Assignment 2

Milan Kuzmanovic, Mark McMahon
Martin Kotuliak, Jakub Polak

April 21, 2020

Task 1

The following table contains the estimates of a logistic regression model.

						95% C.I. for OR	
	Est.	s.e.	z	p-value	OR	lower	higher
X_1	-0.868	**0.367**	-2.365	0.018	**0.42**	0.205	0.865
X_2	2.404	0.601	4.000	<0.001	**11.067**	**3.408**	**35.943**
X_3	**-3.604**	**0.511**	**-7.059**	<0.001	**0.027**	0.010	0.074

Fill in the missing information (Please report formulas and computation.)

$$\begin{aligned}
\text{s.e.}_{X_1} &= \frac{X_1}{z_{X_1}} = \frac{-0.868}{-2.365} = 0.367 \\
\text{OR}_{X_1} &= e^{X_1} = e^{-0.868} = 0.42 \\
\text{OR}_{X_2} &= e^{X_2} = e^{2.404} = 11.067 \\
\text{lower}_{X_2} &= e^{X_2 - z_{0.975} \times \text{s.e.}_{X_2}} = e^{2.404 - 1.96 \times 0.601} = 3.408 \\
\text{lower}_{X_2} &= e^{X_2 + z_{0.975} \times \text{s.e.}_{X_2}} = e^{2.404 + 1.96 \times 0.601} = 35.943 \\
X_3 &= \frac{\ln \text{lower}_{X_3} + \ln \text{higher}_{X_3}}{2} = \frac{0.01 + \ln 0.074}{2} = -3.604 \\
\text{s.e.}_{X_3} &= \frac{-\ln \text{lower}_{X_3} + \ln \text{higher}_{X_3}}{2 \times z_{0.975}} = \frac{-0.01 + \ln 0.074}{2 \times 1.96} = 0.511 \\
\text{z}_{X_3} &= \frac{X_3}{\text{s.e.}_{X_3}} = \frac{-3.604}{0.511} = -7.059 \\
\text{OR}_{X_3} &= e^{X_3} = e^{-3.604} = 0.027
\end{aligned}$$

Task 2

During the lecture, we have considered three systems of hypotheses for the parameters of the MNRM:

1. $H_0 : \beta_{jm} = 0 \quad vs. \quad H_1 : \beta_{jm} \neq 0$
2. $H_0 : \beta_{j1} = ... = \beta_{j(M-1)} = 0 \quad vs. \quad H_1 : \textit{at least one } \beta_{jm} \neq 0, \ \forall m$
3. $H_0 : \beta_{j1} = ... = \beta_{j(M-1)} = 0 \quad vs. \quad H_1 : \textit{at least one } \beta_{jm} \neq 0, \ \forall m, j$

Could you specify another pair of hypotheses H_0 and H_1 for the parameters of the MNRM that we might want to test? Justify your answer.

Firstly, it is important to explain which hypotheses can be tested, i.e. for which hypotheses is there a procedure to define a valid (asymptotic) test. The above hypotheses are tested either based on the asymptotic normality of the maximum likelihood estimator, or on the likelihood ratio statistic whose logarithm is asymptotically Chi-square distributed. It is important to notice that we can extend the hypothesis test $H_0 : \beta_{jm} = 0 \quad vs. \quad H_1 : \beta_{jm} \neq 0$ that is based on the asymptotic normality of the ML estimator to a joint test for any set of parameters. We can do this because the ML estimator is jointly normally distributed, so any subset of the set of all parameters is also jointly normally distributed, which means that by taking the quadratic form of the estimator minus hypothesized value and covariance matrix estimated by the inverse of the Fisher information matrix, we can define a Chi-square test statistic for any null hypotheses that an arbitrary subset of parameters is simultaneously equal to some specific hypothesized values (generally to zero). Therefore, we can in general construct a test for any joint hypothesis on the parameters. Now comes the question of which hypotheses might be of interest, i.e. which questions we might want an answer to? We propose several possible questions of interest defined by the below hypotheses:

1. $H_0 : \beta_{jm} = c \quad vs. \quad H_1 : \beta_{jm} \neq c$ where $c \in \mathbb{R}$ is some constant. Here we have the same logic of the test as for $c = 0$. In fact, this is just a generalization that is useful if we want to answer more specific questions about the parameters. We might be interested in testing not just for the existence of a significant effect ($c = 0$ case), but also about specific magnitude of the effect (e.g. $c = 1$).

2. $H_0 : \beta_{jm} = 0 \quad vs. \quad H_1 : \beta_{jm} > 0$. We can change the alternative hypothesis in case we want to test whether there is a positive effect. Here, we assume that the parameter can't be negative, only zero under the null hypothesis or positive under the alternative hypothesis. This version of the test with different, so-called one-sided alternative, gives us more power in detecting positive (or negative in case $H_1 : \beta_{jm} < 0$) effects because we assume that the effect is either zero or positive, so the rejection region for given significance α is two times larger on the positive (or negative) side compared to the case with the two-sided alternative.

3. $H_0 : \beta_{j1} = ... = \beta_{j(M-1)} = c \quad vs. \quad H_1 : \textit{at least one } \beta_{jm} \neq c, \forall m$, where $c \in \mathbb{R}$ is some constant. This is a variation of the second hypothesis test that was introduced in the lecture. Instead of testing if all of the parameters associated with variable j are zero, which would mean that variable j has no significant influence on the outcome variable, we can test that all of the parameters associated with variable j have the same effect $c \in \mathbb{R}$, on the odds relative to the reference group M. For example, having three groups, and group 3 as the reference, we might be interested if the increase in variable j results in the same change in the odds for group 1 relative to group 3 and for group 2 relative to group 3.

4. $H_0 : \beta_{1m} = ... = \beta_{pm} = 0 \quad vs. \quad H_1 : \textit{at least one } \beta_{jm} \neq 0, \forall j.$ This hypothesis test would test if the variables in the model have any influence on the odds between group m and reference group M. We might be interested in this question if we suspect that for a specific group m the variables in the model don't affect the odds of m happening relative to M happening.

5. $H_0 : \beta_{j_1 m} - \beta_{j_2 m} = 0 \quad vs. \quad H_1 : \beta_{j_1 m} - \beta_{j_2 m} \neq 0$ This hypothesis test would test if the effect of the variable j_1 and the variable j_2 on the odds of group m relative to reference group M is the same.

There are many more possibilities for hypotheses tests. So, if we have a question of interest we can define a hypothesis test for that question, and likelihood theory allows us to construct asymptotic pivots for testing all sorts of hypotheses and thus answering all sorts of questions of interest.

Task 4

Q1. Write down the formulation of the MNRM assuming "fish" as reference category

The general form can be written as:

$$logit[\pi_m(x)] = log[\frac{\pi_m(x)}{\pi_M(x)}] = \beta_{0m} + \beta_{1m}X_{lakeHancock} + \beta_{2m}X_{lakeOklawaha} \\ + \beta_{3m}X_{lakeTrafford} + \beta_{4m}X_{sexMale} + \beta_{5m}X_{sizeSmall}$$

Where m ∈ {invertebrate, reptile, bird, other} and M = fish.

For example, using reptile we have:

$$logit[\pi_{reptile}(x)] = log[\frac{\pi_{reptile}(x)}{\pi_{fish}(x)}] = \beta_{0,reptile} + \beta_{1,reptile}X_{lakeHancock} + \beta_{2,reptile}X_{lakeOklawaha} \\ + \beta_{3,reptile}X_{lakeTrafford} + \beta_{4,reptile}X_{sexMale} + \beta_{5,reptile}X_{sizeSmall}$$

Q2. Estimate the model using R. Interpret the parameters of the log-odds of reptile vs fish

```
fit <- multinom(food ~ lake + sex + size, data, weights = count)

## # weights:  35 (24 variable)
## initial  value 352.466903
## iter  10 value 270.967533
## iter  20 value 268.934907
## final  value 268.932740
## converged
```

```
print(summary)

##                    logit        param   Est. Std..Errors z.stat p.value
## 1          Bird vs. Fish  (Intercept) -1.702       0.769 -2.213   0.027
## 2          Bird vs. Fish  lakeHancock  0.575       0.795  0.723   0.469
## 3          Bird vs. Fish lakeOklawaha -0.551       1.210 -0.455   0.649
## 4          Bird vs. Fish lakeTrafford  1.237       0.866  1.428   0.153
## 5          Bird vs. Fish      sexmale -0.606       0.689 -0.880   0.379
## 6          Bird vs. Fish    sizesmall -0.730       0.652 -1.120   0.263
## 7  Invertebrate vs. Fish  (Intercept) -1.167       0.534 -2.187   0.029
## 8  Invertebrate vs. Fish  lakeHancock -1.781       0.623 -2.857   0.004
## 9  Invertebrate vs. Fish lakeOklawaha  0.913       0.476  1.918   0.055
## 10 Invertebrate vs. Fish lakeTrafford  1.156       0.493  2.345   0.019
## 11 Invertebrate vs. Fish      sexmale -0.463       0.396 -1.171   0.242
## 12 Invertebrate vs. Fish    sizesmall  1.336       0.411  3.250   0.001
## 13        Other vs. Fish  (Intercept) -1.721       0.631 -2.726   0.006
## 14        Other vs. Fish  lakeHancock  0.767       0.569  1.348   0.178
## 15        Other vs. Fish lakeOklawaha  0.026       0.778  0.033   0.973
## 16        Other vs. Fish lakeTrafford  1.558       0.626  2.490   0.013
## 17        Other vs. Fish      sexmale -0.253       0.466 -0.542   0.588
## 18        Other vs. Fish    sizesmall  0.291       0.460  0.632   0.528
## 19      Reptile vs. Fish  (Intercept) -2.859       1.146 -2.496   0.013
## 20      Reptile vs. Fish  lakeHancock  1.129       1.193  0.947   0.344
## 21      Reptile vs. Fish lakeOklawaha  2.530       1.122  2.255   0.024
## 22      Reptile vs. Fish lakeTrafford  3.061       1.130  2.710   0.007
## 23      Reptile vs. Fish      sexmale -0.628       0.685 -0.916   0.360
## 24      Reptile vs. Fish    sizesmall -0.557       0.647 -0.862   0.389
```

Above is the table to estimated parameters for the MNRM, using fish as the reference category. Below is a subset of this table with just the parameters for reptile vs fish, and their respective statistics:

```
print(summary[summary[,"logit"]=="Reptile vs. Fish",])

##              logit        param   Est. Std..Errors z.stat p.value
## 19 Reptile vs. Fish  (Intercept) -2.859       1.146 -2.496   0.013
## 20 Reptile vs. Fish  lakeHancock  1.129       1.193  0.947   0.344
## 21 Reptile vs. Fish lakeOklawaha  2.530       1.122  2.255   0.024
## 22 Reptile vs. Fish lakeTrafford  3.061       1.130  2.710   0.007
## 23 Reptile vs. Fish      sexmale -0.628       0.685 -0.916   0.360
## 24 Reptile vs. Fish    sizesmall -0.557       0.647 -0.862   0.389
```

Now we can interpret these as follows:

$\beta_{0,reptile}$ = -2.859: This indicates that the odds that the food of choice of a large female alligator from lake George is a reptile is expected to change by a factor of $e^{-2.859}$ when compared to the odds that the same alligator's food choice is fish.

$\beta_{1,reptile}$ = 1.129: This indicates that the odds for reptile relative to fish is expected to change by a factor of $e^{1.129}$ if the lake variable is equal to Lake Hancock, while controlling for all other variables in the model. The p-value for this level is given as 0.344, but this should not lead us to belief that this whole variable is not significant as this is just one level in the categorical variable. We would need to do a global test for significance by removing this variable from the model fit and comparing the two models to get a reading on the significance of this variable. The same can be said for the remaining two levels of this lake variable below.

$\beta_{2,reptile}$ = 2.530: This indicates that the odds for reptile relative to fish is expected to change by a factor of $e^{2.530}$ if the lake variable is equal to Lake Oklawaha, while controlling for all other variables in the model

$\beta_{3,reptile}$ = 3.061: This indicates that the odds for reptile relative to fish is expected to change by a factor of $e^{3.061}$ if the lake variable is equal to Lake Trafford, while controlling for all other variables in the model

$\beta_{4,reptile}$ = -0.628: This indicates that the odds for reptile relative to fish is expected to change by a factor of $e^{-0.628}$ if the sex of the alligator is male, while controlling for all other variables in the model. The p-value for this variable can be read directly as 0.360 which, if using the usual 5% level of signifcance, would lead to us failing to reject the hypothesis that the value of this parameter is equal to zero, given all the other variables in the model.

$\beta_{5,reptile}$ = -0.557: This indicates that the odds for reptile relative to fish is expected to change by a factor of $e^{-0.557}$ if alligator is classes as small in size, while controlling for all other variables in the model. Again, this p-value can be read directly as 0.389 and again, this would lead to us failing to reject the hypothesis that the value of this parameter is equal to zero, given all the opther variables in the model.

Q3. What is the odds ratio of bird versus other for an alligator of small size relative to an alligator of large size? Interpret this odds ratio

The odds ratio here can be given by:

$$OR = \frac{\frac{P(Y=bird|alligator=small)}{P(Y=bird|alligator=large)}}{\frac{P(Y=other|alligator=small)}{P(Y=other|alligator=large)}} = \frac{e^{\beta_{5,bird}}}{e^{\beta_{5,other}}} = e^{\beta_{5,bird}-\beta_{5,other}} = e^{-0.730-0.291} = e^{-1.021} = 0.360$$

This signifies that the odds for an alligator's food choice to be bird relative to it being other is expected to change by a factor of 0.360 when the alligator is noted to be of size small instead of large.

Q4. Test the global significance of the variable lake.

```
fit2 <- update(fit, ~ . - lake)

## # weights:  20 (12 variable)
## initial  value 352.466903
## iter  10 value 294.669539
## final  value 294.091727
## converged

anova(fit, fit2, test="Chisq")

## Likelihood ratio tests of Multinomial Models
##
## Response: food
##                Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
## 1        sex + size       308   588.1835
## 2 lake + sex + size       296   537.8655 1 vs 2    12 50.31797 1.228388e-06
```

Here we use the update() function to remove the 'lake' variable from the model, and we then produce an anova table to compare the fit of the two models. We are testing the following:

$$H_0 : \beta_{1,m} = \beta_{2,m} = \beta_{3,m} = 0; \forall m \in \{invertebrate, reptile, bird, other\}$$

versus

$$H_0 : \beta_{jm} \neq 0; \text{ for at least one combination of } j = 1,2,3 \text{ and } m \in \{invertebrate, reptile, bird, other\}$$

In words, we are testing the null hypothesis that the beta values for each level of the categorical variable 'lake' is equal to zero for each of the Y-values, versus the alternative that at least one of the beta values is not equal to zero.

The testing distribution that is used is the chi-squared distribution and in this case we are testing with 12 degrees of freedom. The test statistic has a value of 50.318, which leads to a p-value of 1.228e-06. This signifies strong evidence against the null hypothesis, and therefore we can reject the null hypothesis that the lake variable is not significant in this model set-up.

Q5. Test the fit of the model

Here we fit a model with just the intercept and compare the deviance of the full model with the deviance of this intercept-only model. The hypotheses we are testing is:

$$H_0 : \beta_{jm} = 0; \forall j = 1, ..., 6; m \in \{invertebrate, reptile, bird, other\}$$

and

$$H_0 : \beta_{jm} \neq 0; \text{for at least one} j = 1, ..., 6; m \in \{invertebrate, reptile, bird, other\}$$

In words, we are testing the null hypothesis that the beta values for variable in the dataset is equal to zero for each of the Y-values, versus the alternative that at least one of the beta values is not equal to zero.

```
fit.null <- multinom(food ~ 1, data, weights = count)

## # weights:  10 (4 variable)
## initial  value 352.466903
## final  value 302.181462
## converged

anova(fit.null, fit, test="Chisq")

## Likelihood ratio tests of Multinomial Models
##
## Response: food
##               Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
## 1                 1       316   604.3629
## 2 lake + sex + size       296   537.8655 1 vs 2    20 66.49745 6.723387e-07
```

Again, the testing distribution used is the chi-squared distribution and in this case we are testing with 20 degrees of freedom. From this, we can see that the p-value for this test is 6.723e-07. Assuming we are testing at the 5% significance level, then we can again reject the null hypothesis, in this case the null hypothesis that all beta-values are equal to zero.