PSYC 5143
Spring 2024
Problem Set #3

This is due on Monday, February 12 by 2pm via Blackboard, preferably as a single R file. Ask for help before frustration sets in. When adding new predictors to your data, use fractions rather than rounded decimals if the decimals repeat. For example, if you decide that λ = 1/3 for some group, use 1/3 and not .33.

1) The file linked here has data that represent the effects of food and/or water deprivation on behavior in rats in a learning task for which the reward is food & water. There are three variables. *COND* is the condition in the experiment that rats were assigned to; **make sure this is a factor and not a character variable** when you import data. *TRIALS* is the number of trials it takes the animals to learn a task (low numbers are better). And *group* is a one-letter version of *COND* (from *b* to *e*; don't ask where *a* went), to minimize the amount of typing you have to do when creating new variables. The conditions are: two per day = fed twice on training day (a control group); food deprived (no food on training day); water deprived (no water on training day); food & water deprived (no food or water on training day).

   a. Find group means, the overall mean, and the mean of the group means.
   b. Add three predictors to the data (call these X1, X2, and X3, please) that test the following predictions/contrasts.
      i. Does deprivation of any kind lead to improved learning (relative to the control)?
      ii. Does double deprivation lead to better learning than either kind of single deprivation?
      iii. Is there a difference between food and water deprivation?
   c. Are these contrasts orthogonal? Whether yes or no, show that this the case. (Please use the `cor` function to find the correlations among X1, X2, and X3; in a balanced design like this, with equal-size groups, if the correlations are 0, these predictors are orthogonal; otherwise, they are not.)
   d. Fit the model with all three predictors you added in part b and say what the values of the intercept and slopes are with respect to the means you found in part a.
   e. What is R^2 for the whole model above?
   f. Use `lmSupport::modelEffectSizes` (or some other function you like) to find ΔR^2(i.e., sr^2) for each predictor the model using the contrast codes above. Do they sum to R^2 from part e? (They should. Try to say why.)
   g. Add three more predictors to the data using dummy coding (call them D1, D2, and D3, please) with the control group as the reference group.
   h. Fit the model and say what the values of the intercept and slopes are with respect to the means you found in part a.
   i. What is R^2 for the whole model with dummy codes as predictors?
   j. Find ΔR^2(i.e., sr^2) for each predictor in the dummy-coded model. Do they sum to R^2 for the whole model? (They won't. Try to say why if you can.)
   k. Making sure *COND* is a factor, execute `summary(lm(TRIALS ~ COND, d)` on the data (where `d` = whatever *you* named the data). How do the parameter estimates (intercept and slopes) map onto the means from part a?
   l. Going back to the contrast codes you used to fit the model in part d, redo the coding so that you use only the numerators of the λs (e.g., if a particular λ was = ½, use 1; if a particular λ = -2/3, use -2; etc.).
   m. Using the non-fractional λs from part l, refit the model and compare the slopes and *p*-values for this model to the slopes and the *p*-values from the model in part d. What changed and what stayed the same? For the things that changed, did they change systematically (e.g., were they doubled? tripled? halved?)?

don't miss the next page

2) Using the data in threeGroups.csv (which is based on the three-group design presented in class recently), complete the following.

   a. Find the group means and mean of the group means.

   b. Create two orthogonal contrasts of your choosing (you're welcome to use the ones used in class) called C1 and C2, please. Fit the model, report the parameter estimates, say how they're related to the means in part a, report *SE* for each of the two slopes, and report R^2 for the whole model.

   c. Now fit the model using only one (your choice) of the contrasts from part b, either C1 or C1 (but not both). Report the parameter estimates (the intercept and one slope this time), say how they're related to the means from part a, report *SE* for the slope, and report R^2 for the whole model.

   d. Between parts a and b, what changed? Try to say *why* the changes occurred (they might be small or large, depending on the contrasts you created in part b). A couple hints for answering part of this might come from generating CIs for the slopes in parts b and c, and then examining the formula for a CI for a slope presented in the slides from class on January 31.

   e. Now create a third contrast (it won't be orthogonal to both of the two you created in part b, though it might be orthogonal to one of them) called C3 and fit a model with all three contrasts included. Does the summary for the model include a slope for all three predictors?

3) The data in p4n3.csv contains four variables: *group1* (a numeric variable numbered 1, 2, and 3 to indicate group membership); *group2* (a numeric variable numbered 1, 3, and 2 to indicate group membership); *groupF* (a character variable indicating group membership with the labels *g1*, *g2*, and *g3*); and *dv* (an outcome variable.) Import these data, **make sure groupF is treated by R as a factor**, and answer the following questions.

   a. Fit the linear model `lm(dv ~ groupF)`. Report the *F*-ratio for the whole model along with its *df*. Also find the means of the three groups. Also report *SSE* for this model (also known as $SS_{residuals}$), which you can get with the `anova()` function, among others.

   b. Fit the linear model `lm(dv ~ group1)`. Report the *F*-ratio for the whole model along with its *df*, as well as *SSE*.

   c. Fit the linear model `lm(df ~ group2)`. Report the *F*-ratio for the whole model along with its *df*, as well as *SSE*.

   d. Which of the three models you've fit has the lowest *SSE* value? Try to say why.

   e. You should find that the models in parts b and c have different *F*-ratios and *SSE* values. They both model the outcome on a numeric variable that simply codes group membership with a number. Why are they giving different results? It might help to look at the slopes from these models and possibly scatterplots of the outcome (on the y-axis) and the predictor (group1 or group2) on the x-axis, maybe with best-fitting lines added.