

Week 8: Statistical Thinking

The data in the file `galapagos.csv` are counts of total observed plant species and the count of those species that are endemic, for 30 islands in the Galápagos (Johnson and Raven. 1973. *Science* 179:893–895). The columns are:

ISLAND Name of the island

TOTAL Number of observed plant species

NATIVE Number native (endemic) to the island

AREA Area of the island (km²)

ELEV Average elevation (m)

DISTNEAR Distance to nearest island (km)

DISTSC Distance to Santa Cruz island (km)

AREANEAR Area of nearest island (km²)

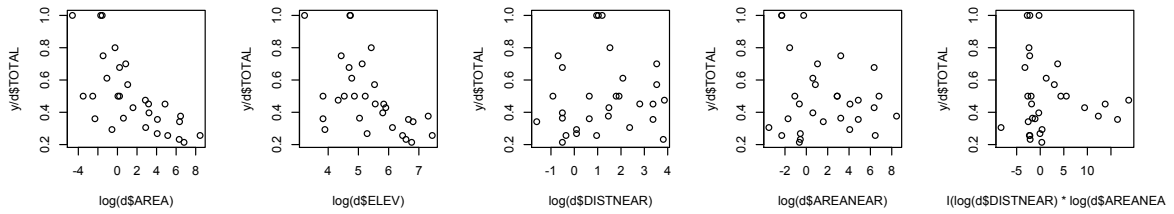
I want you to model the counts in **NATIVE**. Consider three hypotheses. (1) **NATIVE** depends positively upon the **AREA** of each island. Larger islands have more endemic species. (2) **NATIVE** depends positively upon the **ELEV** of each island. Higher islands have more endemic species. (3) **NATIVE** will be lower for islands with close, large neighboring islands. When you have a large neighbor that is nearby, you end up with invading species that reduce the number of endemics, the theory goes. You will need to consider the logs of **AREA** and **ELEV** and **DISTNEAR** and **AREANEAR**, because these variables are have exponential-like distributions.

I want you to report you analyses in the form of the six steps of statistical thinking:

1. Specify the models. State the hypothesis that corresponds to each model. These are casual mathematical models, like $y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma)$ or $y_i \sim \text{Nbinom}(\alpha \exp(\beta x_i), \exp(n))$.
2. Simulate data from the models. Display the simulated data using the code below:

```
par(mfrow=c(1,5))
plot( y/d$TOTAL ~ log(d$AREA) )
plot( y/d$TOTAL ~ log(d$ELEV) )
plot( y/d$TOTAL ~ log(d$DISTNEAR) )
plot( y/d$TOTAL ~ log(d$AREANEAR) )
plot( y/d$TOTAL ~ I(log(d$DISTNEAR)*log(d$AREANEAR)) )
```

where `y` contains your simulated counts of **NATIVE**. This will produce a row of scatter plots like this (in this case for the real data):



If you are having trouble finding good parameters for the simulations, then skip ahead to fitting, and then come back once you have reasonable values for the parameters.

3. Specify likelihood functions for the models, in the form of R code.
4. Validate your code, by recovering known parameters from simulated data (step 2).
5. Fit your models to the actual data.
6. Compare the models and interpret their estimates.

PLEASE work in groups. Feel free to turn in joint reports, with everyone's name on it. You can work alone, if you prefer, but there are benefits to working in teams and dividing up the six jobs.