

Multinomial Regression

Regarding your data, what you have is a response with multiple categories, and anytime you are trying to model a response which is categorical you are right to try and use some type of generalized linear model (GLM). In your case you have additional information which you must take into account regarding your response and that is that your response levels have a natural ordering good \bar{c} middle \bar{c} bad, notice how this is different from trying to model a response such as what color balloon someone is likely to buy (red/blue/green), these values have no natural ordering. When doing this type of model with an ordered response you may want to consider using a proportional odds model.

I haven't used it myself, but the `polr()` function in the MASS package is likely to be of some use, alternatively I have used the `lrm()` function in the rms package to do similar types of analysis, and have found it quite useful. If you load these packages just use `?polr` or `?lrm` for the function information.

Alright enough background, on to your questions:

This should be covered above, check out these packages/functions and read up on ordinal logistic regression and proportional odds models

Any time you have a covariate which is categorical (Race/Sex/Hair color) you want to treat these as 'factors' in your R coding in order to model them appropriately. It's important to know what a factor is and how they are treated, but essentially you treat each category as a separate level and then model them in an appropriate way.

Just read up on factors in models and you should be able to tease out whats going on. Keep in mind that treating categorical variables as factors is not unique to glm models or proportional odds models, but is typically how all models deal with categorical variables. <http://www.stat.berkeley.edu/classes/s133/factors.html>

Missing values can sometimes be a hassle to deal with but if you're doing a fairly basic analysis its probably safe to just remove data rows which contain missing values (this isn't always true, but based on your current experience level I'm guessing you need not be concerned with the specifics of when and how to deal with missing values). In fact this is pretty much what R does. If you have a data which you are using to model, if you are missing information in a row for your response or any covariate in the model R is just going to exclude this data (this is the warning your seeing).

Obviously if you're excluding a large proportion of your data due to missingness, your results could be biased and it's probably good to try and get some more info on why there are so many missing values, but if you're missing 162 observations in 10,000 rows of data I wouldn't sweat it too much. You can google up on methods for handling missing data if you're interested in some more specifics.

Almost all R model objects (`lm`, `glm`, `lrm`,...) will have an associated `predict()` function which will allow you to calculate the predicted values for your current modeling dataset and additionally for another dataset which you wish to predict an outcome for. Just search `?predict.glm` or `?predict.lm` to try and get some more info for whatever model type you want to work with. This is a very typical thing people wish to do with models so rest assured that there are some built in functions and methods that should make doing this relatively straightforward.