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)

```
# Read in data
dataset <- read.csv("Contraception315.csv")
# Convert District variable into factor
dataset$district <- as.factor(dataset$district)
dataset$urban <- as.factor(dataset$urban)
# Fit a Poisson Regression
poisson_model <- glm(formula = livch ~ district + use + urban + age,
                     data = dataset, family = poisson)
summary(poisson_model)
```

```
##
## 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)

```
library(MASS)
# The minimum model
poisson_model1 <- glm(formula = livch ~ 1, family = poisson, data = dataset)
# Apply the backward selection
poisson_back <- stepAIC(poisson_model, direction = 'backward',
                scope = list(upper = poisson_model,lower = poisson_model1))
```

```
## Start:  AIC=1592.67
## livch ~ district + use + urban + age
## 
##            Df Deviance    AIC
## <none>          618.13 1592.7
## - district  4   627.19 1593.7
## - urban     1   623.30 1595.8
## - use       1   644.47 1617.0
## - age       1   946.05 1918.6
```

```
summary(poisson_back)
```

```
## 
## 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
```

The backward selection suggested three variables: *use*, *urban*, and *age*. The equation is shown as below:

$$Livch = 0.65 + 0.31 * use - 0.24 * urban + 0.06 * age$$

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)

The coefficient of variable **Use** is 0.28, with the standard deviation of 0.06350. \ Use: $e^{0.31} = 1.36$ \ Urban: $e^{-0.24} = 0.79$ \ Age: $e^{0.06} = 1.06$

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)

```
summary(poisson_back)
```

```
## 
## 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
```

Overdispersion can be detected by dividing the residual deviance by the degrees of freedoms.

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)

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

```
##
## 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)

```
model_lr1 <- glm(child ~ 1, family = binomial(), data = dataset1)
# Apply the backward selection
model_lr_back <- stepAIC(model_lr, direction = 'backward',
                scope = list(upper = model_lr,lower = model_lr1))
```

```
## 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
```

```
summary(model_lr_back)
```

```
##
## 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
```

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)