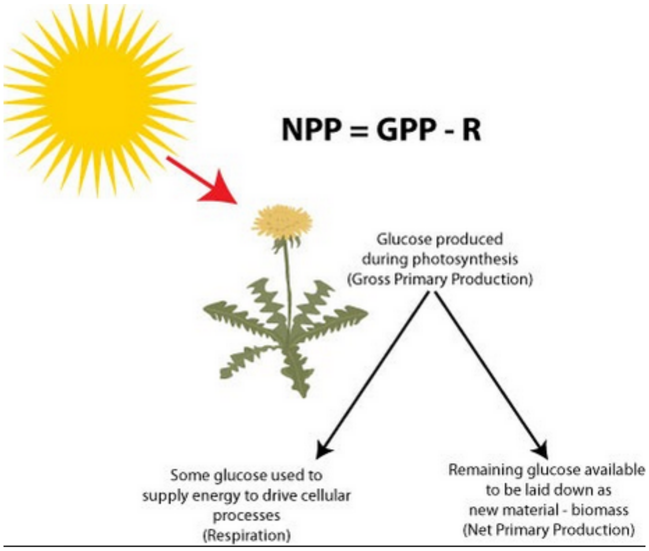
STA2007F Project 1: Regression
Factors driving vegetation productivity

Due Date: Friday 22nd March before 4:00 pm


NPP = GPP - R
Glucose produced during photosynthesis (Gross Primary Production)
Some glucose used to supply energy to drive cellular processes (Respiration)
Remaining glucose available to be laid down as new material - biomass (Net Primary Production)

Creative Commons License

The project is meant to be done in groups of 3 people and involves analysing the data described below and writing a short report (see the guidelines for Statistical report writing on Vula). Your analysis should include all the steps that have been covered in the course, starting with coming up with a set of questions/hypotheses that you will investigate, exploratory data analysis, model selection, model checking, interpretation and explanation of the final model chosen.

Please hand in a type-written hard copy at Stats reception, as well as an electronic version including a pdf file of your report and your R script file (via the Assignment tab on Vula). We require only one report per group. Decide on one person in your group to do the electronic submission on behalf of the group. *Please make sure that the names and student numbers of all contributing group members are clearly shown on the first page of the report on both the hard copy and the electronic version.* It is your responsibility to make sure that your name appears on the reports of your group.

The work that you will hand in must be your own. This means that you can discuss problems and difficulties or ideas with your tutor or an expert in the field. However, you must acknowledge any help received, and each group should do the analysis and write the report on its own. Please be sure to sign the plagiarism declaration (see the plagiarism document on Vula for more information).

The project uses data from a fictional study on land degradation where 163 grassy land parcels were randomly sampled from across South Africa. The aim is to better understand the factors driving vegetation productivity in South Africa. The data are on Vula in a file called "GPP.csv". The response variable is gross primary productivity (*GPP*) which is a measure of the amount of photosynthesis taking place and is measured in grams of carbon per m^2 per year. Other variables include:

Variable	Description	Codes/Values
`rainfall`	mean annual rainfall	53 - 918 (mm)
`temp`	mean summer temperature	18 - 25 (°C)
`livestock`	livestock units	3 - 47 (per ha)
`size`	size of land parcel	1302 - 3942 (ha)
`soil`	soil nutrient richness index	10 - 95
`seasonality`	predominant type of rainfall	Summer / Winter
`land.use`	land use classification	Communal grazing / Commercial livestock / Protected area

Please note the following:

- In addition to content, the layout and presentation of your report are important. Do NOT cut-and-paste R output into the report. Rather present R output in neatly constructed tables.
- Your project will NOT be marked if you do not include a plagiarism declaration.
- Although there is no formal page limit for this assignment, you should bear in mind that more does not mean better!
- The assignment must be typed except for formulae that may be added by hand if desired.

Hints

1. For Microsoft Word, save R graphs as metafiles, with .emf extension!! Everything else looks grainy and just not nice.
2. You can hand in a simple stapled report, with double-sided printing.
3. Harvard reference style: Give references in text, e.g. Erni et al. (2012) found that . . . (summarised in your own words), or: . . . (Erni et al. 2012).
4. All tables and figures should have a number and a caption and should be referred to in the main text.

5. Justify your hypotheses, either using findings and suggestions of other studies or your own reasons. Write out your hypotheses in words, e.g. the factors predicted to influence density were average winter temperature and habitat type.

6. The way to cite R:

   R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

   See citation() in R.

7. Steps in statistical modelling:

   (a) Which models will I fit and why?

   (b) Exploratory data analysis. Look for distributions of response variable, form of relationship with explanatory variables, multi-collinearity.

   (c) Using the exploratory data analysis, adapt my above models if required (but say so and why, i.e. do I need a transformation?)

   (d) Fit the above adapted models, and examine which models are best supported by the data.

   (e) Model checking: all models that do not meet all requirements can be discarded or adapted.

   (f) Present results, coefficients, plots, interpretations only for the best model (or 2 in some cases where it is hard to choose between 2 models for biological/practical/statistical reasons).

   (g) Interpret the final model. What does the model tell us about the response variable, is it a good/useful model, plots with fitted curves. Show that you can read the output from a statistical model and explain to somebody what it says about the data. Answer the questions initially asked. Which are the variables that had the greatest effect on the response? At this stage we are not interested in the significance, just how exactly are response and explanatory variables related?

8. Avoid naked p-values! Always add the size of the effect as well, with standard error! We want to know how and by how much the explanatory variables influence the response. The significance level just ensures that you don't make claims that could have been just due to random fluctuations that occurred purely by change.