

Exer 6.1

[H/H124]

Here are some (made-up) data from an experiment growing trees. The height was measured for trees in different locations that had been watered and fertilized in different ways.

height	water	light	compost	nitrogen
5	2	shady	none	little
4	1	bright	none	lot
5	1.5	bright	some	little
6	3	shady	rich	lot
7	3	bright	some	little
6	2	shady	rich	lot

1. In the model expression $\text{height} \sim \text{water}$, which is the explanatory variable?

- A height
- B water
- C light
- D compost
- E Can't tell from this information.

Exer 6.1-1

2. Ranger Alan proposes the specific model formula

$$\text{height} = 2 * \text{water} + 1.$$

Fill in the table showing the model values and the residuals.

height	water	model values	resids
5	2		
4	1		
5	1.5		
6	3		
7	3		
6	2		

3. Ranger Bill proposes the specific model formula

$$\text{height} = \text{water} + 3.$$

Again, make a table of model values and residuals.

height	water	model values	resids
5	2		
4	1		
5	1.5		
6	3		
7	3		
6	2		

4. Which of the two models is better? Give a specific definition of “better” and explain your answer quantitatively.

Exer 6.1-2

5. Write down the set of indicator variables that arise from the categorical variable compost.

Exer 6.1-3

6. The fitted values are exactly the same for the two models $\text{water} \sim \text{compost}$ and $\text{water} \sim \text{compost}-1$. This suggests that the $\mathbf{1}$ vector $(1, 1, 1, 1, 1, 1)$ is redundant with the set of indicator variables due to the variable compost. Explain why this redundancy occurs. Is it because of something special about the “compost” variable?

Exer 6.1-4

7. Estimate, as best you can using only very simple calculations, the coefficients on the model $\text{water} \sim \text{compost}-1$. (Note: there is no intercept term in this model.)

Exer 6.1-5

8. Ranger Charley observes that the the following model is perfect because all of the residuals are zero.

$$\text{height} \sim 1 + \text{water} + \text{light} + \text{compost} + \text{nitrogen}$$

Charley believes that using this model will enable him to make excellent predictions about the height of trees in the future. Ranger Donald, on the other hand, calls Charley’s regression “ridiculous rot” and claims that Charley’s explanatory terms could fit perfectly any set of 6 numbers. Donald says that the perfect fit of Charley’s model does not give any evidence that the model is of any use whatsoever. Who do you think is right, Donald or Charley?

Exer 6.1-6