BI5302 Dealing with temporal non-independence Practical

Alex Douglas

04 November, 2020

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.)

Environmental impacts on Hawaiian black-necked stilt abundance

These data were collected from bird surveys conducted on two Hawaiian islands (Maui and Oahu) from 1956 - 2003. The annual abundance of black-necked stilts (*Himantopus mexicanus knudseni*) was measured each winter using transect surveys on each island. Along with bird counts, annual rainfall data for the region was also obtained from the National Climate Data Center. The researchers were interested in understanding whether levels of rainfall impacted on bird abundance and whether any impact was different between the two islands.

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 `hawaii3.txt` dataset into R and assign it to a suitably named variable. Examine the structure of the dataframe. 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.

3. How many observations are there for each island?

4. Explore these data graphically. Are there any obvious outliers in the `abund` variable for each of the `locations` variable levels (perhaps the `dotchart()` function with the `group` argument might help)? Next, use an `xyplot` (from the `lattice` package) or a `coplot` to explore any relationships between bird abundance and rainfall for each of the two islands Finally, create a plot to examine how bird abundance changes over time (`year`) for each of the two islands.

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 (including `year`). 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 or correlation structures to refit your linear model specified above (the model you fitted using the `lm()` function). This GLS model is equivalent to a standard linear model. You will use this GLS model to compare with models you subsequently fit. Store your GLS model in an object with a suitable name (`birds_gls1` or similar).

9. OK, let's try to deal with the heterogeneity of variance issue we identified in Q6 above. Remember, it looks like the cause of our heterogeneity of variance was due to differences between location. Hopefully you remember how to deal with this situation using the `varIdent` variance covariate and the `weights =` argument when fitting a GLS model. Fit this model and call it something like `birds_gls2`.

10. Compare the `birds_gls1` and `birds_gls2` models using AIC to identify the 'best' model.

11. Extract the residuals using the `resid()` function from this model and plot against the the `location` variable to check whether this model has sorted out the issue of heterogeneity of variance between our two islands. Don't forget to use the argument `type = "normalized"` to extract the normalised residuals form the GLS model.

12. OK, now that we've dealt with the heterogeneity of variance issue, its time to look at whether the residuals are independent. Use the extracted residuals from Q11 and plot these over time (`year`)

for both of the island (`locations`). You can use either the `xyplot()` function from the `lattice` package or the `coplot()` function from base R.

13. So now that we've identified non-independence in our residuals we need to decide what type of correlation structure to use in our GLS model to account for this non-independence. One of the best ways of doing this is to use the `acf()` function to plot the autocorrelation function and the `pacf()` function to plot the partial autocorrelation function on our residuals.

14. Fit a new GLS model that incorporates both our variance covariate to deal with heterogeneity of variance and also a first order autoregressive correlation structure to deal with non-independence in our residuals at each site. Remember to include this AR(1) structure you will need to use the `corAR1` function and the `correlation =` argument. The form of the `corAR1` structure should include the `year` variable conditional (`|`) on the `location`. Assign your model to an appropriately named object (`birds_gls3`?).

15. Now that we've fitted our model let's once again extract our residuals and plot them against year for each location to see if the first order autoregressive correlation structure has dealt with our non independence. You should also re-plot the ACF and PACF plots.

16. Compare all of your fitted models so far using AIC. Which model has the most support?

17. So, now that we've dealt with the main issues of residual variance heterogeneity and non-independence it's time to turn our attention to performing model selection of our explanatory variables (remember this is the fixed part of the model). However, before we can perform model selection we must first refit our model using maximum likelihood (ML) rather than restricted maximum likelihood (REML) which is the default. To do this we refit the model using the `gls()` function once again but this time include the argument `method = "ML"`.

18. Now we can perform model selection using either AIC or by performing a likelihood ratio test (LRT). It's up to you how you want to perform this. If you want to automate this process you can use the `drop1()` function, but remember you need to use this sensibly and with care and don't forget about the biology! If you use the argument `test = "none"` with the `drop1()` function this will return the AIC values for each iteration. Conversely, if you use the argument `test = "Chisq"` this will return the LRT statistics.

19. Once you have determined your minimum adequate model, refit this model using REML once again (remember, this is the default for the `gls()` function).

20. Almost there! Now that we've determined our minimum adequate model we now need to revalidate the model using our usual residuals plots.

21. Now that we're happy with our model it's time to see what the model is telling us. Use the `summary()` function with our final model to obtain the parameter estimates. What is the estimate of the residual variance for Maui and Oahu? (hint: take a look at the `Variance function` section of the output - you will need to do a little maths!).

22. Once again looking at the output of the summary table what is the estimate of the correlation of residuals at lag 1, 2 and 3?

23. OK, let's finish this beast! Take a look at the parameter estimates and see if you can figure out what they are telling you. What does the `intercept` parameter represent? What do the `rainfall` and `locationOahu` parameters represent?

24. A picture paints a thousand words as they say, so finally, let's create a graph of our predicted values along with 95 % confidence intervals. There are many ways to do this but perhaps the easiest way to also contain the standard errors of the fitted values is to use the `predictSE()` function from the `AICcmodavg` package (you will need to install this). You use the `predictSE()` function in much the same way as you would use the standard `predict()` function. See `?predictSE()` for more details about this function.

[End of exercise]