UNIVERSITY *of* WEST FLORIDA

Nominal Logistic Regression

Statistics for Data Science II

Introduction

Suppose we now have an outcome with more than two possible nominal outcomes.

e.g., type of account at bank: mortgage, credit card, personal

When we have a response variable with c categories, we can create multicategory logistic models simultaneously.

We will choose a reference category and create $c - 1$ models.

Each model will compare outcome j to outcome c (reference group).

Model

The baseline-category logit model (or the multinomial logit model):

$$\ln\left(\frac{\pi_j}{\pi_c}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

where $j = 1, \ldots, c-1$.

Again, each model is comparing outcome j to outcome c.

Model

Example:

Let us examine foods that alligators in the wild choose to eat. For 59 alligators sampled in Lake George, Florida, the alligator data shows the primary food type (in volume) found in the alligator's stomach. Primary food type has three categories: Fish, Invertebrate, and Other. The invertebrates were primarily apple snails, aquatic insects, and crayfish. The "other" category included amphibian, mammal, plant material, stones or other debris, and reptiles. Let's model food choice as a function of alligator length.

Model

Example:

```
head(data)
```

```
## # A tibble: 6 x 2
##   length food
##    <dbl> <chr>
## 1   1.24 I
## 2   1.3  I
## 3   1.3  I
## 4   1.32 F
## 5   1.32 F
## 6   1.4  F
```

```
data$food <- factor(data$food, levels = c("O", "I", "F"))
```

Model

Example:

```
m1 <- multinom(food ~ length, data = data)
```

```
## # weights:  9 (4 variable)
## initial  value 64.818125
## iter  10 value 49.170785
## final  value 49.170622
## converged
```

Model

Example:

```
coefficients(m1)
```

```
##   (Intercept)     length
## I    5.697543 -2.4654695
## F    1.617952 -0.1101836
```

This results in two models:

$$\ln\left(\frac{\pi_\text{I}}{\pi_\text{O}}\right) = 5.70 - 2.47\text{length}$$
$$\ln\left(\frac{\pi_\text{F}}{\pi_\text{O}}\right) = 1.62 - 0.11\text{length}$$

Interpretations

Interpretation for continuous predictors:

For a 1 [predictor's unit] increase in [predictor name], the odds in favor of [response category j] over [response reference category] are multiplied by $e^{\hat{\beta}_i}$.

For a 1 [predictor's unit] increase in [predictor name], the odds of [response category j] are [increased or decreased] by [100($e^{\hat{\beta}_i}$-1)% or 100(1-$e^{\hat{\beta}_i}$)%] as compared to the [response reference category].

Interpretations

Interpretations for categorical predictors:

> As compared to [predictor reference category], the odds of [predictor category of interest] in favor of [response category *j*] over [response reference category] are multiplied by $e^{\hat{\beta}_i}$.
>
> As compared to [predictor reference category], the odds of [predictor category of interest] in favor of [response category *j*] over [response reference category] are [increased or decreased] by [$100(e^{\hat{\beta}_i}-1)\%$ or $100(1-e^{\hat{\beta}_i})\%$].

Interpretations

Example:

Let's convert the $\hat{\beta}_i$ to odds ratios and provide brief interpretations.

```
round(exp(coefficients(m1)), 2)
```

```
##   (Intercept) length
## I      298.13   0.08
## F        5.04   0.90
```

For a 1 meter increase in alligator length, the odds of choosing invertebrates over other food are multiplied by 0.08, or decreased by 92%.

For a 1 meter increase in alligator length, the odds of choosing fish over other food are multiplied by 0.90, or decreased by 10%.

Inference

We will first test for overall (global) significance, as we saw in previous lectures, using the `anova()` function.

Example:

```
full <- multinom(food ~ length, data = data)
reduced <- multinom(food ~ 1, data = data)
```

```
anova(reduced, full)
```

```
##    Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
## 1      1       116  115.14186           NA       NA           NA
## 2 length       114   98.34124 1 vs 2     2 16.80061 0.0002247985
```

Yes, length of alligator is a significant predictor of food choice ($p < 0.001$).

Inference

Like in binary logistic regression, we can construct Wald Z statistics that will allow us to test for significance within each model constructed.

Example:

```
coeftest(m1)
```

```
##
## z test of coefficients:
##
##                Estimate Std. Error z value Pr(>|z|)
## I:(Intercept)  5.69754    1.79382  3.1762 0.001492 **
## I:length      -2.46547    0.89965 -2.7405 0.006135 **
## F:(Intercept)  1.61795    1.30729  1.2376 0.215851
## F:length      -0.11018    0.51708 -0.2131 0.831259
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Length is a significant predictor of choosing invertebrates over other food choices ($p = 0.001$) but not when choosing fish over other food choices ($p = 0.831$).

Inference

Like in binary logistic regression, we can construct confidence intervals using $\hat{\beta}_i$, $z_{1-\alpha/2}$, and $\text{SE}_{\hat{\beta}_i}$. We will run the model results through the `confint()` function.

Example:

```
round(exp(confint(m1)),2)
```

```
## , , I
##
##               2.5 %   97.5 %
## (Intercept)  8.86 10030.31
## length       0.01     0.50
##
## , , F
##
##               2.5 % 97.5 %
## (Intercept)  0.39  65.38
## length       0.33   2.47
```

Inference

Example:

```
round(exp(confint(m1)),2)[,,1]
```

```
##                2.5 %   97.5 %
## (Intercept)   8.86 10030.31
## length        0.01     0.50
```

The 95% CI for the OR for length when choosing invertebrates over other food choices is (0.01, 0.50).

Inference

Example:

```
round(exp(confint(m1)),2)[,,2]
```

```
##              2.5 % 97.5 %
## (Intercept)   0.39  65.38
## length        0.33   2.47
```

The 95% CI for the OR for length when choosing fish over other food choices is (0.33, 2.47).

Predictions

We can construct predicted probabilities for the non-baseline categories as follows:

$$\pi_i = \frac{\exp\{\beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki}\}}{1 + \sum_h \exp\{\beta_{0h} + \beta_{1h} X_{1i} + \ldots + \beta_{kh} X_{ki}\}}$$

Then, for the baseline category,

$$\pi_i = \frac{1}{1 + \sum_h \exp\{\beta_{0h} + \beta_{1h} X_{1i} + \ldots + \beta_{kh} X_{ki}\}}$$

Predictions

Example:

```
c1 <- coefficients(m1)
data <- data %>% mutate(pred_I = exp(c1[1] + c1[3]*data$length)/
                               (1+exp(c1[1] + c1[3]*data$length)
                               +exp(c1[2] + c1[4]*data$length)),
                        pred_F = exp(c1[2] + c1[4]*data$length)/
                               (1+exp(c1[1] + c1[3]*data$length)
                               +exp(c1[2] + c1[4]*data$length)),
                        pred_O = 1/(1+exp(c1[1] + c1[3]*data$length)
                                 +exp(c1[2] + c1[4]*data$length)))
```

Predictions

Example:

```
head(data)
```

```
## # A tibble: 6 x 5
##   length food  pred_I pred_F pred_O
##    <dbl> <fct>  <dbl>  <dbl>  <dbl>
## 1   1.24 I      0.722  0.227 0.0515
## 2   1.3  I      0.692  0.250 0.0573
## 3   1.3  I      0.692  0.250 0.0573
## 4   1.32 F      0.682  0.258 0.0593
## 5   1.32 F      0.682  0.258 0.0593
## 6   1.4  F      0.640  0.293 0.0677
```

Visualization

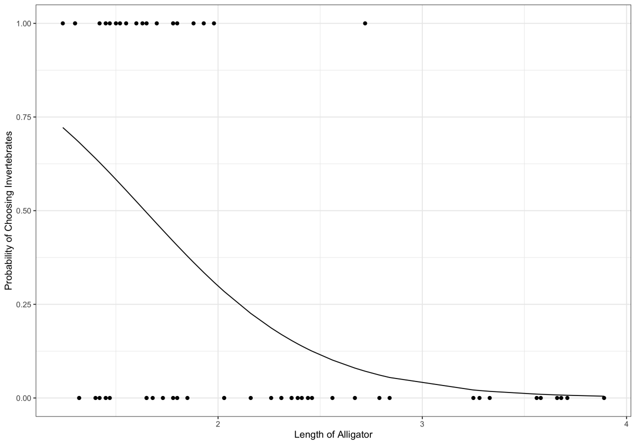
Example:

Create visualizations for the probability of food choice vs. the length of the alligator.

```
data <- dummy_cols(data, select_columns = "food")
p1 <- data %>% ggplot(aes(x = length, y = food_I)) +
  geom_point() +
  geom_line(aes(y = pred_I)) +
  ylab("Probability of Choosing Invertebrates") +
  xlab("Length of Alligator") +
  theme_bw()
```

Visualization

Example:



Visualization

Example:

```
p2 <- data %>% ggplot(aes(x = length, y = food_F)) +
  geom_point() +
  geom_line(aes(y = pred_F)) +
  ylab("Probability of Choosing Fish") +
  xlab("Length of Alligator") +
  theme_bw()
```

Visualization

Example: