

Prac 5.14

[f2007/f2007-126]

For the simple model $A \sim G$ where G is a categorical variable, the coefficients will be group means. More precisely, there will be an intercept that is the mean of one of the groups and the other coefficients will show how the mean of the other groups each differ from the reference group.

Similarly, when there are two grouping variables, G and H , the model $A \sim G + H + G:H$ (which can be abbreviated $A \sim G*H$) will have coefficients that are the groupwise means of the crossed groups. Perhaps “subgroup-wise means” is more appropriate, since there will be a separate mean for each subgroup of G divided along the lines of H . The interaction term $G:H$ allows the model to account for the influence of H separately for each level of G .

However, the model $A \sim G + H$ does **not** produce coefficients that are group means. Because no interaction term has been included, this model cannot reflect how the effect of H differs depending on the level of G . Instead, the model coefficients reflect the influence of H as if it were the same for all levels of G .

To illustrate these different models, consider some simple data.

Suppose that you found in the literature an article about the price of small pine trees (either Red Pine or White Pine) of different heights in standard case/variable format, which would look like this:

Case #	Color	Height	Price
1	Red	Short	11
2	Red	Short	13
3	White	Tall	37
4	White	Tall	35

and so on ...

Commonly in published papers, the raw case-by-case data isn’t reported. Rather some summary of the raw data is presented. For example, there might be a summary table like this:

SUMMARY TABLE

Mean Price			
Color			
Height	Red	White	Both Colors
Short	\$12	\$18	\$15
Tall	\$20	\$34	\$27
Both Heights	\$16	\$26	\$21

The table gives the mean price of a sample of 10 trees in each of the four overall categories (Tall and Red, Tall and White, Short and Red, Short and White). So, the ten Tall

and Red pines averaged \$20, the ten Short and White pines averaged \$18, and so on. The margins show averages over larger groups. For instance, the 20 white pines, averaged \$26, while the 20 short pines averaged \$15.

The average price of all 40 trees in the sample was \$21.

Based on the summary table, answer these questions:

1. In the model $\text{price} \sim \text{color}$, which involves the coefficients “intercept” and “color-White”, what will be the values of the coefficients?

- Intercept 12 15 16 18 20 21 26 27 34 Prac 5.14-1
- colorWhite -10 -8 0 5 8 10 Prac 5.14-2

2. In the model $\text{price} \sim \text{height}$, which involves the coefficients “intercept” and “height-Tall”, what will be the values of the coefficients?

- Intercept 0 4 8 12 15 16 18 20 21 26 27 34 Prac 5.14-3
- heightTall 0 4 8 12 15 16 18 20 21 26 27 34 Prac 5.14-4

3. The model $\text{price} \sim \text{height} * \text{color}$, with an interaction between height and color, has four coefficients and therefore can produce an exact match to the prices of the four different kinds of trees. But they are in a different format: not just one coefficient for each kind of tree. What are the values of these coefficients from the model? (Hint: Start with the kind of tree that corresponds to the intercept term.)

- Intercept 0 4 6 8 10 12 16 Prac 5.14-5
- heightTall 0 4 6 8 10 12 16 Prac 5.14-6
- colorWhite 0 4 6 8 10 12 16 Prac 5.14-7
- heightTall:colorWhite 0 4 6 8 10 12 16 Prac 5.14-8

4. The model $\text{price} \sim \text{height} + \text{color}$ gives these three coefficients:

- Intercept : 10
- heightTall : 12
- colorWhite : 10

It would be hard to figure out these coefficients by hand because they can’t be read off from the summary table of Mean Price.

According to the model, what are the fitted model values for these trees:

- Short Red 10 12 15 16 20 22 32 34 Prac 5.14-9
- Short White 10 12 15 16 20 22 32 34 Prac 5.14-10
- Tall Red 10 12 15 16 20 22 32 34 Prac 5.14-11

- Tall White 10 12 15 16 20 22 32 34 Prac 5.14-12

Notice that the fitted model values aren't a perfect match to the numbers in the table. That's because a model with three coefficients can't exactly reproduce a set of four numbers.