Assignment

Created by Blas Mola, last modified on Dec 08, 2022

Assignment: analysing and reporting data

Marina PERIS-LLOPIS, Blas MOLA-YUDEGO

Submission instructions

Send the final report by email (blas.mola@uef.fi) before the agreed deadline.

Data descriptions

In this study you will use two data sets:

1) Modeling dataset (csv) (each group has own data set, see group-wise file names above)
2) Validation dataset (csv) (all groups use the same validation data)

Both data sets include the following variables:

ID – stand id

LAT – latitude

LONG - longitude

SP_GROUP – dominant species (1: Scots pine, 2: Norway spruce, 4: Birch and other deciduous)

AGE – stand age (years)

BA – stand basal area (m2/ha)

D – average diameter (cm)

H – average height (m)

FOREST_TYPE – site fertility class (1: very fertile, 2: fertile, 3: semi-fertile, .., 7: very poor)

VOLUME – stand volume (m3/ha)

p0 - Annual precipitation (mm)

Objectives

To build a linear regression model, which can be used **for estimating total stem volume in the stand** *(m^3 ha^{-1})* as a function of basic stand characteristics (e.g. diameter, height, possibly other variables).

Task: Create linear regression models with different independent variables (and possibly converted variables) and **select the best model** among your trials. You can follow the next instructions:

Modeling

1. Import the **modeling data set**
2. Calculate **summary statistics** for ratio scale variables in your subset (min, max, arithmetic mean, standard deviation, total number of plots). Draw histogram of the total volume, and boxplots on the total volume by dominant species and different forest types. *(summary(DATASET), sd(DATASET$VAR1), length(DATASET$VAR1), hist(DATASET$VAR1), boxplot(DATASET$VAR1~DATASET$VAR2).* Use this data in “Material” chapter where you describe your modeling data.
3. Examine relationships of **the total stem volume** and **other variables** by drawing scatter plots between them. *(plot(DATASET$X, DATASET$Y, xlab='your_x_label_name',ylab='your_y_label_name'))*
4. Try building a model for estimating your **dependent Y variable (the total stem volume)** as a function of other stand variables **(independent X variables)**. Examine outputs (R^2, p-values, standard error of residual etc) and normality and homoscedasticity of residuals. *(your_model_name=lm(DATASET$Y~DATASET$X), summary(your_model_ name), plot(your_model_name))*

Instructions

Each student can work individually or in groups of two, and belongs to a group number, that will deal with a specific dataset. The group number correspond to the last digit of the student number. For groups of two persons, to either last digit of the student numbers of the participants.

Modelling dataset

Group1.csv
Group2.csv
Group3.csv
Group4.csv
Group5.csv
Group6.csv
Group7.csv
Group8.csv
Group9.csv
Group10.csv
Group11.csv
Group12.csv
Group13.csv
Group14.csv
Group15.csv
Group16.csv
Group17.csv
Group18.csv
Group19.csv

Validation dataset

ValidationDataset.csv

Further information

Visualisation [ppt]
Validation [ppt]
RMSE and BIAS [pdf]

*You can build a model with **multiple X variables**, if needed. Remember, that the X variables should be at ratio scale. Variables like site class or tree species can be included in the model only as dummy variables (e.g. factor(your_dataframe$FOREST_TYPE)).*

1. Select 3 best models among your trials. Save the following data:

- outputs of the selected models (summary(your_model_name)
- scatter plots (independent variables plotted against dependent variable)
- residual figures (plot(your_model_name)).

2. Take the best model and calculate total stem volume for each plot by applying the selected volume model (add the result as new column to your data frame).

3. Compare the modeled volume to the reference volume:

- Draw a figure where reference volume is in x-axis, and modeled volume in y-axis. (plot(DATASET$VOLUME_ORIGINAL, DATASET$VOLUME_MODELED))

Validating the selected model

1. Import the **validation data set** to R. This file contains another data set in Finland. The idea is to check, how well your newly fitted model works with different data set.
2. Calculate summary statistics also for this subset (min, max, arithmetic mean, standard deviation, total number of plots). Do this for all the ratio scale variables. Use this data also in the report "Material" chapter.
3. Add validation data set a new column, where you calculate stand volume estimate using the best model among your models (the one selected in step 6). Compare the new modeled stand volume with the original stand volume in the validation data. Draw scatter plot between the modeled and original volumes. (plot(DATASET_2$VOLUME_ORIGINAL, DATASET_2$VOLUME_MODELED))

4. Calculate absolute bias and absolute RMSE for the modeled stand volume in this data set (see ppt on validation in case you doubt)

Writing a report

Write a report (~5 pages) based on your work. Report should follow the normal scientific article style. Report should include the following chapters:

1**) Introduction:** Give here some introduction to the topic and explain the goals for the study

2) Materials: Report here the two data sets you used: modeling data set and the validation data set. Include their summary statistics as tables. Include also some figures.

3) Methods: Explain here, how you did the modeling work. Give here also the equations for calculating bias, RMSE etc.

4) Results: This chapter should include the main results of your study. Don't give any discussion/conclusions here, just tell the **main results.** Include here:

Modeling results:

- Details of the 3 best models you found (equations, significance of variables (p-values), R^2, relationships)
- Figures showing the model residuals (are they homoscedastic and are they normally distributed)
- Figures, where your independent variables are plotted against the dependent variable.
- Figures, where modeled volumes (your model) and original volumes in modeling data set are plotted against each other

Validation results:

- Bias and RMSE values in validation data set
- Scatter plot of the volume calculated using your volume model plotted against the volume available in the validation data set

5) Discussion (recommended): In this part you can comment all the problems you have identified and propose solutions for it.

Groups for the assignment

Students working together (max. 2) or individually. We will assign the data sets to each group/person.