Understanding `factors` in `R` Practice

Ryan Safner

11/19/2018

Let's generate some random data. We will make a `data.frame` called `df` (feel free to call it something else.)

```
set.seed(1) # this makes data reproducible, if we set the same seed #, we all get the same "random" dat

colors<-c("red", "orange","yellow", "green", "blue", "purple")
color<-sample(colors,100,replace=TRUE)

educ.levels<-c("high school","college", "graduate degree")
education<-sample(educ.levels, 100, replace=TRUE)


# make a data frame of three variables, x, y, and color
df<-data.frame(x=rnorm(100,2,1), # x is 100 draws from a normal distr with mean 2 and sd 1
                 y=rnorm(100,5,1), # y is 100 draws from a normal distr with mean 5 and sd 1
                 color=factor(color), # color is a factor
                 education=ordered(education, levels=c("high school", "college", "graduate degree")) #
               )
```

1. Get a summary of `df`. Also check its `str()` and `head()` to get a closer look. Look at the data itself with `View(df)`.

2. Look more closely at `color`. Check its `class()`, `nlevels()` and the actual `levels()`. Finally, make a `table()` of the counts of each category.

3. Make a barplot of `color`.

4. Look more closely at `education`. Check its `class()`, `nlevels()` and the actual `levels()`. Finally, make a `table()` of the counts of each category.

5. Make a barplot of `education`.

6. Now let's try looking at plots by different categories. Make a scatterplot of `x` and `y` and add `color=color` to your base layer `aes()`.

a. Now let's try subsetting. Plot only data for the color `green`.

b. In addition to your `geom_point()`, add a `geom_smooth(method="lm")` regression line. Notice it makes a regression line for each color. If we want an overall regression line, we need to redo our scatterplot as follows. In the base layer, don't include `color` in your `aes()`, move it instead inside `geom_point(aes(color=color))`. Then add a `geom_smooth()`, it will do it for the overall plot.

c. Now let's try subsetting. Make a scatterplot and regression line only with data points for the color `green`.

d. Let's simply use the `facet_grid()` command to plot all the different colors as different plots. Reuse your commands from the first plot in this question, and then add a `facet` layer with `+facet_grid(cols=vars(color))`

7. Run through problem #6 again, but using `education` instead of `color`.

8. Now let's try some regression.

a. Run a regression of `y` on `x` and `education`. What happens?

b. `R` was generous and did the work for you! But let's do the same thing ourselves manually. What we need to do is convert `education` into a three dummy variables, one for each level of education. It's easiest to use the `ifelse()` command here. Remember the syntax: `ifelse(condition, do.this.if.true, do.this.if.false)`. Check your data `df` again with `head()` or `View()` to make sure you properly coded the variables.

c. Now run a regression of `y` on `x` and all of your new dummy variables. What happens, and why?

d. Run three different regressions, each one omitting one of the different categories of education. Interpret your coefficients.