# COMPLETE EXAMPLE

MULTINOMIAL LOGISTIC

**Data set:** multi log.csv

**Research Question:** Can we correctly predict anorexic only patients as compared to patients with bulimia?

**Classification variable/DV:**

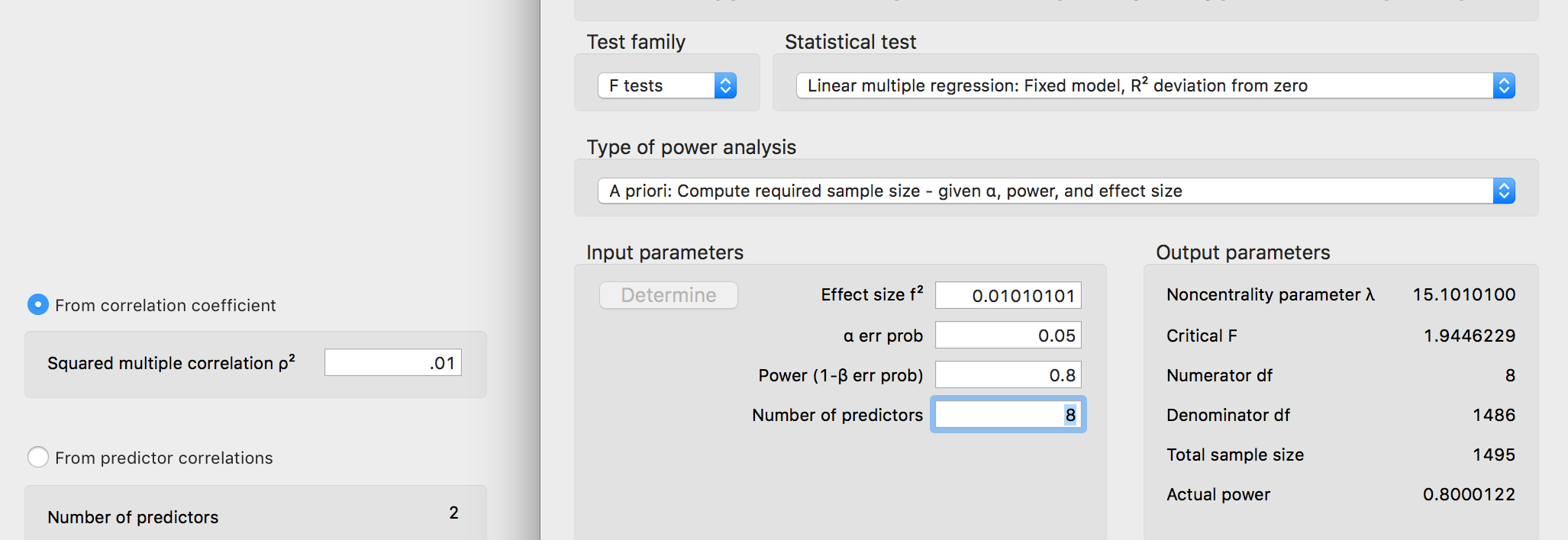
* Diag – anorexia, both, or bulimia after anorexia.

**IVs (1 to 4 Likert scale with 4 as high):**

* Weight
* Fasting
* Binging
* Vomiting
* Purging
* Hyperactivity
* Preoccupation with Body Weight
* Body Views

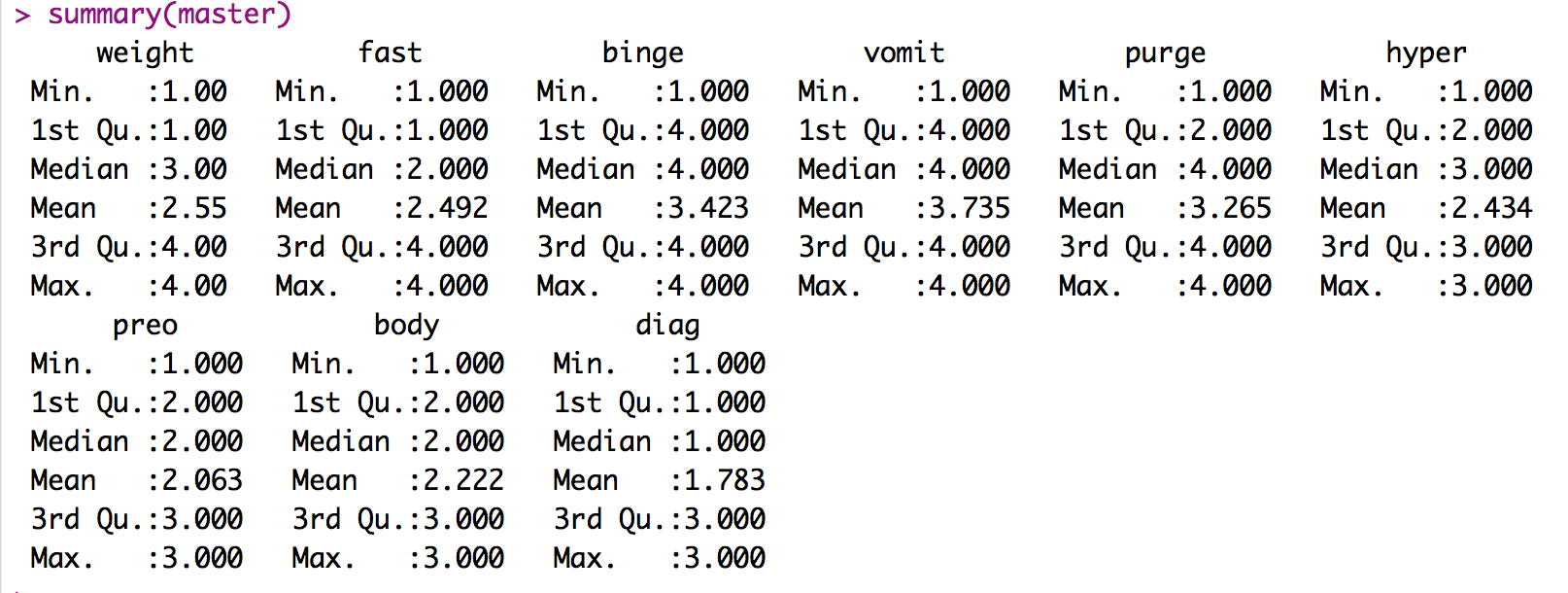
**Power:**

1. G\*Power Options:
   1. Test family: F Test
   2. Statistical test: Linear multiple regression: fixed model R2 deviation from zero
   3. Effect size: click determine 🡪 from correlation coefficient 🡪 estimate R2 🡪 calculate and transfer to main window.
   4. Alpha = .05
   5. Power (1-beta of .20) = .80
   6. Number of predictors = number of IVs.
2. Let’s try a small effect:
   1. R2 = .01
   2. Number of predictors: 8
   3. We need 1495 individuals in our study with a small effect size (yikes!).

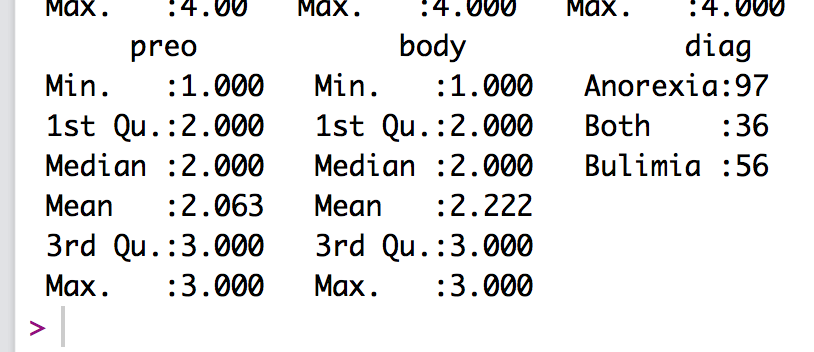


**Data screening and Assumptions:**

1. Accuracy and Ratio of Cases:
   1. Use the summary(*dataset name*) function to get the basic information for the data.
   2. In our dataset, we need the IVs to be 1-4, and they seem to be within that range.
   3. However, the DV is not a categorical variable, so we should factor that first.

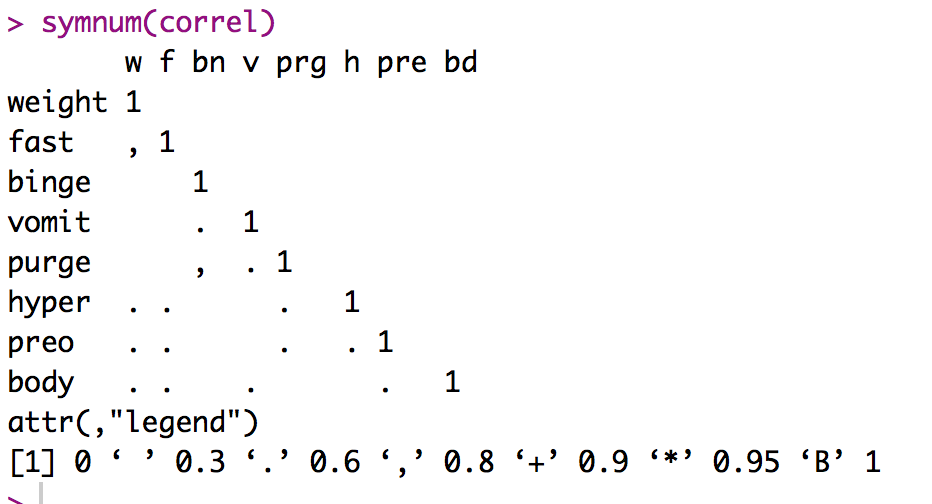


* 1. If I rerun the summary, I see that my ratio of cases is ok – not the best for one because it’s 3X the size, but it should be doable.



* 1. Important note here:
     1. Since we are running a multinomial log regression, we are going to dummy code the DV, so we will get “dummy coded” IVs for each DV comparison.
     2. One predictor will look at Anorexia versus Both, then Anorexia and Bulimia. If you wanted different comparisons, you can use the reflevel function in the mlogit package to change which one is the base (see below).

1. Missing:
   1. I can see from my summary function that I do not have missing data. Remember that you will need at least twenty variables to estimate missing data for participants – so mostly you won’t be estimating for regression.
2. Additivity
   1. Get the correlations:
      1. correl = cor(*dataset*, use = “pairwise.complete.obs”)
   2. Get the symbols chart:
      1. symnum(correl)
   3. Appears we do not have serious issues (*r* = .90) or potential suppression issues (*r* = .70).

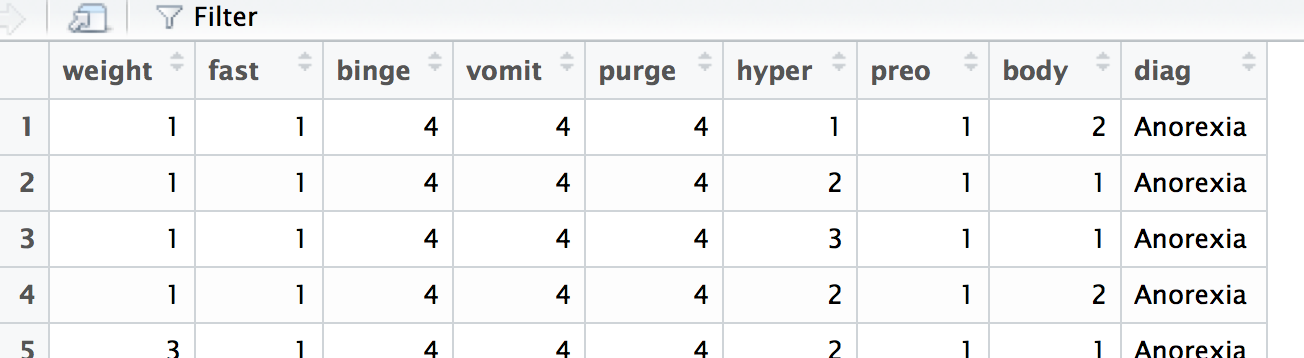


1. Install the mlogit package, and open the library.
2. Restructure the data – this package requires oddly stacked data (sorry it’s not reshape!).
   1. longdata = mlogit.data(*dataset*,

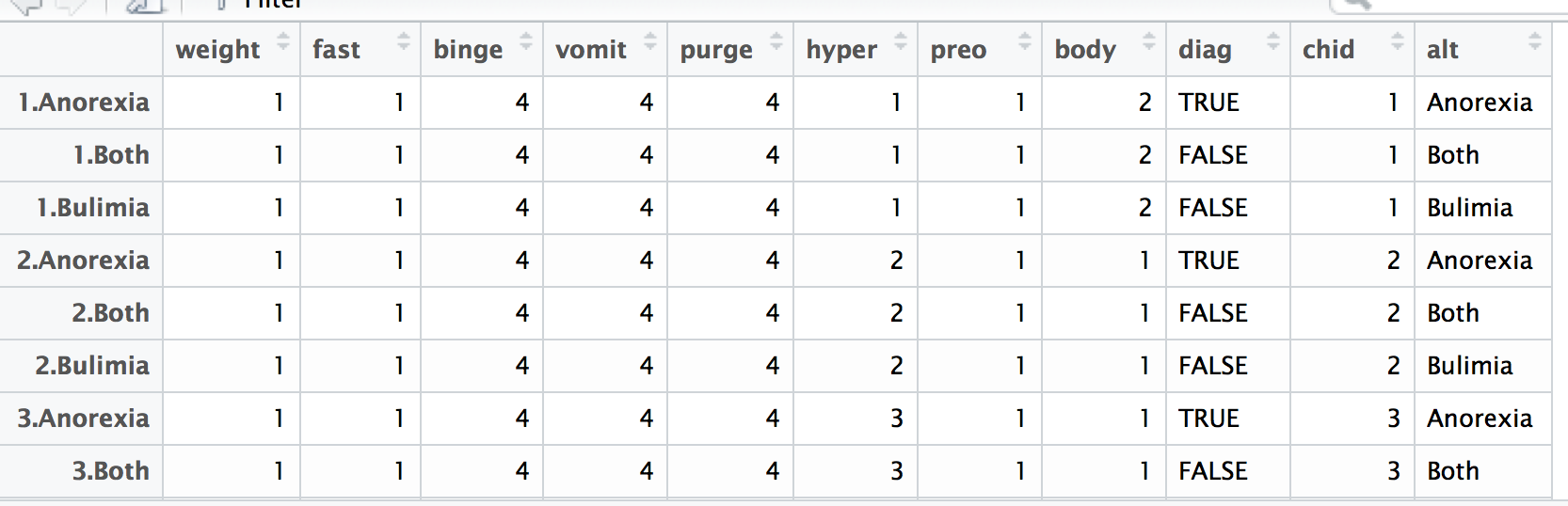
choice = "*column with DV*", shape = "wide")

* 1. Essentially, you want to fill in the name of the dataset, and the column with the DV in it. This package gets more confusing with multiple repeated measures (i.e. time 1 purge, time 2 purge) – but this restack is pretty simple with only one set of measured variables. You just leave them out of the code, and it repeats them in the final dataset.

Before:



After:

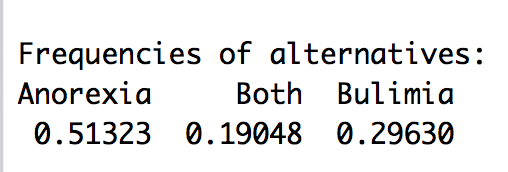


* 1. It took each person and created several lines for them – each diagnosis got it’s own line (chid, alt), and the actual diag column was restructured to be true or false for that person.
     1. This format is LONG formatted dummy coded. (ugh).

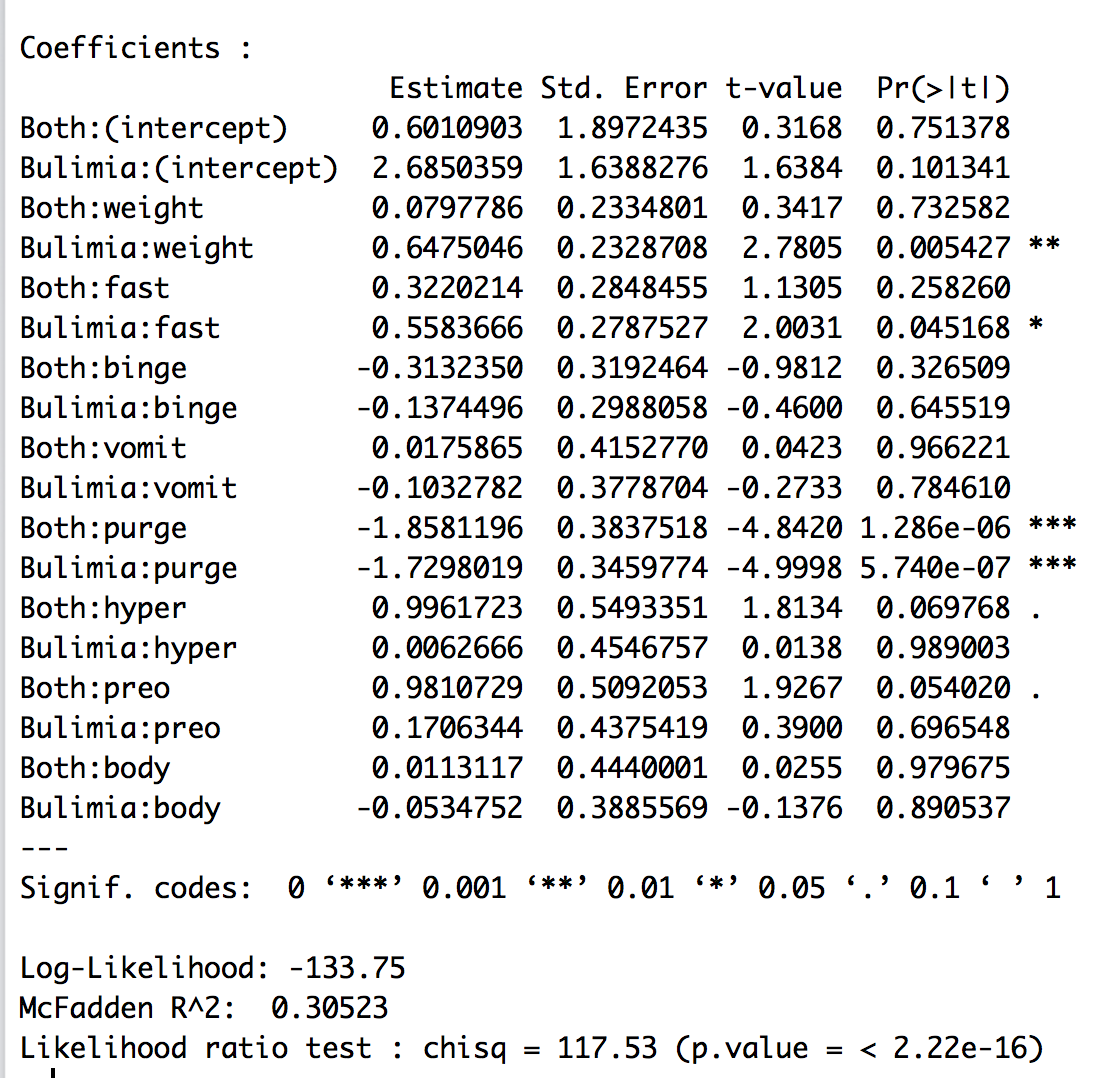
1. Run the model:
   1. model = mlogit(*DV* ~ 1 | *IV + IV + IV … ,*

data = longdata, reflevel = "*reference group*")

* 1. In the output, you get the percentages, in case you forgot to look at ratio of cases:
     1. In our example, Both is going to be hard to predict because it is the smaller group.



* 1. Then you get an unfortunate amount of output in a not so pretty format:

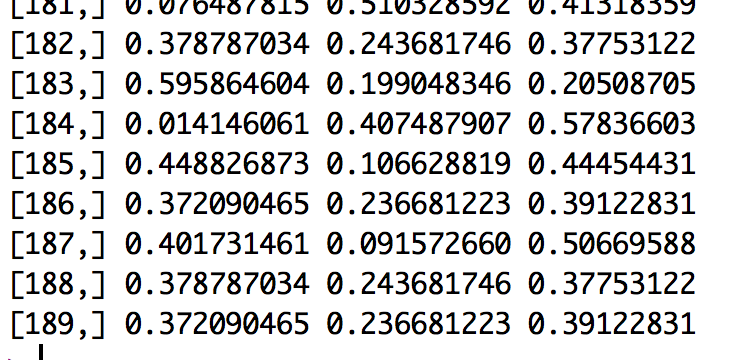


* 1. For each comparison, you get a line for each variable (i.e. anorexia v both weight, anorexia v bulimia weight, etc.).

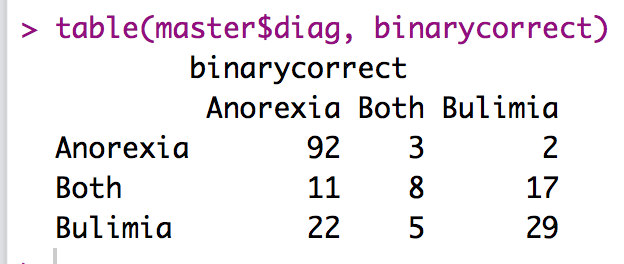
1. Is my model significant?
   1. Thankfully – we don’t even have to do any model comparisons, just read the output!
      1. Yes! X2(18) = 117.53, *p* < .001, McFadden *R*2 = .31
      2. Why McFadden? It’s right there on the output ☺.
2. Are my predictors significant?
   1. Let’s make a table!

|  |  |  |
| --- | --- | --- |
| Predictor | Anorexia V Both | Anorexia V Bulimia |
| Weight | NS | Bulimia  (more weight 🡪 higher prob of bulimia) |
| Fasting | NS | Bulimia  (more fasting 🡪 higher prob of bulimia) |
| Binge | NS | NS |
| Vomit | NS | NS |
| Purge | Anorexia  (more purge 🡪 higher prob of Anorexia) | Anorexia  (more purge 🡪 higher prob of Anorexia) |
| Hyper | Both (marginal)  (more hyperactivity 🡪 higher prob of both) | NS |
| Preoccupation with body | Both (marginal)  (more preoccupation 🡪 higher prob of both) | NS |
| Body | NS | NS |

1. Look at percent correct.
   1. How many people did we correctly classify with our results?
   2. These values are taken together to calculate a likelihood of a person being in a particular category.
      1. Remember, it’s not 50 / 50 now because we are talking about 3 categories – it’s 33% at chance for this example.
      2. Those can be found at model$probabilities.
      3. Because we have probabilities of each group, we can just categorize them using the highest likelihood.
         1. Just like the 50.1/49.9 split with binomial, that can be messy.



* 1. Use some gross code to pick which group they were classified in:
     1. correct = model$probabilities
     2. This code pulls the table of probabilities shown above for each group.
     3. binarycorrect = colnames(correct)[apply(correct,1,which.max)]
        1. This code picks the max value for each row (apply looks at each row and then picks the max using which.max), then pulls the column name for that max value.
  2. Run a table of correct versus predicted.
     1. Real is the left hand column, predicted across the top.



* 1. We want the diagonal middle to be the highest numbers (we aren’t doing so hot at both).
  2. To get the percent correct for each group use the following formula:
     1. Percent match / total for that row
     2. 92 / (92 + 3 + 2) \* 100
     3. 8 / (8 + 11 + 17) \* 100
     4. 29 / (29 + 22 + 5) \* 100To get the overall correct:
  3. Percent match of all of them / total N
     1. (92 + 8 + 29) / nrow(master) \* 100

1. You can also make a plot of the fitted values (in a continuous way) and see how well you did visually.
   1. Not in quite the same way we did before – you could plot the probabilities of one of the groups – as compared to both groups.
   2. This procedure would allow you to look for crazy scores (outliers in the solution).
   3. You wouldn’t necessarily publish this type of plot, but it helps see what’s happening.
      1. The x axis is probability of being in the selected coded group.
      2. Frequency is a bit of a mislabel, but each dot is a person.
   4. The code is included below.
      * 1. Code explanation:
   5. Open ggplot2:
      1. library(ggplot2)
   6. Run some theme coding to turn off the ugly graph defaults.

cleanup = theme(panel.grid.major = element\_blank(),

panel.grid.minor = element\_blank(),

panel.background = element\_blank(),

axis.text.y=element\_blank(),

axis.ticks=element\_blank(),

axis.line.x = element\_line(color = "black"),

axis.line.y = element\_line(color = "black"),

text = element\_text(size=20),

legend.key = element\_blank())

* 1. Create a histogram:
     1. hist = ggplot(*dataset*, aes(*probability*, color = *DV*, fill = *DV*))
  2. Add things to the histogram:
     1. Geom dotplot makes the dots rather than histogram bars. The jitter just makes sure that you can see the dots, otherwise they all stack together.
     2. Coord\_cartesian sets the x limit to 0 and 1 (because it’s probability).
     3. X lab adds the x axis label.
     4. Y lab adds the y axis label.
     5. Scale color and scale fill allow you to change the legend and dot colors – you have to include both.

hist = ggplot(master, aes(correct$Anorexia, color = diag, fill = diag))

hist +

theme +

geom\_dotplot(binwidth = .01, position = "jitter") +

coord\_cartesian(xlim = c(0,1)) +

xlab("Likelihood of Anorexia") +

ylab("Frequency") +

scale\_color\_manual(values = c("Blue", "Green", "Red"),

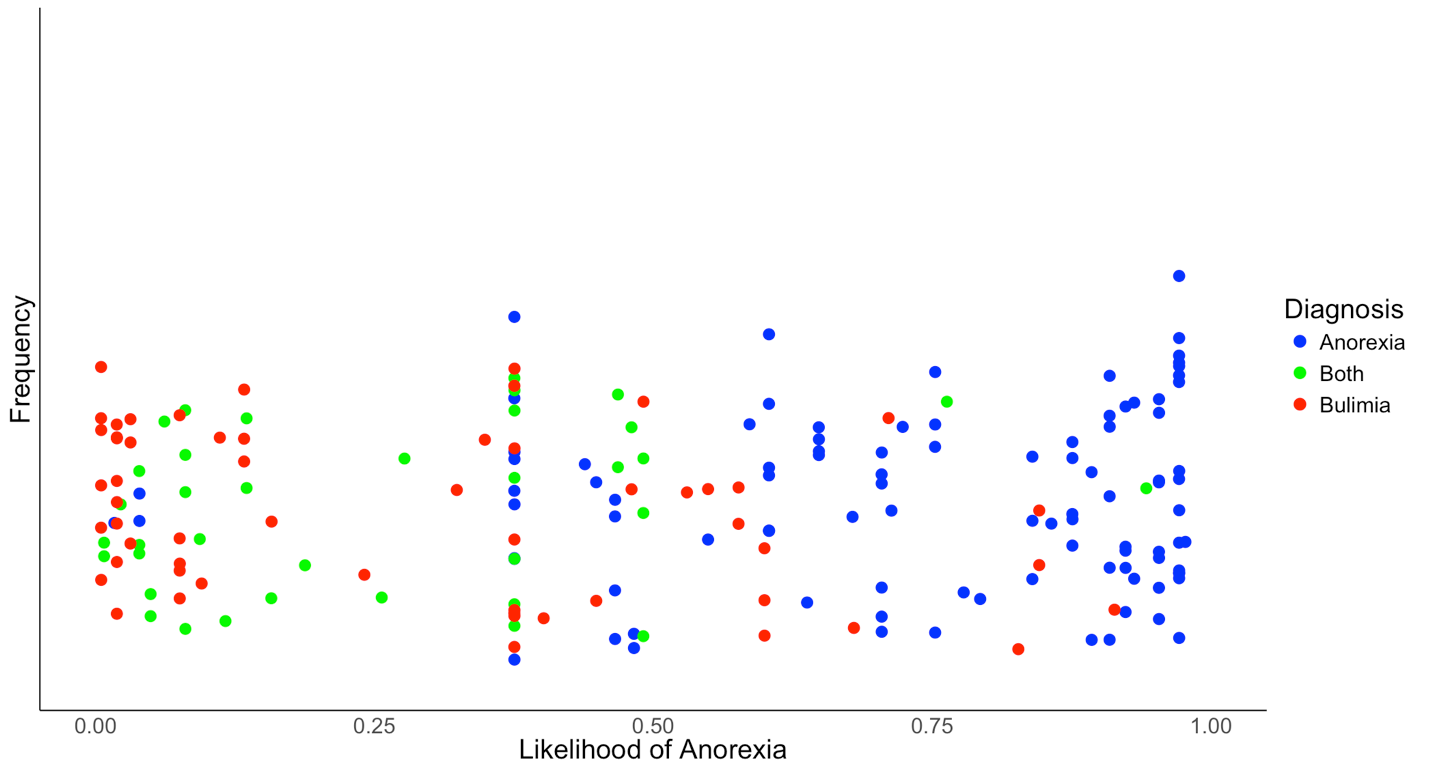
labels = c("Anorexia", "Both", "Bulimia"),

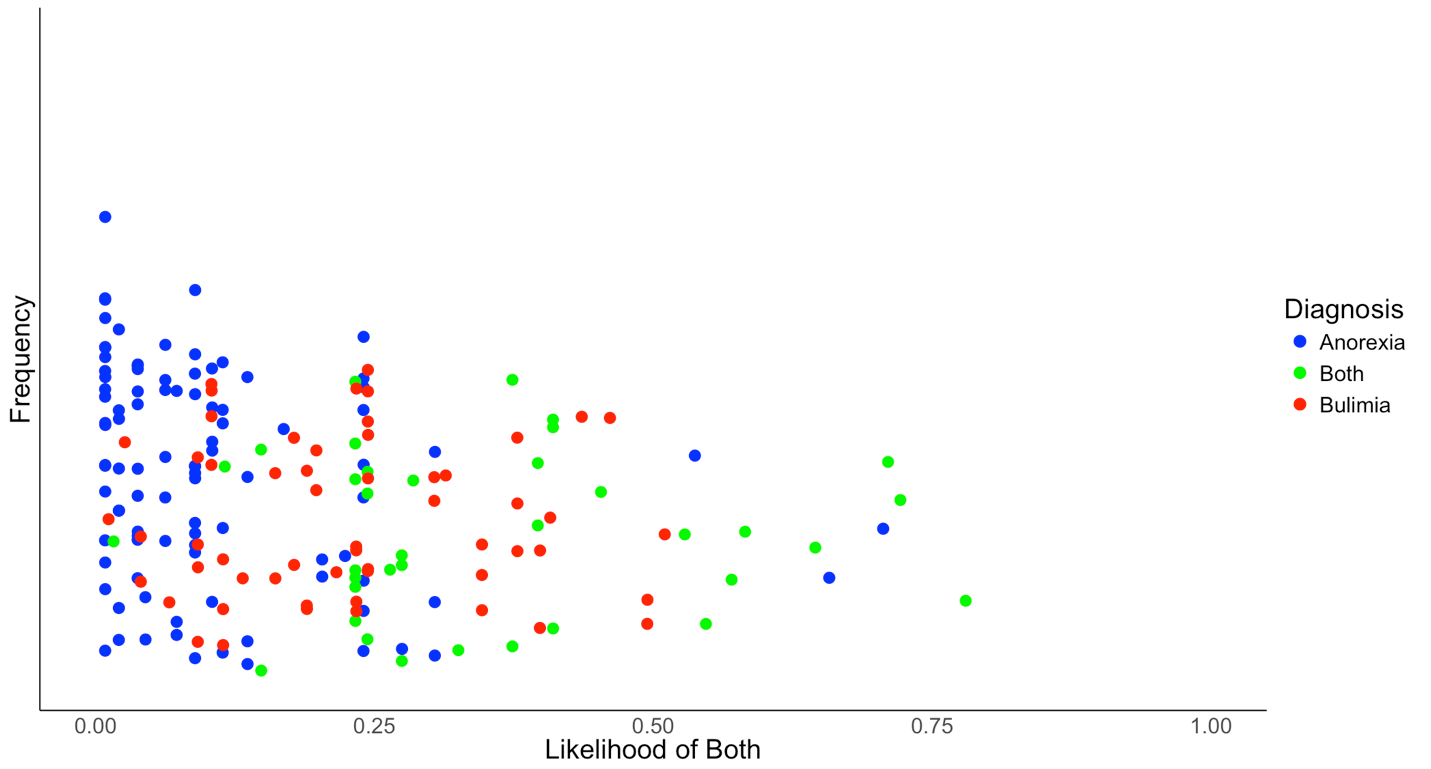
name = "Diagnosis") +

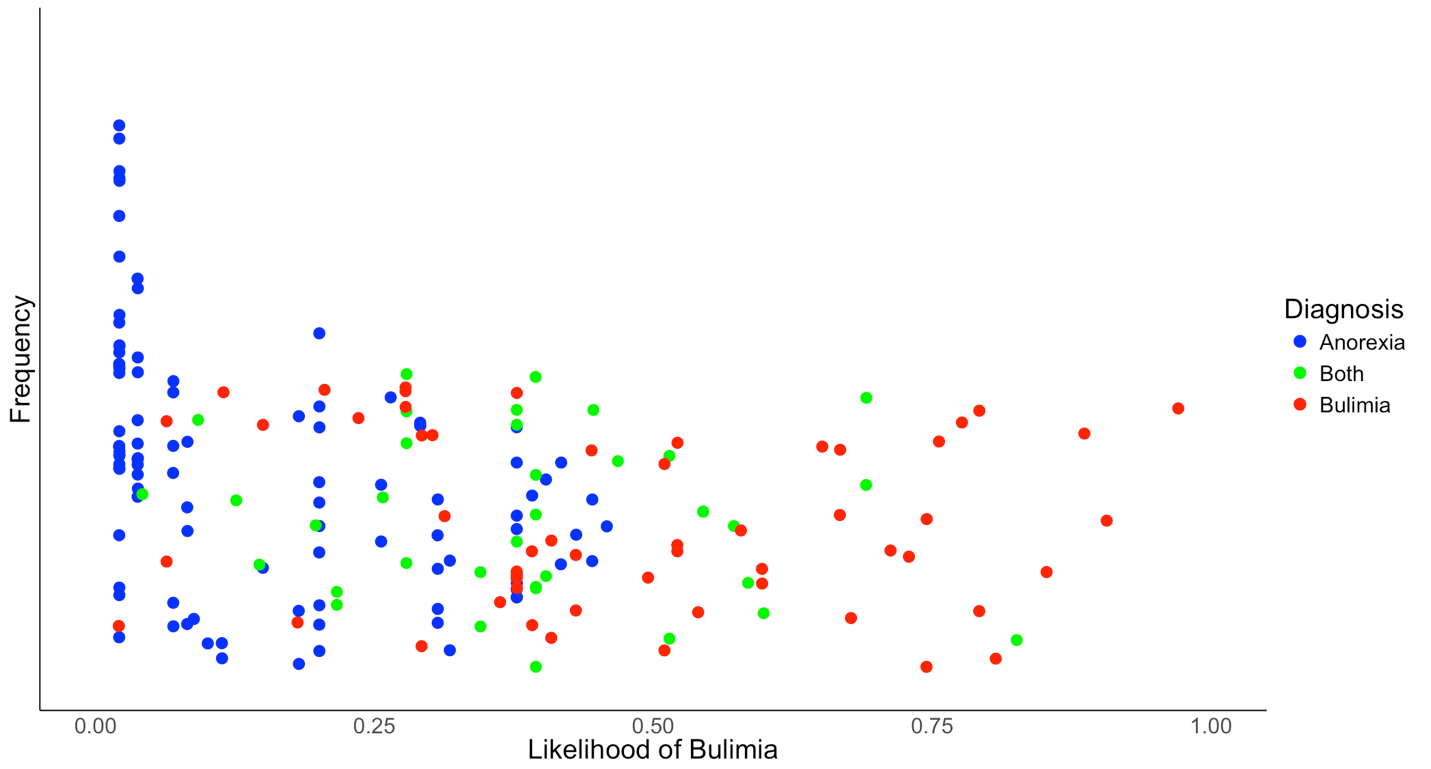
scale\_fill\_manual(values = c("Blue", "Green", "Red"),

labels = c("Anorexia", "Both", "Bulimia"),

name = "Diagnosis")







**Write up example:**

Results

A multinomial logistic regression was used to predict the differences between anorexia and bulimia and anorexia combined, as well as the difference between anorexia and anorexia that turns into bulimia. Scales were taken to determine a participant’s weight (low to normal), fasting, binging, vomiting, purging, hyperactivity, body view, and preoccupation with body weight (all none to obsessive). The data were screened and no problems (multicollinearity, missing data) appeared to be present.

Overall, the model was predictive of group classification, *X2*(16) = 117.52, *p*<.001; McFadden *R2* = .31. The model correctly classified 68.3% of participants, with 94.8% of the Anorexic patients being classified correctly, 22.2% of the patients with both, and 51.8% of patients with Bulimia.

Table 1 contains the parameter estimates for both equations. When classifying the difference between bulimia after anorexia patients and patients with anorexia, only purging and weight preoccupation were significant. Anorexics were more likely to purge, while patients with bulimia were more likely to be obsessed with weight and fast. When predicting the bulimia after anorexia and patients with both diagnoses, bulimia only patients were again more likely to be weight obsessed, while patients with both were more likely to be hyperactive. See Table 1 for predictors.

Table 1.

*Predictors for the Classification of Anorexia*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Patient Diagnosis |  | *b* | Std. Error | *t* | Sig. |
| Both | Intercept | 0.60 | 1.90 | 0.32 | 0.75 |
|  | Binge | -0.31 | 0.32 | -0.98 | 0.33 |
|  | Body | 0.01 | 0.44 | 0.03 | 0.98 |
|  | Fast | 0.32 | 0.28 | 1.13 | 0.26 |
|  | Hyperactivity | 1.00 | 0.55 | 1.81 | 0.07 |
|  | Preoccupation with Body | 0.98 | 0.51 | 1.93 | 0.05 |
|  | Purge | -1.86 | 0.38 | -4.84 | <0.01 |
|  | Vomit | 0.02 | 0.42 | 0.04 | 0.97 |
|  | Weight | 0.08 | 0.23 | 0.34 | 0.73 |
| Bulimia | Intercept | 2.69 | 1.64 | 1.64 | 0.10 |
|  | Binge | -0.14 | 0.30 | -0.46 | 0.65 |
|  | Body | -0.05 | 0.39 | -0.14 | 0.89 |
|  | Fast | 0.56 | 0.28 | 2.00 | 0.05 |
|  | Hyperactivity | 0.01 | 0.45 | 0.01 | 0.99 |
|  | Preoccupation with Body | 0.17 | 0.44 | 0.39 | 0.70 |
|  | Purge | -1.73 | 0.35 | -5.00 | <0.01 |
|  | Vomit | -0.10 | 0.38 | -0.27 | 0.78 |
|  | Weight | 0.65 | 0.23 | 2.78 | 0.01 |
| *Note.* Anorexia Nervosa is the reference category, *df* = 1. | | | |  |  |