BI5302 Dealing with variance heterogeneity exercise

Note: The first 6 questions in this practical are the same questions that you answered in the case study group work where you identified issues of residual heterogeneity of variance when fitting a standard linear model to these data. You can either add to your previous R markdown document (recommended) when completing this exercise or you can create a new R markdown document which only contains your answers to questions 7 onwards (don't forget though that you will need to import the `Hediste.txt` dataset again and perform any data transformations required.)

If you would prefer to work from the pdf version of this document you can find this here.

Benthic Biodiversity experiment

These data were obtained from a mesocosm experiment which aimed to examine the effect of benthic polychaete (*Hediste diversicolor*) biomass on sediment nutrient (NH_4, NO_3 and PO_3) release. At the start of the experiment replicate mesocosms were filled with homogenised marine sediment and assigned to one of five polychaete biomass treatments (0, 0.5, 1, 1.5, 2 g per mesocosm). Mesocosms were allowed to acclimatise for 24 h after which the concentration of either NH_4, NO_3 or PO_3 was measured in the water. The concentration of only one nutrient was measured in each mesocosm. The researchers were particularly interested in whether the nutrient concentration differed between polychaete biomass treatments and whether these effects were dependent on the nutrient type.

1. Create a new R markdown document in your BI5302 RStudio project and save it using a suitable file name. I suggest you specify the default output format as html but feel free to experiment with pdf (you can always change this later). Use this R markdown document to record your data exploration, statistical analysis (including graphs and tables) and commentary. For this exercise I would also suggest that you embed your R code as visible chunks within the document (use `echo = TRUE`) for later reference.

2. Import the `Hediste.txt` dataset into R and assign it to a suitably named variable. Remember if you're using R version > 4.0.0 (most of you will be) then columns containing character strings will be imported into R as character type variables not as factors by default. You can either use the argument `stringsAsFactors = TRUE` when you use the `read.table()` function to automatically convert character type variables to factors when you import your data or you can use the `read.table()` function without the `stringsAsFactors = TRUE` argument and then covert them after you import your data. Examine the structure of the dataframe and convert the `biomass` variable to a factors and store it as a new variable in your dataframe.

3. How many replicates are there for each biomass and nutrient combination?

4. Explore these data graphically. Are there any obvious outliers in the `concentration` variable for each of the `biomass` or `nutrient` variable levels (perhaps the `dotchart()` function with the `group` argument might help)? Use an appropriate plot to examine whether there are any biomass and/or nutrient effects on concentration (perhaps a boxplot?). Do you notice a potential issue regarding the between group variances?

5. With reference to the study aims stated above, fit an appropriate linear model to these data using the `lm()` function.

6. Use appropriate residual plots to identify whether the modelling assumptions are met. Don't forget to also plot the residuals from this model against all explanatory variables. Can you see a problem? Can you assume homogeneity of variance of the residuals from your model? If not, then try to identify the cause of this problem. Make sure you describe and discuss this process in your R markdown document.

7. Import the `nlme` package into R.

8. Use the `gls()` function from the `nlme` package without any variance covariates to refit your linear model specified above (the model you fitted using the `lm()` function). This GLS model with no variance structure is equivalent to a standard linear model. You will use this GLS model to compare with models you subsequently fit. If you are not convinced that these two models are equivalent then compare the output from both models (and replot the residuals if you're still not convinced!). Why do you need to refit this model using the `gls()` function? (Hint: take a look at the AICs from both models).

9. With reference to the residual plots you created above, fit models with different variance covariate structures to account for the heterogeneity of variance you identified previously.

You will first need to decide which variance structure to use (and which R function to use). I suggest that you start with a simple variance structure and then build up to a more complicated structure. Remember to use the `weights =` argument with the `gls()` function to specify your variance structure. Use the `AIC()` function to compare each of your models with different variance structures. Identify the model with the most appropriate variance covariate structure. Describe and justify your process in your R markdown document.

10. Now that your model has the optimal variance covariate structure it's time to perform model selection to identify the optimal fixed effects structure. To do this you will have to refit your optimal model (in terms of variance structure) using maximum likelihood estimation (ML) rather than the default restricted maximum likelihood estimation (REML). To do this use the argument `method = "ML"` with the `gls()` function.

11. Perform model selection using AIC to compare model fit. Select the model with the most appropriate fixed effects structure. Describe this process in your R markdown document. If you are feeling adventurous (this is optional!) you can also perform model selection by comparing nested models using likelihood ratio tests (using the `anova()` function). Does this result in the same final model?

12. Once you have your model with optimal variance covariate and fixed effects structures refit this model using REML. This is your final model.

13. Extract the normalised (variance scaled) residuals and fitted values from your final model. Plot the normalised residuals against the fitted values. Can you assume homogeneity of residual variance now? Also plot your normalised residuals against each of your explanatory variables? Comment on the model assumptions and contrast these with the naive linear model.

14. Now (and only now!) you can go ahead and interpret the the output from your final model (use the `anova()` and `summary()` functions). Summarise your interpretation in your R markdown document.

That's more or less it. Well done and congratulations!

[End of exercise]

If you want to work alongside my solutions you can find the relevant files below. Download the R markdown file and open it in your RStudio project.

- Dealing with variance heterogeneity R markdown file
- Dealing with variance heterogeneity final rendered html file
- Dealing with variance heterogeneity final rendered pdf file