0.000446 ***
## age          0.26172    0.02916   8.976  < 2e-16 ***
## urbanY      -0.69073    0.28373  -2.434 0.014916 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 515.43  on 452  degrees of freedom
## Residual deviance: 334.93  on 449  degrees of freedom
## AIC: 342.93
##
## Number of Fisher Scoring iterations: 6
```

Here the backward selection was used. The chosen variables were quite similar to the previous model. The equation in this case is shown below:

$$Logit[child] = 1.978 + 0.262 * age + 1.024 * useY - 0.691 * urbanY$$

8. If Y is Poisson distributed with parameter λ, then $\Pr(Y > 0) = 1 - e^{-\lambda}$. Use this to construct estimates of the probability of having at least one living child for each women in the study from the Poisson regression model chosen in part 2. Compare this to the estimate of the probability of having at least one living child for each women in the study using the Logistic regression model chosen in part 7. (Note: Due to the number of women in the study, do not provide a table of all the estimated probabilities. Instead focus on graphical visualizations to determine any similarities and differences.) (4 marks)

```r
pred_glm <- predict(poisson_model1, type = "response")
prob_glm <- (1 - exp(-pred_glm)) * 100
summary(prob_glm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   43.54   69.84   83.95   81.33   94.77   99.99
```

```r
pred_lr <- predict(lr1, type="response")
prob_lr <- exp(pred_lr) / (1+exp(pred_lr)) * 100
summary(prob_lr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52.36   63.76   70.31   67.53   72.60   73.10
```

From the two summaries above, it is noticeable that the estimates of probabilities using poisson model are much higher than the ones using logistic regression.