Topic: Linear Regression
Predicting Soil Fertility for Agricultural Planning and Soil Management

BIDS, gwendolin.wilke@hslu.ch

Agricultural production depends on the nature and quality of soil used to grow crops. Farmers have come to rely on high quality soil to improve their crop yield and ensure a sustained production of food crops. A loss in the fertility of soil used for cultivation can harm the yield per acre, resulting in loss of agricultural production. There is hence a need for efficient soil testing, inspection, and certification services that check the quality of the soil. Your data set contains 147 soil profile observations from southwestern Cameroon.

Data Set Description

Name: `soil.csv`
Description: Southwestern Cameroon soil samples.
Attributes:

Three different attributes have been sampled, each for three different depth layers (0–10 cm, 10–20 cm, and 30–50 cm):

- *Clay1, Clay2, Clay5*: Clay content in weight %.
- *OC1, OC2, OC5:* Organic carbon in volume %.
- *CEC1, CEC2, CEC5:* Cation exchange capacity in cmol+(kg soil)-1.

CEC is a measure of soil fertility. The higher the CEC value, the better the soil can store nutrients, and consequently the higher is its fertility. The CEC value of a soil depends on clay content and/or on organic carbon (such as humus). The CEC is important for agricultural planning and soil management, since it controls how much added artificial or natural fertilizer will be retained by the soil for a long-lasting effect on crop growth. Heavy doses of fertilizer on soils with low CEC will be wasted, since the extra nutrients will leach. For farmers, it is essential to know the CEC content of all soil depth layers.

Assignment

1. Load the data set `soil.csv` and view it using the `View()` command.
2. Discuss the dataset based on `str()` and `summary()`.[1]
3. Your goal is to use `Clay` and OC to predict `CEC`. Get a first impression on the predictive capabilities of your data by plotting all the `Clay` and OC variables against all `CEC` variables.[2] Discuss your plots.
4. Build *simple*[3] linear regression models for all of the above variable pairs. Plot your results.
5. Based on the R-squared value, what is the best predictor for top-soil CEC (CEC1), mid-soil CEC (CEC2) and sub-soil CEC (CEC5), respectively?
6. Based on the R-squared value, what is the best predictor for the sub-soil CEC value, given top-soil samples of Clay, OC and CEC? Did you expect this outcome? What is the straight-line equation of this predictor?
7. Do a residual plot for this predictor and interpret it.
8. Using the above predictor, what is the sub-soil CEC value predicted for a soil with no topsoil clay? What is the sub-soil CEC value predicted for soil with 70 weight-% of topsoil clay? Plot and interpret your result.

---

[1] If you have never heard of Quartiles or the Median, look it up on Wikipedia. What does it tell you?

[2] Use `plot(yourData)` to displays a matrix of scatter plots.

[3] You may use multiple linear regression. Yet, in this case make sure that your inputs are not correlated.