CameraGLM

2023-08-30

# \* \* BACKGROUND \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \*

This script sets up and runs a series of GLMs to explain why we misclassify the final fate of certain nests in the field (e.g., which variables determine whether we can classify a nest at all, and whether we correctly classify it).

**My notes/interpretations of the model output are in boxes to differentiate them from the text created by the R script (which just kind of walks you through the script).**

Predictor variables:

1. nest\_age (age of nest at the time that fate was assigned)
2. fate\_date (day of the season on which fate was assigned)
3. obs\_int (number of days between penultimate and final observation of nest)
4. species (CONI – Common Nighthawk or LETE – Least Tern)
5. cam\_fate (the true fate as discovered from camera footage)

Response variables:

1. HF\_mis (0/1 was nest misclassified with respect to hatch vs fail?)
2. misclass (0/1 was nest misclassified, including all fates?)
3. is\_u (0/1 was nest unclassifiable from the field evidence?)

Prepare data:

* Examine and remove NAs
* Make sure categorical variables are coded as factors in R
* Create datasets with and without 2021 data so we can compare

Examine the data:

* View frequency tables for predictor and response variables
* View histograms of response variable for each proposed analysis
* Check for correlation between continuous predictor variables

Run the models:

* Create functions to simplify running so many models
* For each proposed analysis/question:
  + examine output of univariate model for each predictor
  + create full, null, and nested models with explanations
  + generate odds ratios and AIC table

# \* \* PREPARE THE DATA \* \* \* \* \* \* \* \* \* \* \* \* \*

**NOTES:**

I didn’t notice anything too weird in my examination/visualization of the data. Clearly, without the 2021 data, some of the categories for the cam\_fate predictor become too small to run a logistic regression. I combined some categories, but may have to combine more going forward. I address this in the next note.

As far as what should be included in the chapter, I am planning to mention the raw numbers of hatched and failed nests, with and without the cameras. I will also mention the breakdown of the failure categories, and what predators were seen. A lot of this is probably best represented in a table, but I think the frequency tables I created below are too busy. I’m not sure if I should remove the percentages, split the table into three tables (one per response variable), or just add better spacing.

## Read in the data.

## Data wrangling

First, we will select only the nests with cameras, because the response variable is fate misclassification (difference between camera and field fate).

We then reduce back to one row per nest by condensing the observation history to one cell.

##   
## CONI LETE   
## 47 107

##   
## RUTE RUTW   
## 43 111

##   
## 2019 2020 2021   
## 46 54 54

##   
## A Ca D F H Hu S U   
## 12 12 21 12 62 1 6 1

## Add variables representing misclassification.

* Combine some levels of the cam\_fate variable and make sure it’s coded as a factor with correct number of levels.
* Create functions to label nests as misclassified or not.
* Add variable for nest misclassification based on hatch or fail, not multiple failure causes.
  + Coded as 0 (fail) 1 (hatch) and 7 (unknown)
  + Nests that were uncertain (e.g. “most likely hatched”) were classified as though there was no uncertainty for this analysis (that nest would just be “hatched”)

## without cameras:  
## 0 1 7   
## 56 61 37

##

## with cameras:  
## 0 1 7   
## 64 62 28

* Add variable for nest misclassification including all fate categories. For this variable, the uncertain fates were grouped into two categories (“U-H” and “U-F”) for probably hatched and probably failed, respectively.
* Add ariable for how the nest was misclassified (correct, newly assigned from unknown, incorrect)
* Visually inspect new fate categories for discrepancies.
* Create new variable (1/0), assigned fate in field or not

## Examine data; find and remove missing values (NA).

* Ungroup dataframe and make categorical variables into factors; make sure other data classes are correct
* Examine and remove missing data (NA) for relevant variables

Including 2021: 36 nests with NA for at least one analysis variable (24%) Excluding 2021: 20 nests (20%)

## All the missing values (with 2021):

## nest site species status cov\_5m fate\_date   
## 0 0 0 0 49 41   
## nest\_age obs\_int i j k final\_fate   
## 26 0 3 9 7 0   
## fate cam\_fate cfate year fdate hatchfail   
## 0 27 27 0 7 0   
## c\_hatchfail HF\_mis misclass how\_mis is\_u c.fate   
## 0 0 28 28 0 0   
## numNA numNA1920   
## 0 0

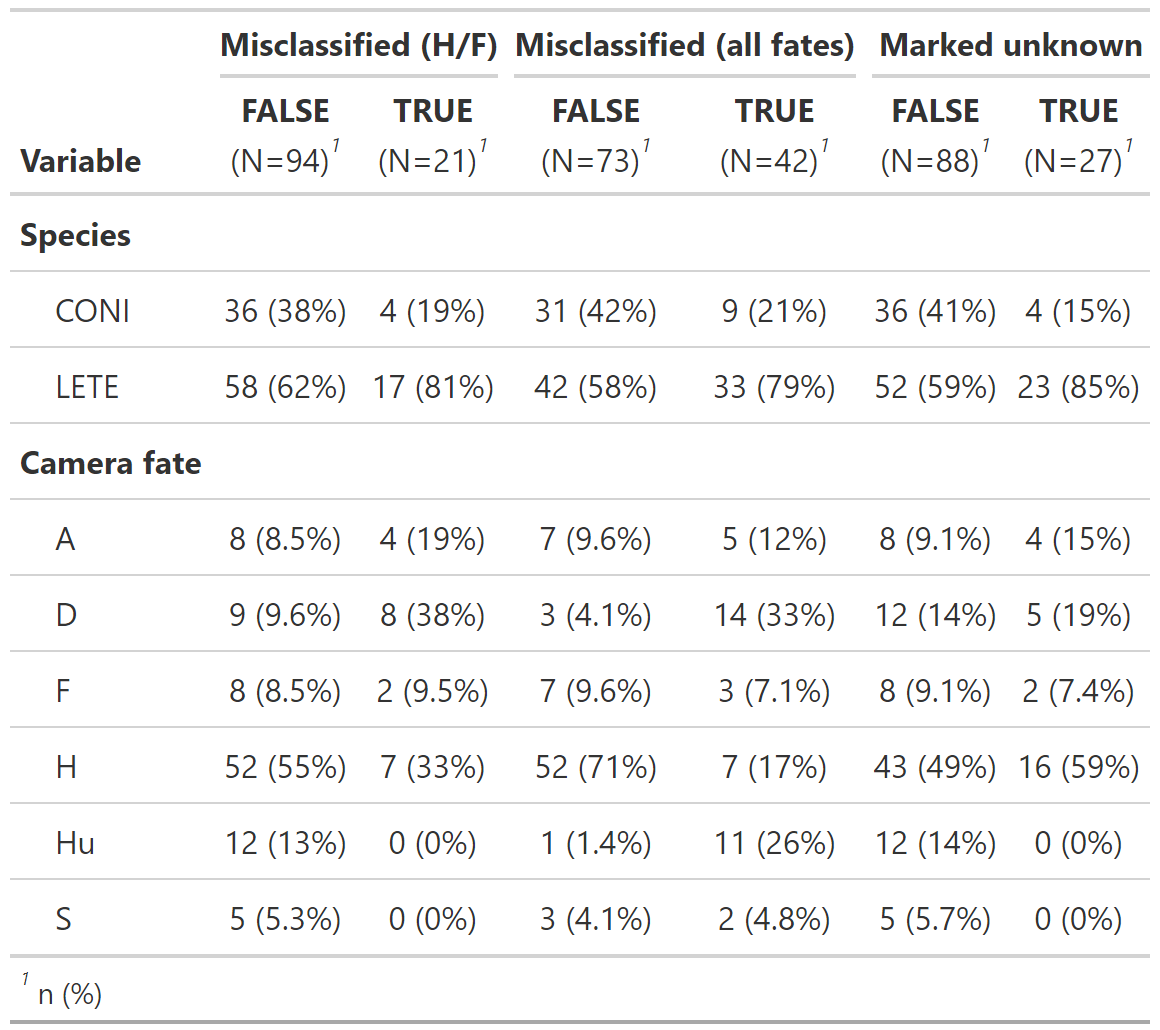
##   
##   
## Without 2021:

## nest site species status cov\_5m fate\_date   
## 0 0 0 0 37 16   
## nest\_age obs\_int i j k final\_fate   
## 10 0 1 2 2 0   
## fate cam\_fate cfate year fdate hatchfail   
## 0 16 16 0 2 0   
## c\_hatchfail HF\_mis misclass how\_mis is\_u c.fate   
## 0 0 16 16 0 0   
## numNA numNA1920   
## 0 0

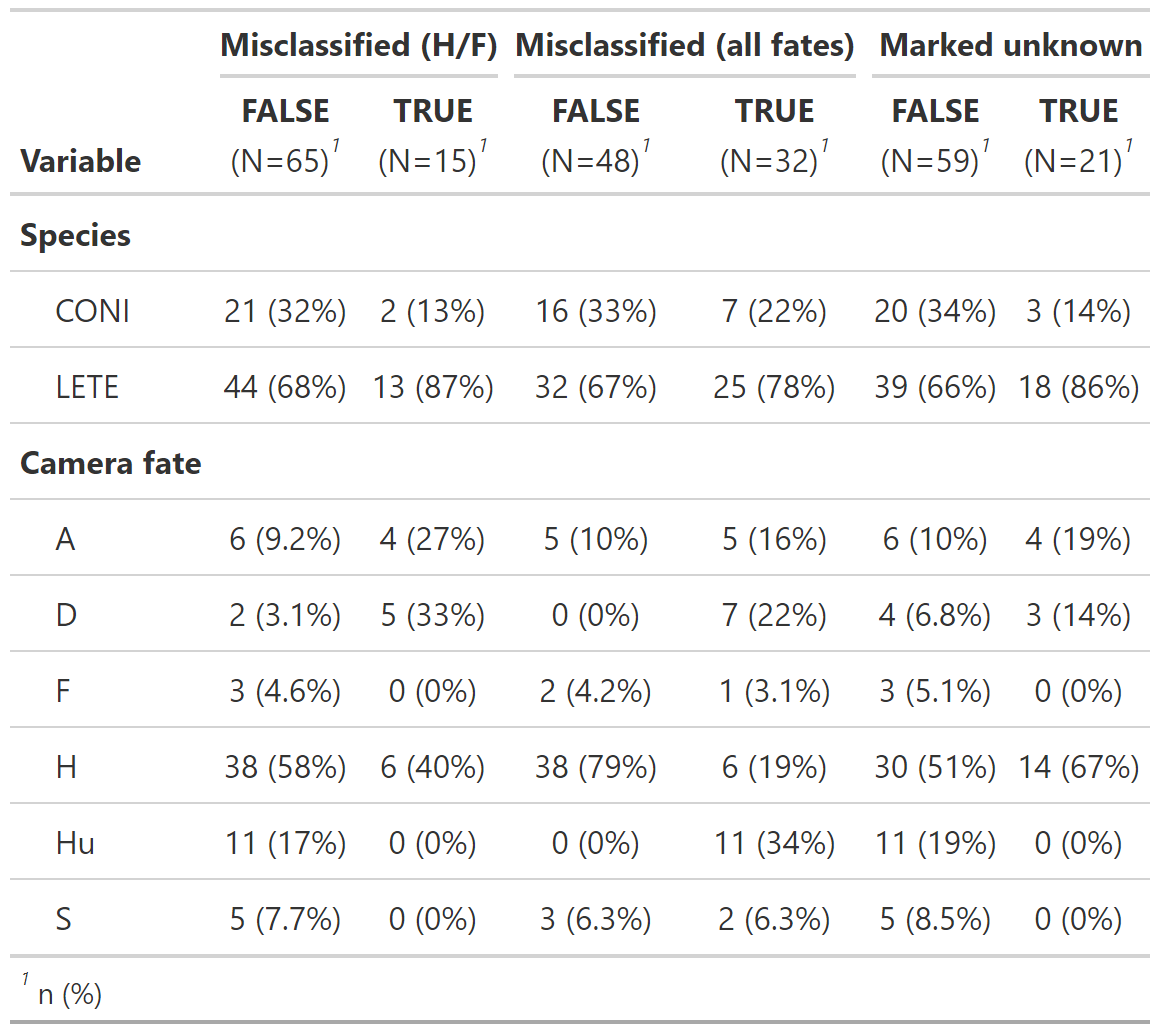
* Create new dataset w/o 2021

## Create frequency tables for vars in nest data

## Frequency table including 2021



## Excluding 2021



# \* \* PLOTS \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \*

## HISTOGRAMS

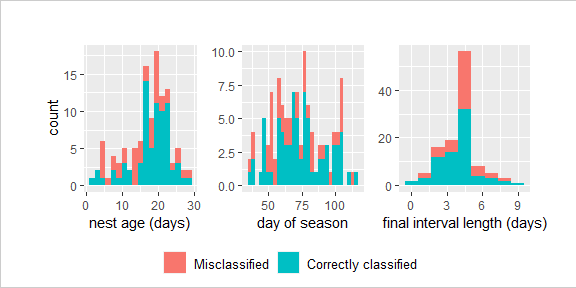
Plot the distributions of each continuous predictor with and without 2021 data. Make sure the bin sizes make sense.

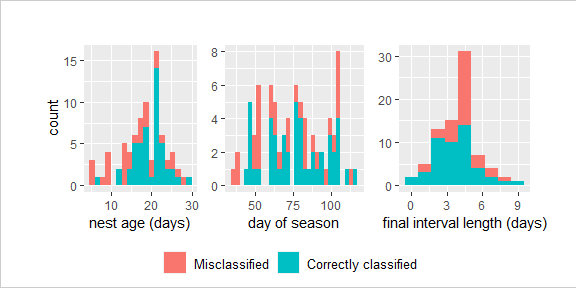
### First, create a function that makes the plots.

* Can change response variable / data for the plot without having to copy-paste and edit

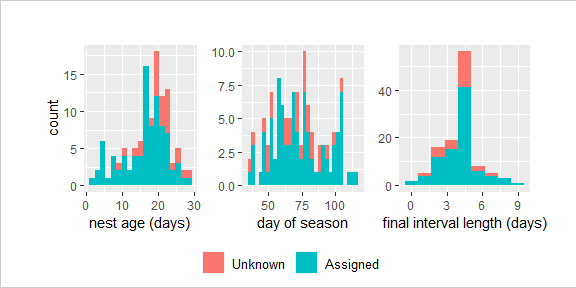
### Nest fate misclassification

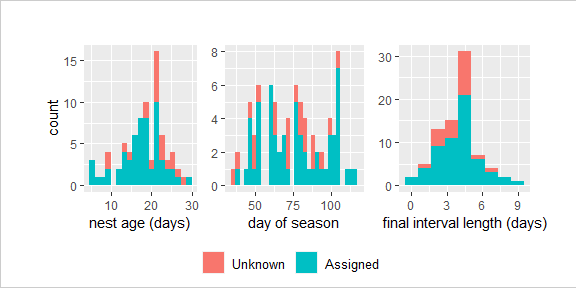
* Ends up with NAs for nests w/ unknown fate in the field





### Assigned vs unknown





## CORRELATION PLOTS

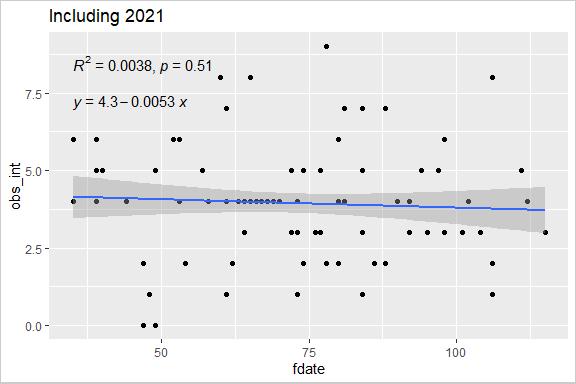
* Check for correlation amongst continuous predictors

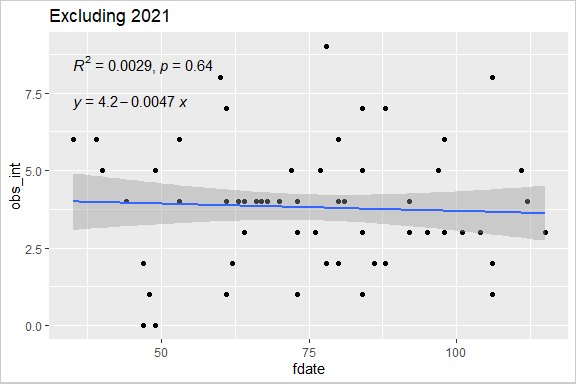
**NOTES:**

I am not sure if the correlation in the third plot (nest age vs. fate date, including 2021) requires further investigation, although my instinct was that it did not. Is there a cutoff number for R-squared above which you would say that the correlation should be looked into?

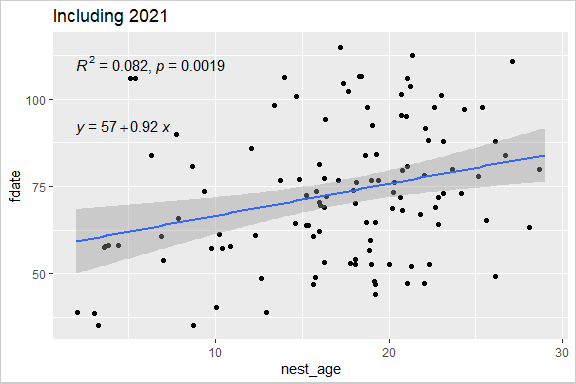
I will make these into correlation matrices so they don’t take up so much room.

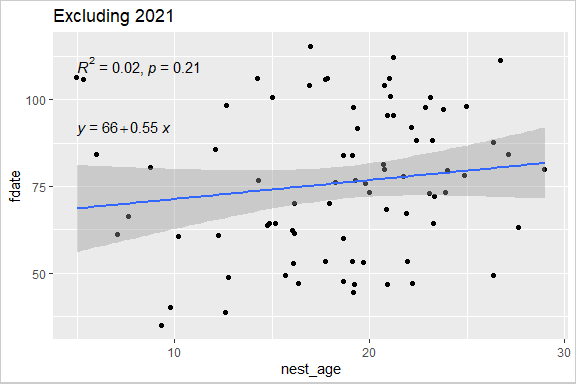
### Fate date vs obs int



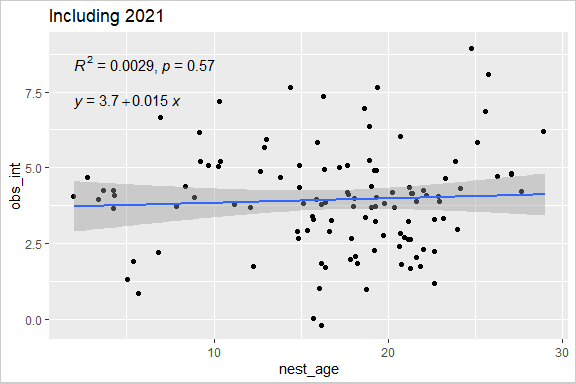


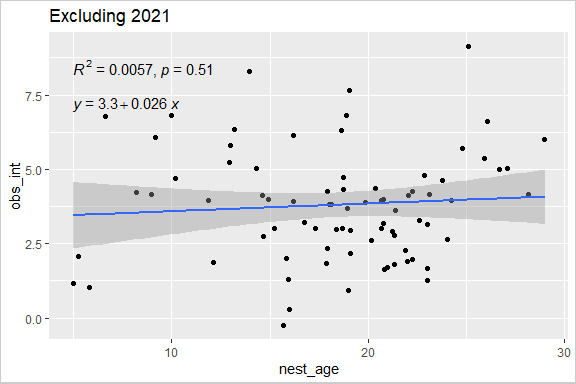
### Nest age vs fate date





### Nest age vs obs int





# \* \* MODELS \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \* \*

## FUNCTIONS

Functions to do the following:

* Make a univariate model from each predictor variable.  
  Print the summaries of all univariate models
* Compare coefficients of any 2 different models
* Generate odds ratios
* Calculate AIC and BIC
* Create combo AIC/BIC tables
* Save tables to file

## PREPARE THE DATA

Center the (numerical) predictor variables (don’t scale, just center)

## QUESTION 1: How do the predictor vars affect classification accuracy?

Response: HF\_mis

## Including 2021 data

**NOTES:**

For this first model (question 1, including 2021 data), I have included all the R output, but for subsequent models I have included only a table summary of the univariate models and an AIC table for all models evaluated. I can easily have R print all the output for all models if something warrants further attention.

**Main effects model:** The best model appears to be the full model, as demonstrated in the step-by-step procedure and the AIC values. This means that all of the predictor variables have some effect on the variability of the response variable, even if it’s indirectly through another variable.

There are other goodness-of-fit statistics I could also apply, if relevant.

**Adding interactions:** There are two interaction models that are almost as well ranked as the main effects model, and none of these three models is smaller than any of the others. I don’t feel comfortable saying one is the best fit unless parsimony also applies to interactions, in which case I would say the simplest model (no interactions) is still the main effects model.

### Initialize - Choose the data and response var, create list to store models.

### Step 1: Univariate models

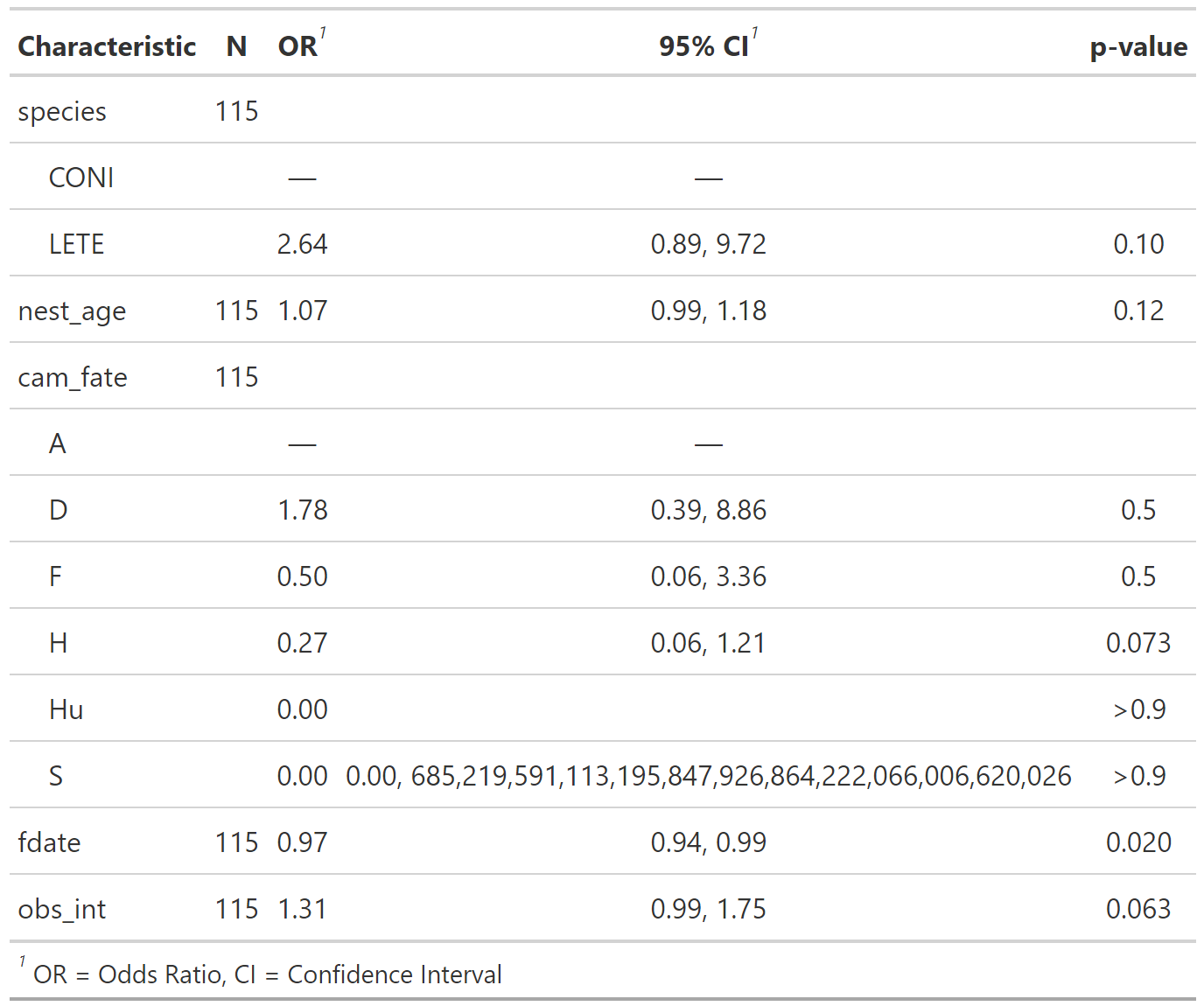
**NOTES**

Analyze the relationship between each predictor variable and the response variable when the other variables are not accounted for. The output can be used in different ways depending on how you choose to do variable selection.

The output table below shows that some categories in the camera fate predictor have ridiculous SE/confidence intervals. I think they have too few observations. I already merged cattle into human impacts, and I could merge more categories, but the question is: Is it better to have unreliable parameter estimates, or split the variable into fewer categories that are less informative, or just exclude the variable altogether?

## HF\_mis ~ species HF\_mis ~ nest\_age HF\_mis ~ cam\_fate HF\_mis ~ fdate HF\_mis ~ obs\_int

## [[1]]  
## [[1]][[1]]  
##   
## Call:  
## glm(formula = f, family = binomial, data = modData)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.1972 0.5270 -4.169 3.06e-05 \*\*\*  
## speciesLETE 0.9700 0.5948 1.631 0.103   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 109.33 on 114 degrees of freedom  
## Residual deviance: 106.29 on 113 degrees of freedom  
## AIC: 110.29  
##   
## Number of Fisher Scoring iterations: 4  
##   
##   
## [[1]][[2]]  
##   
## Call:  
## glm(formula = f, family = binomial, data = modData)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.55560 0.25402 -6.124 9.12e-10 \*\*\*  
## nest\_age 0.07042 0.04477 1.573 0.116   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 109.33 on 114 degrees of freedom  
## Residual deviance: 106.59 on 113 degrees of freedom  
## AIC: 110.59  
##   
## Number of Fisher Scoring iterations: 4  
##   
##   
## [[1]][[3]]  
##   
## Call:  
## glm(formula = f, family = binomial, data = modData)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.6931 0.6124 -1.132 0.2577   
## cam\_fateD 0.5754 0.7817 0.736 0.4617   
## cam\_fateF -0.6931 1.0000 -0.693 0.4882   
## cam\_fateH -1.3122 0.7329 -1.790 0.0734 .  
## cam\_fateHu -17.8729 1882.9237 -0.009 0.9924   
## cam\_fateS -17.8729 2917.0127 -0.006 0.9951   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 109.33 on 114 degrees of freedom  
## Residual deviance: 91.77 on 109 degrees of freedom  
## AIC: 103.77  
##   
## Number of Fisher Scoring iterations: 17  
##   
##   
## [[1]][[4]]  
##   
## Call:  
## glm(formula = f, family = binomial, data = modData)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.62402 0.26855 -6.047 1.47e-09 \*\*\*  
## fdate -0.03278 0.01405 -2.333 0.0197 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 109.33 on 114 degrees of freedom  
## Residual deviance: 103.20 on 113 degrees of freedom  
## AIC: 107.2  
##   
## Number of Fisher Scoring iterations: 4  
##   
##   
## [[1]][[5]]  
##   
## Call:  
## glm(formula = f, family = binomial, data = modData)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.5617 0.2540 -6.149 7.79e-10 \*\*\*  
## obs\_int 0.2672 0.1436 1.861 0.0628 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 109.33 on 114 degrees of freedom  
## Residual deviance: 105.78 on 113 degrees of freedom  
## AIC: 109.78  
##   
## Number of Fisher Scoring iterations: 4



### Step 2: Multivariate models

**NOTES**

Method: First, I created the null and full models.

To select variables, I opted to start with the variables with significant p-values with an arbitrary cutoff of 0.25 (cite). I then compared to the full model via a likelihood ratio test, and examined the percent change in coefficients between the full model and the reduced model. If the percent change is large or the LRT p-value is low, then some variables that were removed were providing explanatory power.

I then added those variables back in one by one and again compared using LRT and percent coefficient change. If the p-value for the Likelihood Ratio Test is large (or the percent coefficient change is small), this means both models are equally good at explaining the variation (e.g. the variable does not improve the model). If any variables did not contribute significantly to the explanatory power of the model, I removed them to create my proposed main effects model.

To look for the presence of interactions, I first decided what interactions would make sense biologically, and then added those to the main effects model one at a time. I evaluated the difference compared to the main effects model as explained above.

Finally, I compared the remaining models.

#### Null and full models

## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ 1  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 105 75.659   
## 2 114 109.325 -9 -33.666 0.0001022 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
## (Intercept)   
## -0.401368  
## ---------------------------------------------------------------

#### Other models

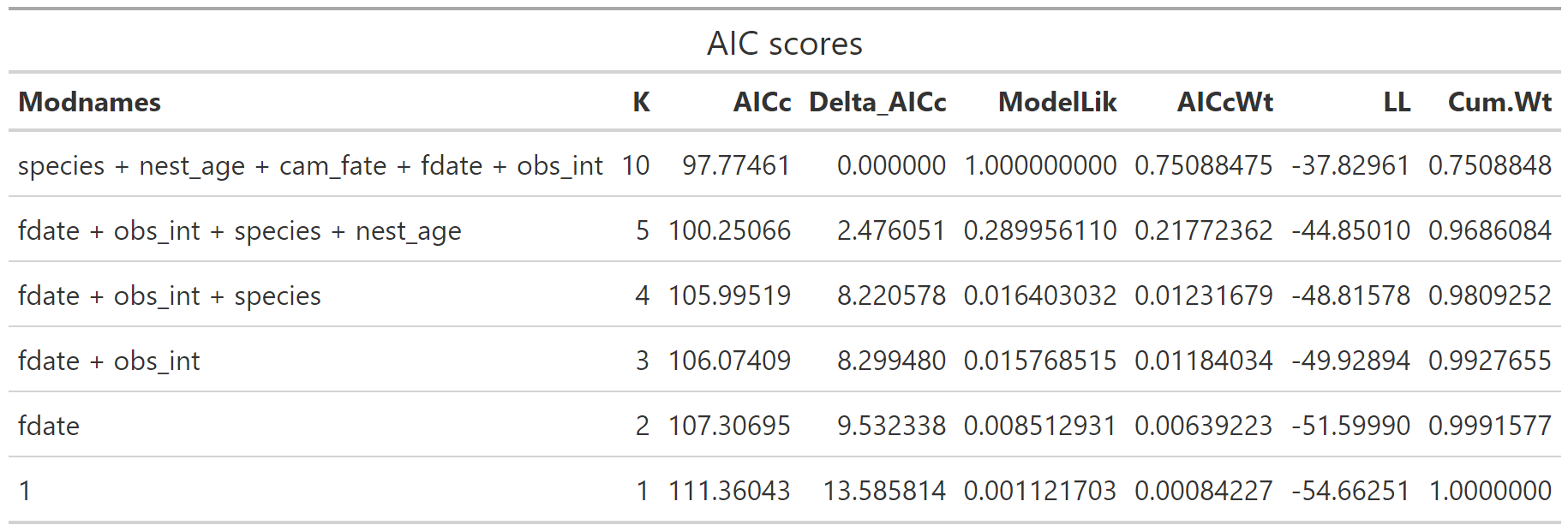
## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ fdate  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 105 75.659   
## 2 113 103.200 -8 -27.541 0.00057 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
## (Intercept) fdate   
## -0.3513417 -0.176406  
## ---------------------------------------------------------------

## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ fdate + obs\_int  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 105 75.659   
## 2 112 99.858 -7 -24.199 0.001051 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
## (Intercept) fdate obs\_int   
## -0.3234969 -0.1660251 0.3961847  
## ---------------------------------------------------------------

## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ fdate + obs\_int + species  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 105 75.659   
## 2 111 97.632 -6 -21.972 0.001225 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
## (Intercept) speciesLETE fdate obs\_int   
## -0.08391833 1.029512 -7.456802 3.52492  
## ---------------------------------------------------------------

## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ fdate + obs\_int + species + nest\_age  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 105 75.659   
## 2 110 89.700 -5 -14.041 0.01535 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
## (Intercept) speciesLETE nest\_age fdate obs\_int   
## -0.05942387 1.048524 0.2446208 -28.44609 0.318324  
## ---------------------------------------------------------------

#### Analysis, part 1: select main effects model



#### Interaction models

Specify the main effects model, and compare interaction models to it. In this case, it’s the same as the full model.

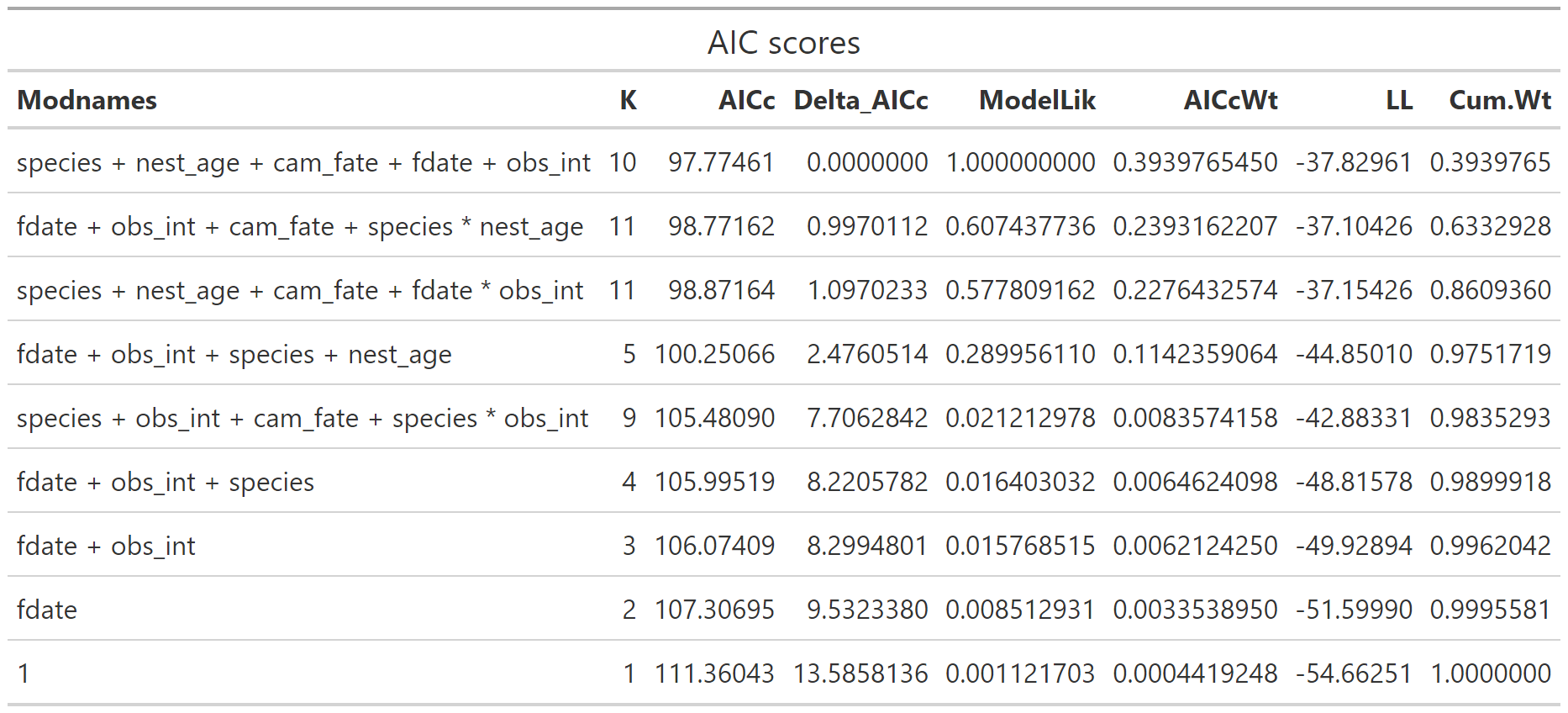
## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ fdate + obs\_int + cam\_fate + species \* nest\_age  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 105 75.659   
## 2 104 74.209 1 1.4507 0.2284  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
##   
##   
## ---------------------------------------------------------------

## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ species + obs\_int + cam\_fate + fdate \* nest\_age  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 105 75.659   
## 2 104 75.233 1 0.42598 0.514  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
##   
##   
## ---------------------------------------------------------------

## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ species + nest\_age + cam\_fate + fdate \* obs\_int  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 105 75.659   
## 2 104 74.309 1 1.3507 0.2452  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
##   
##   
## ---------------------------------------------------------------

## Compare to full or main effects model:  
##   
## Analysis of Deviance Table  
##   
## Model 1: HF\_mis ~ species + nest\_age + cam\_fate + fdate + obs\_int  
## Model 2: HF\_mis ~ species + obs\_int + cam\_fate + species \* obs\_int  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 105 75.659   
## 2 106 85.767 -1 -10.107 0.001477 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
##   
## Percent difference in model coefficients (vs full or main effects model):  
##   
## (Intercept) speciesLETE cam\_fateD cam\_fateF cam\_fateH cam\_fateHu cam\_fateS obs\_int   
## -0.4127639 0.3016698 1.024008 0.2080796 -0.08902561 -0.9382872 -0.07898081 93.63552 -1.196229  
## ---------------------------------------------------------------

#### Analysis



## Question 1, excluding 2021 data

**NOTES:**

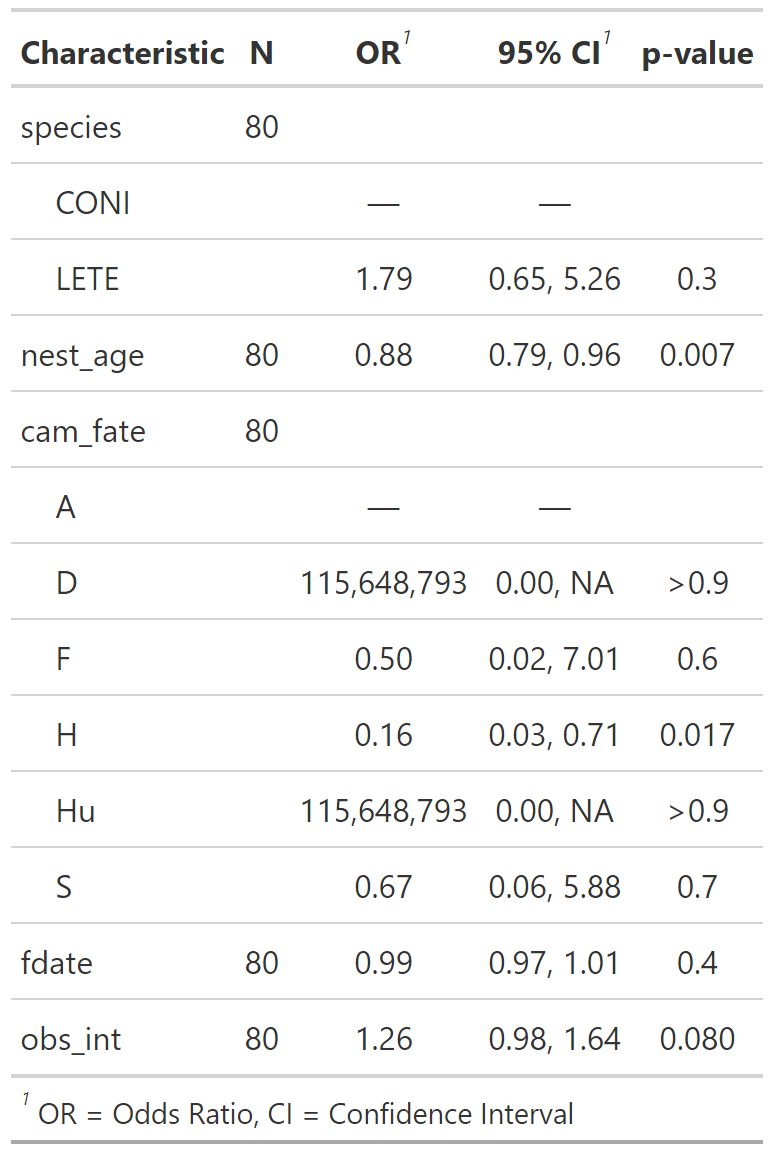
The best fit model for this data set is smaller than when we include the 2021 data. I’m not sure if this means that (with respect to this response variable) the 2021 data is anomalous, or if it’s something we can explain.

**Model selection:** The best three models are all very close together, so for the sake of parsimony I would pick the simplest of the three.

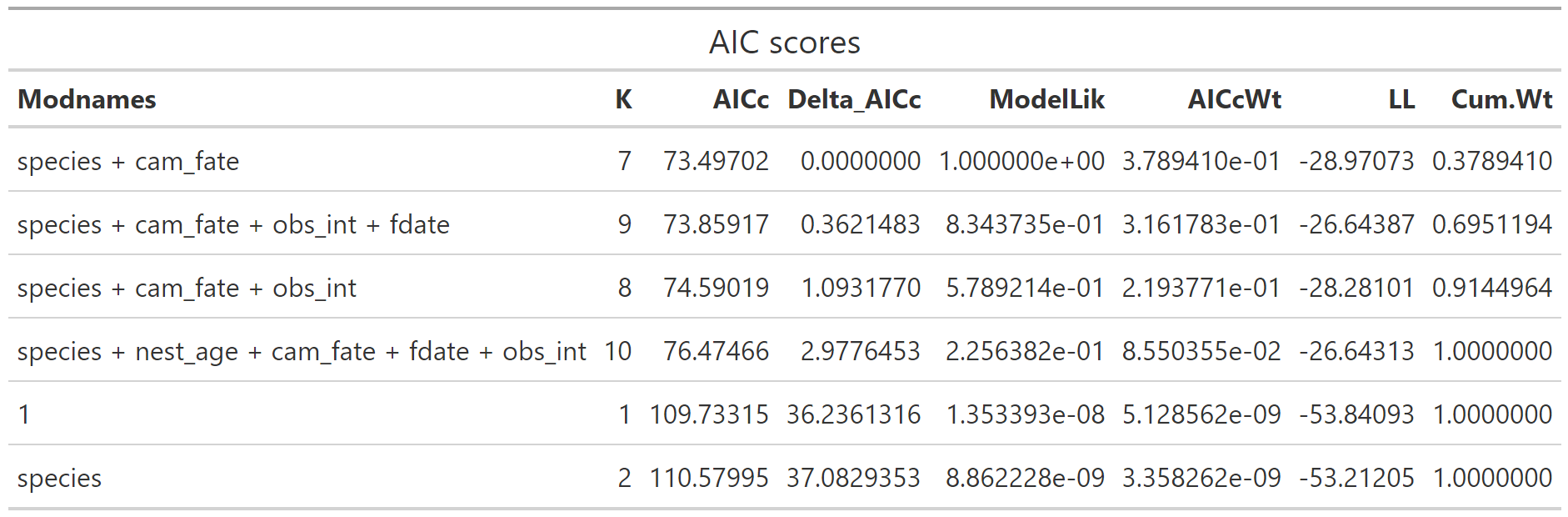
I did not test any interactions for this model.

#### Univariate models

## misclass ~ species misclass ~ nest\_age misclass ~ cam\_fate misclass ~ fdate misclass ~ obs\_int



#### Multivariate models



## QUESTION 2: If nest has failed, what affects classification accuracy?

* Remove hatched nests
* Response: misclass

## Including 2021 data

**NOTES:**

For this analysis, because the first and second models (ranked by AIC) are so close, I would choose the model with only camera fate as the best model.

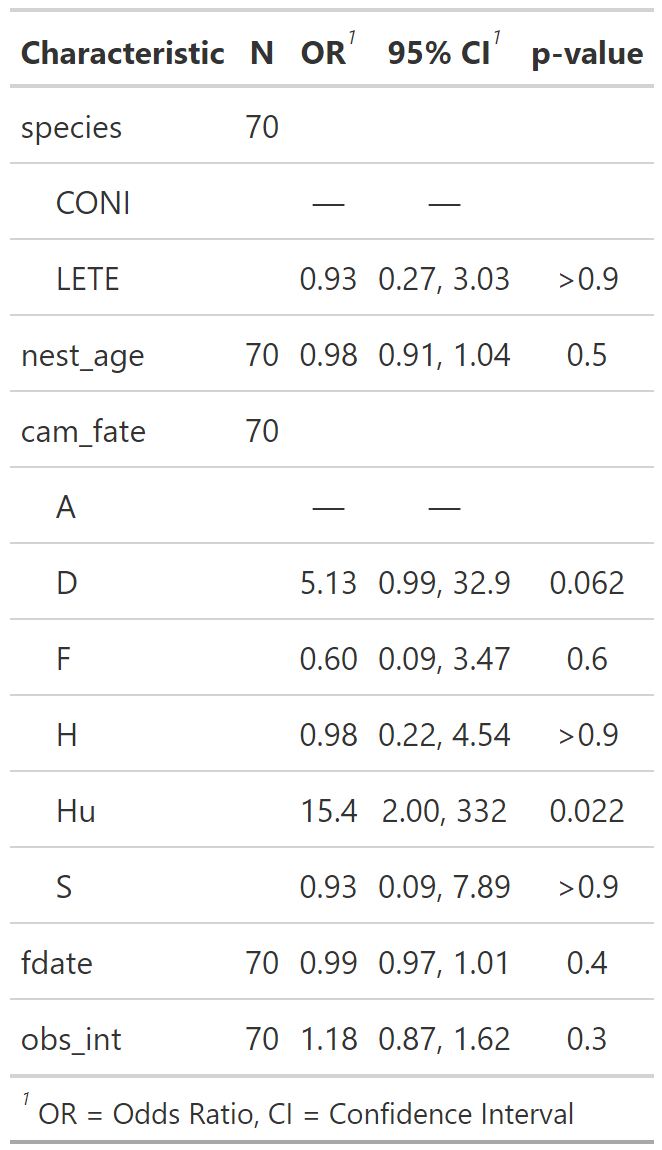
In the step-by-step procedure, I reduced the model to just cam\_fate, and when I added back the next most significant variable (obs\_int) the Chi-sq p-value (compared to full model) almost doubled to about 0.5. So I probably didn’t need to add back the other variables (models 3 and 4 on the AIC table)? Or is it possible they had some indirect effect, so better to include them?

### Initialize

Choose the data and response var, create list to store models

### Univariate models

## misclass ~ species misclass ~ nest\_age misclass ~ cam\_fate misclass ~ fdate misclass ~ obs\_int

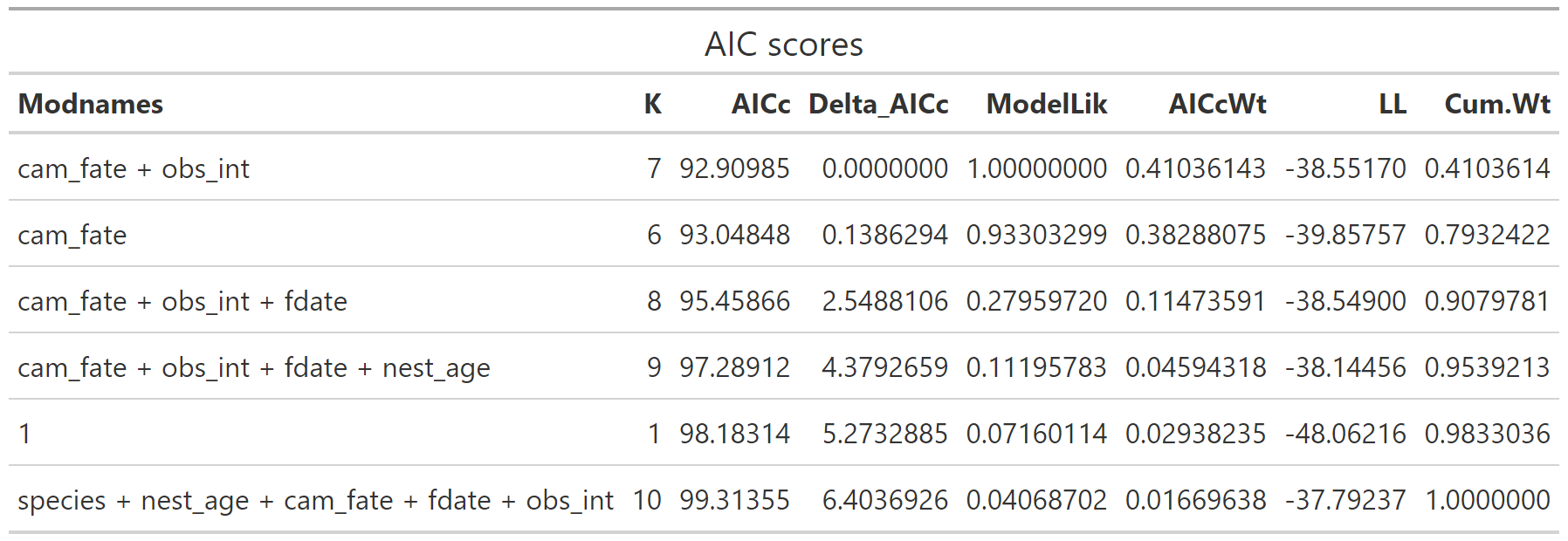


### Multivariate models

#### Null and full models

#### Other models

#### Analysis



## Excluding 2021 data

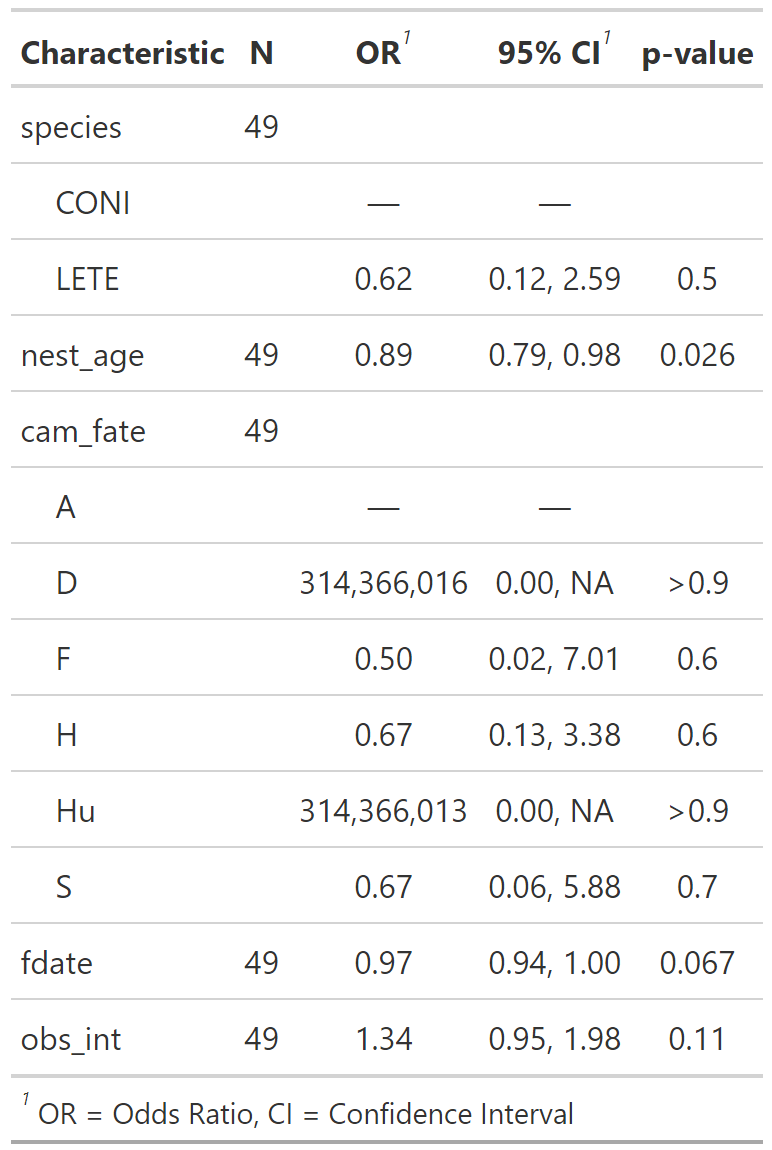
**NOTES:**

The best model without the 2021 data is again different from the model including the 2021 data, and so are the other top-ranked models. In part, this is clearly because of my choice of model, which was based on the output of the univariate analysis. Should I be comparing the exact same model set with and without 2021 data, or going based on the output as I did here?

#### Initialize

#### Univariate models

## misclass ~ species misclass ~ nest\_age misclass ~ cam\_fate misclass ~ fdate misclass ~ obs\_int



#### multivariate models

## 

## QUESTION 3: What affects whether nest is classifiable from field evidence?

* Response: is\_u

## Including 2021 data

**NOTES:**

**Main effects model:** This one is slightly more straightforward, as long as a delta AIC of >2 is considered enough to be a different model. Luckily, the top-ranked model is also smaller (fewer terms) so it would be preferable regardless.

**Adding interactions:** The best model still seems to be the same, unless there’s some reason to give extra weight to interaction effects.

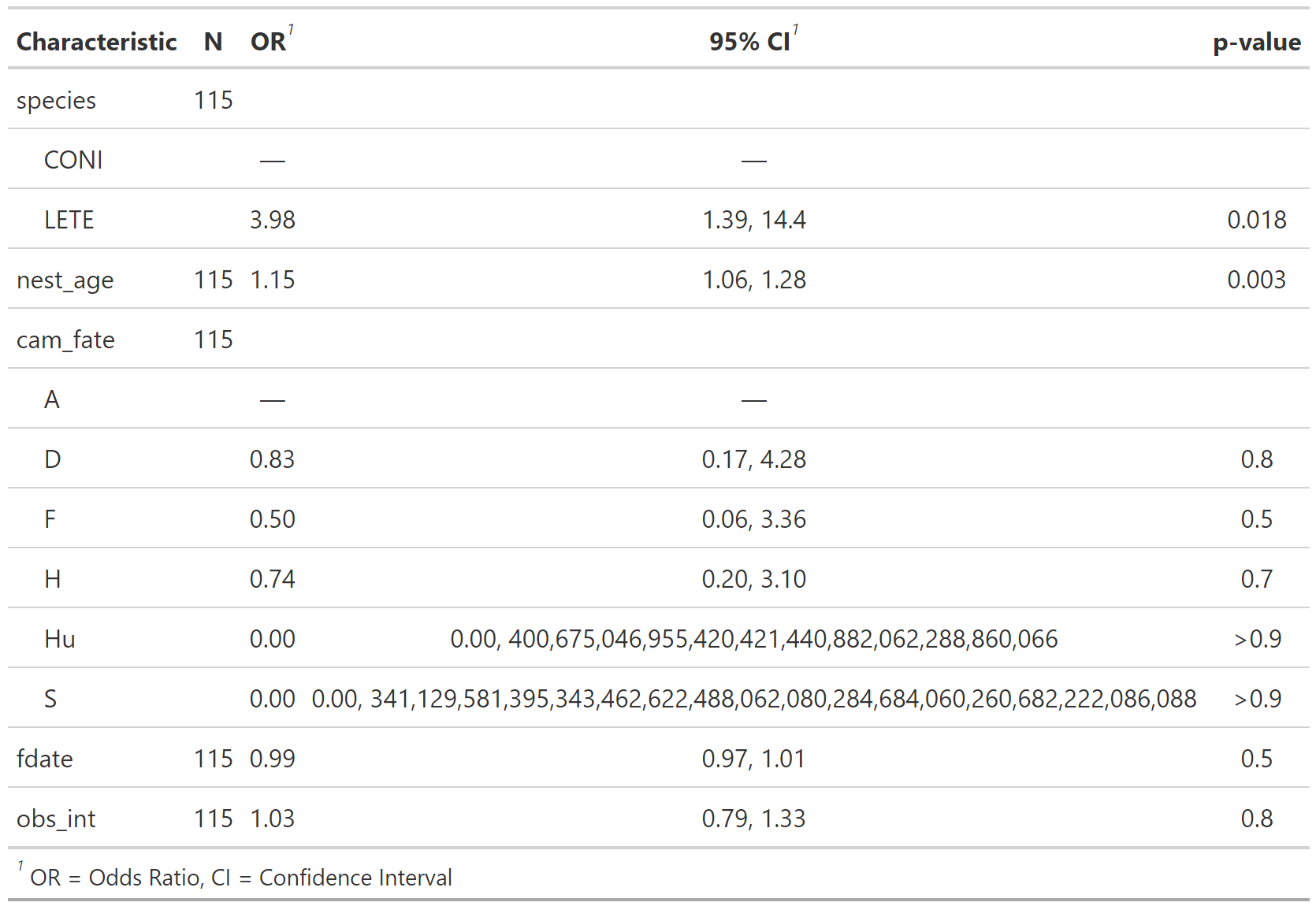
### Initialize

Choose the data and response var, create list to store models

### Univariate models

Again with the crazy large confidence intervals.

## is\_u ~ species is\_u ~ nest\_age is\_u ~ cam\_fate is\_u ~ fdate is\_u ~ obs\_int

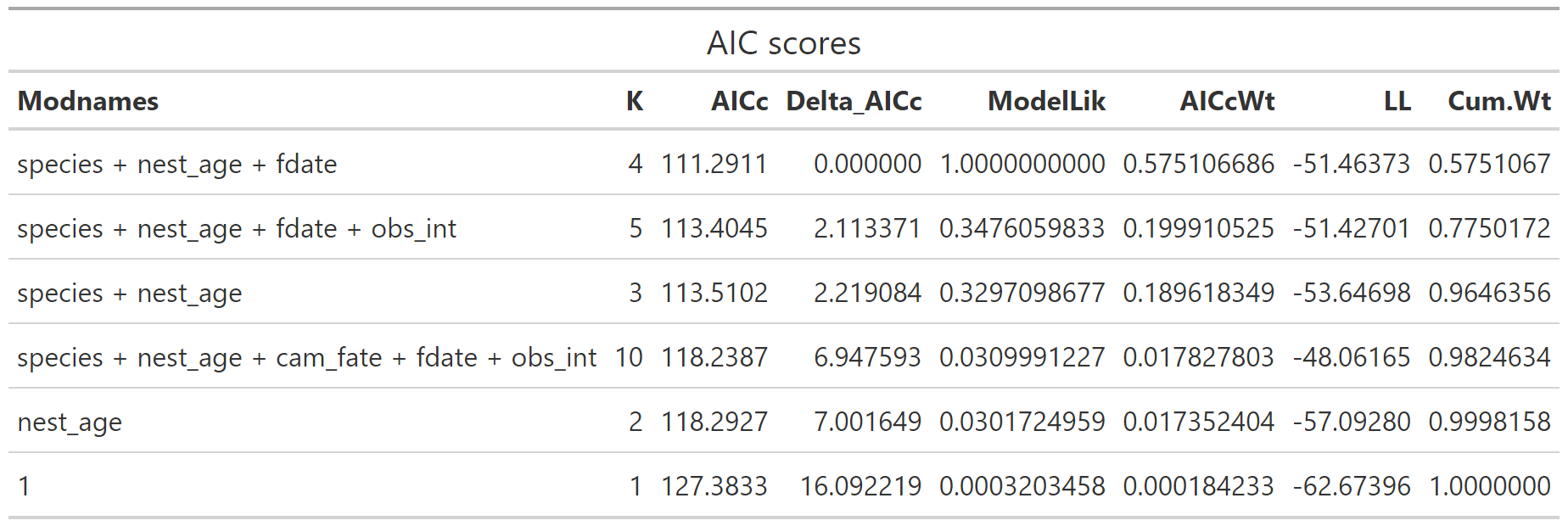


### Models

#### Null and full models

#### Other models

#### Analysis 1: main effects model



#### Interaction models

Specify chosen main effects model, compare to proposed interaction models

#### Analysis 2: including interaction models

## A screenshot of a computer Description automatically generated

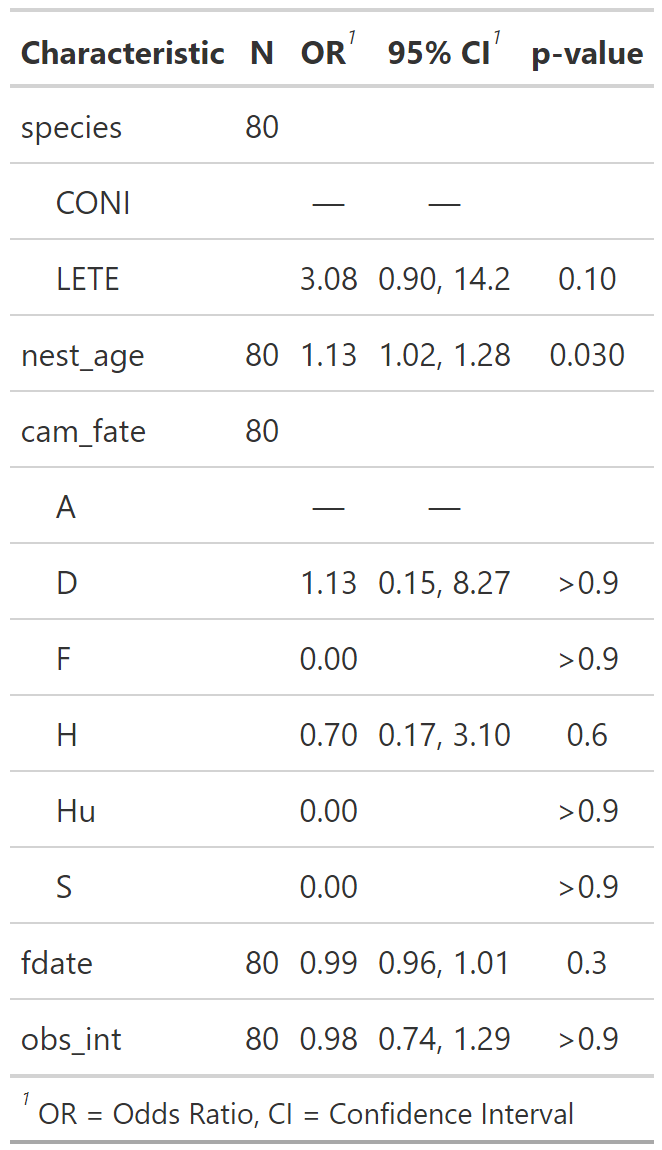
## Excluding 2021 data

**NOTES:**

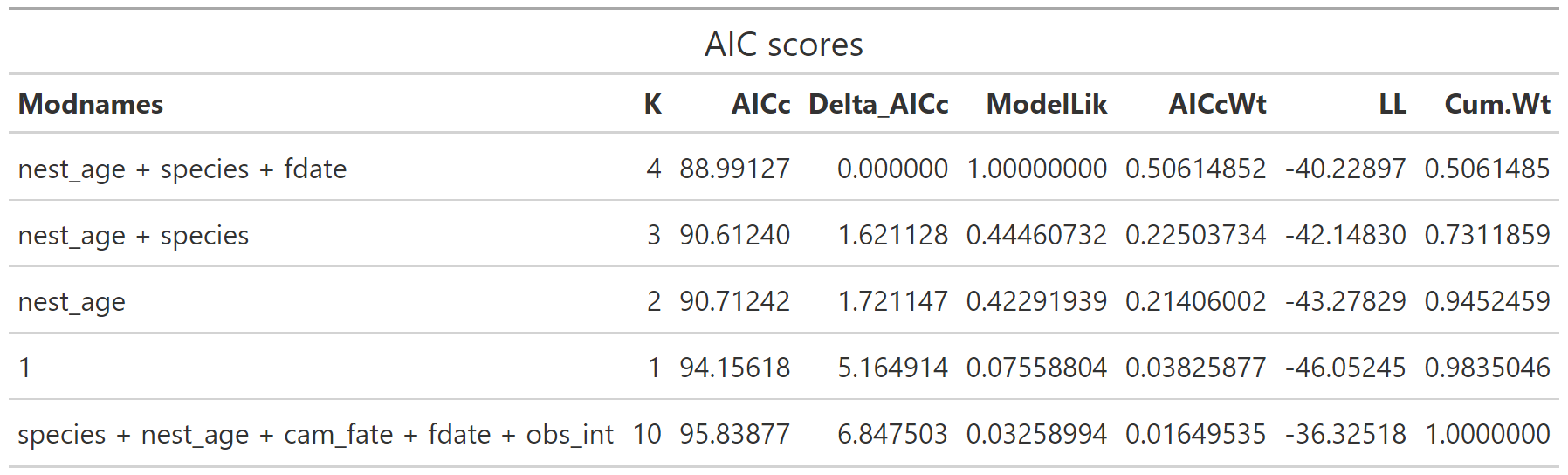
The best model excluding the 2021 data is the same as the best model including the 2021 data, which I am taking to mean that 2021 is not anomalous in some way (for this response variable) and so at least for this question, including 2021 isn’t biasing the analysis.

#### Univariate models

## is\_u ~ species is\_u ~ nest\_age is\_u ~ cam\_fate is\_u ~ fdate is\_u ~ obs\_int



#### Models



|  |
| --- |
|  |
|  |
|  |
|  |