Dummy Coding in Regression

**Description:**

* Dummy coding is factorial coding that creates pairwise comparisons for categorical variables in regression.
* Dummy coded variables are created by creating individual comparison variables from the group that is coded first.
* You will get Levels – 1 comparisons for a dummy coded variable.
  + For example, if you had four levels: Catholic, Jewish, Protestant, Other.
  + You would get THREE predictors (X variables):
    - Catholic v Jewish
    - Catholic v Protestant
    - Catholic v Other
  + These variables are the same thing as pairwise post hoc t-test comparison.
  + What if you wanted Jewish versus Protestant?
    - You would need to recode the variables using *factor* to get those comparisons.

**Dataset:** Dummy Coding

**IV:** Group health – excellent, fair, poor

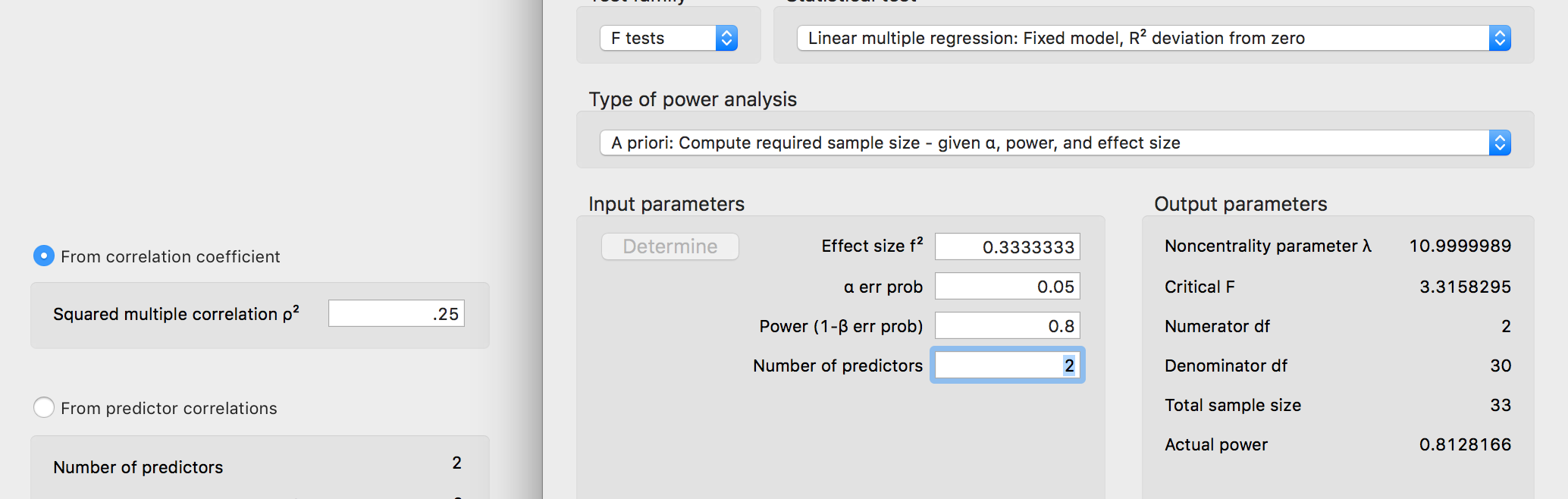
**DV:** Number of Friends

**Research question:** Does a participant’s health predict the number of friends they have?

**Regression type:** In theory, we only have one predictor (group health), but because it has multiple levels, we actually have a **set** of predictors (2 pairwise comparisons). Therefore, this analysis is a multiple linear regression – simultaneous.

**Power:**

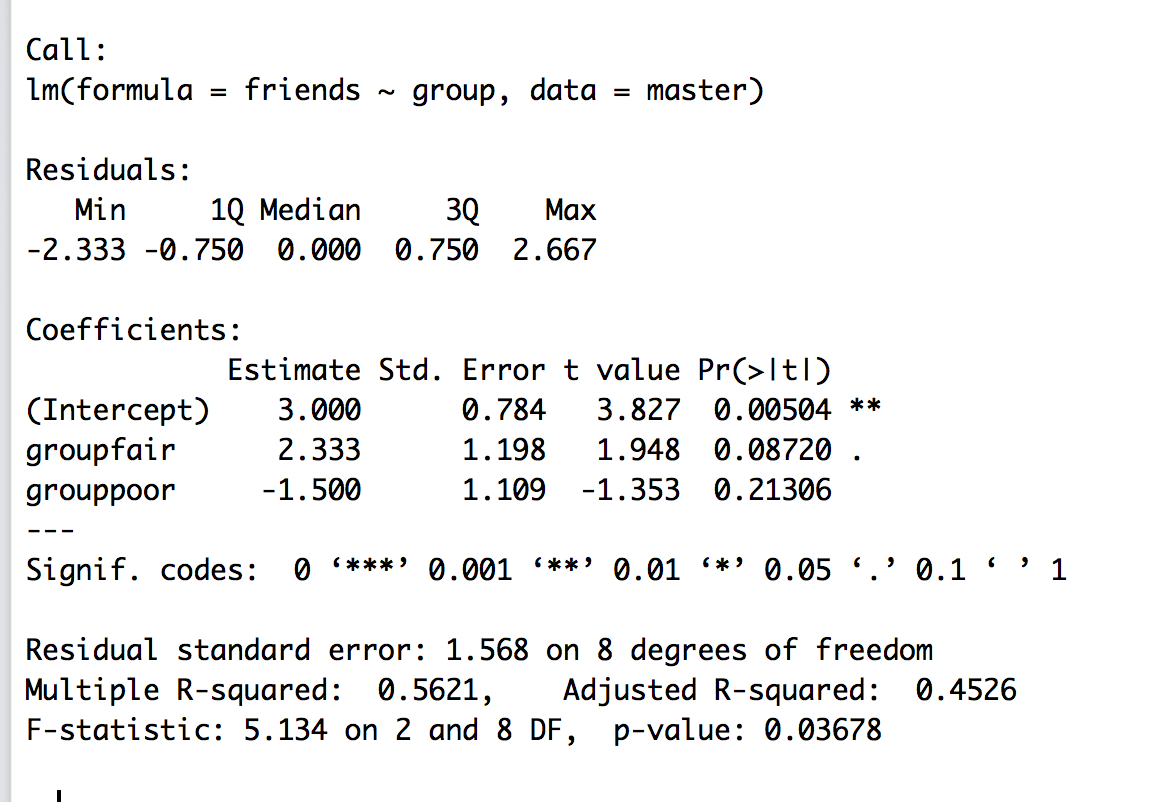
1. Open Gpower!
   1. Test family: F-test
   2. Statistical Test: Linear multiple regression: fixed model, R2 deviation from zero.
      1. We are using multiple regression because we have more than one predictor.
      2. R2 deviation from zero indicates that we are interested in the overall model, rather than asking if the addition of more predictors to previous model are useful.
   3. Estimate an effect size: click determine 🡪 use R square sizes you think might be accurate, remember small, medium, and large estimates from the notes.
   4. Alpha = .05
   5. Power (1-beta .20) = .80
   6. Number of predictors: number of IVs/X variables.
2. Let’s estimate the following:
   1. Large effect size (*R2* = .25)
   2. Number of predictors: 2
3. Says we needed to run 33 people to find a significant effect with a large effect size.



**Assumptions are discussed in more detail in the multiple linear regression notes. This guide is for interpreting dummy coded variables more specifically.**

**Run the regression:**

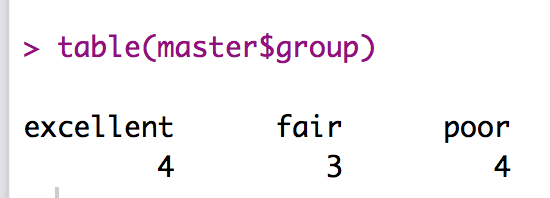
1. Regression set up:
   1. output = lm(*DV* ~ *IV*, data = *data*).
   2. summary(output)



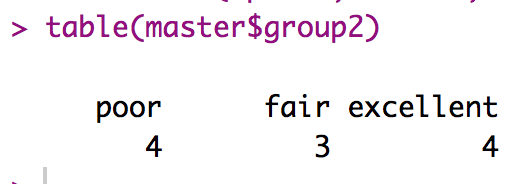
**Interpret the output:**

1. First, is the overall model significant?
   1. Yes, *F*(2, 8) = 5.13, *p* = .04, *R2 =* .56
   2. What does that mean?
      1. It means the *overall* the group variable is a significant predictor of the number of friends.
      2. Because this variable is categorical, that means that there are differences in the number of friends by group.
2. Second, are the individual predictors significant?
   1. Groupfair = comparison between group excellent and group fair. Basically, the *b* value is the mean difference between the two groups.
      1. Marginally significant difference, *b* = , *t*(8) = 1.95, *p* = .09
   2. Grouppoor = comparison between group excellent and group poor.
      1. Non significant difference, *b* = -1.50, *t*(8) = -1.35, *p* = .21.
3. What about fair to poor?
   1. We would need to recode the variable to get those values.
   2. group2 = factor(*data$column*, levels = c(“stuff”, “stuff”)
      1. Put the levels in a different order than they already are.

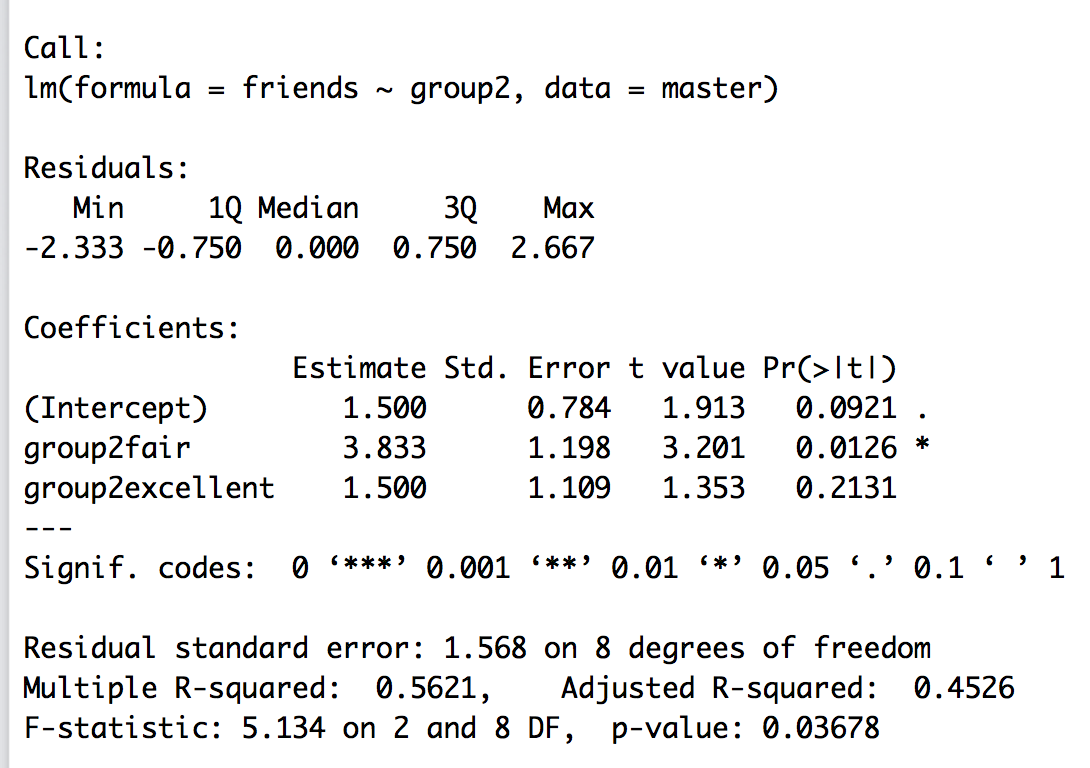
Before factoring:



After refactoring:



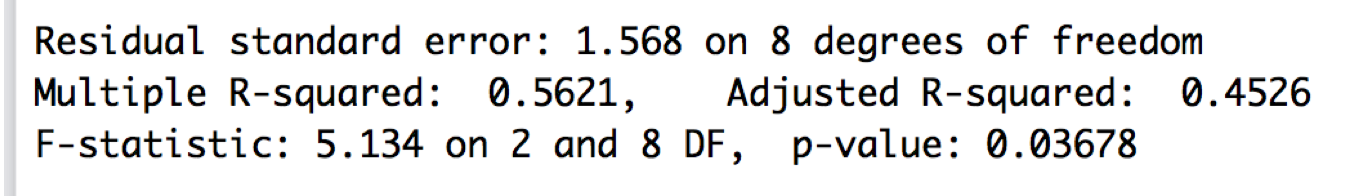
Second regression:



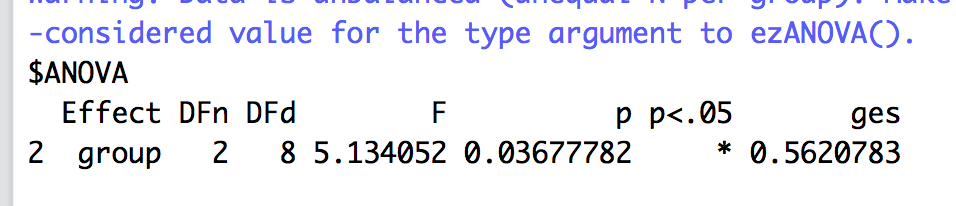
* 1. Group2fair is now group fair compared to group poor.
     1. Significant difference, *b* = 3.83, *t*(8) = 3.02, *p* = .01

1. We already did this example as an ANOVA – compare the *F* values:

Regression:



ANOVA:



How to interpret the categorical predictors:

