M.Sc. Practical MT2005 week 7

Initial Feedback

The data are taken from

McMaster, R. T. (2005) Factors influencing vascular plant diversity on 22 islands off the coast of eastern North America, *Journal of Biogeography* **32**, 475–492.

who used a least-squares regression analysis of the numbers of native and of non-native species. It provides extensive background information, but poorly informed statistical analysis.

Reading in the data

A number of you had trouble with this, despite this being a topic covered in Chapter 2 of your S-PLUS notes. Do read carefully the options for `read.table` and make sure you select them correctly. Here the columns are tab-separated (you were told this), and missing values are indicated by `-`. I used

```
islands <- read.table("islands.dat", sep="\t", na.strings="-")
names(islands) <- c("total", "native", "nonnative", "perc", "area",
  "latitude", "elevation", "mainland", "nearest", "soil.types",
  "isolation", "deglaciation", "popln", "popln.density")
```

Note that the only missing values are soil type for 4 islands (the most northerly), so these are not missing at random. A quick look at the map would show that these are off the coast of Canada and marked (NB) (for New Brunswick), and in fact that is the reason the data are missing: the US has not surveyed them. Only one of you spotted that they were clustered and outside the main body of the data.

Visualization

We are principally interested in `total` and `100 - perc`:

```
pairs(islands[, -c(2,3)])
```

now consider a smooth curve for continuous variables (population is mainly zero)

```
splom(~islands[, -c(2,3,10,13,14)], panel=function(...) {
      panel.splom(...); panel.loess(...)})
```

and some individual regressions

```
nm <- names(islands)
attach(islands)
par(mfrow=c(3, 3))
for(i in c(5:9, 11:12)) {
  x <- get(nm[i])
```

```
  scatter.smooth(x, total, xlab=nm[i])
  abline(lm(total ~ x), lty=2)
}
par(mfrow=c(3, 3))
prop <- 1 - perc/100
for(i in c(5:9, 11:12)) {
  x <- get(nm[i])
  scatter.smooth(x, prop, xlab=nm[i])
  abline(lm(prop ~ x), lty=2)
}
```

Remember these are just univariate regressions and fitted by least-squares, which is not appropriate. But they do indicate that `total` varies strongly and non-linearly with `area` and `elevation`, and at most linearly with the remaining variables. For the proportion there is possibly a non-linear effect of elevation, and otherwise perhaps linear effects. (As the proportions are not far from 0.5, least-squares is not badly wrong.)

Note that some of the explanatory variables are quite highly related, for example time since deglaciation being related to latitude.

Missing values

We have to do something about the missing values. We have two choices, to drop the Canadian islands or to predict values from the remaining variables. I decided to first exclude `soil.types` and then see if it had an effect on the predictions for the non-Canadian islands.

Those of you who did try predicting came up with a wide range of predictions, in fact from 0 to 80 for one of the islands. (Surely neither value is sensible, and although 0 was changed to 1 it indicates the model was not sensible.) Since there must be at least one soil type, a Poisson model is not suitable.

Question (a)

explaining the numbers of species. All of you interpreted this as the total number, but that is not the only way to read the question.

For the total, we must exclude the numbers of native and non-native species, and also the percentage of non-native species since these are not independent variables.

We first need to tackle the non-linear relationship with area.

```
options(digits=5, width=60)
isA <- islands[, -c(2:4,10)]
attach(isA)
plot(total ~ area)
plot(total ~ sqrt(area))
plot(total ~ log(area))
```

suggest $\log$ is about the right transformation. Note that this says that the mean number of species varies linearly with log area.

Since these are counts, the obvious model is a Poisson regression, which by default is a model for log means (and not log counts, as several of you had). However, plotting `log(total)` against transformations of `total` still suggests `log(area)` as reasonable. It is not obvious that a multiplicative model is optimal here, especially as we have the sum of two types of species introduced by different mechanisms. It is also not obvious that the independence assumptions behind a Poisson model are satisfied, as species might arrive together, or be a host and a parasite for example. So one could consider exploring other models such as a negative binomial glm (see the Statistical Methods exercises).

```
fit <- glm(total ~ log(area) - area + ., isA, family=poisson)
summary(fit, cor=F)
fit1 <- glm(total ~ log(area) - area + ., isA, family=poisson(identity))
summary(fit1, cor=F)
```

This does produce a large reduction in the residual deviance, but only to 515 on 12 degrees of freedom, which is a lot of over-dispersion. We can also consider a linear model which fits about equally well, but selects different variables. (You may want to replace elevation by log elevation.)

We need to be careful assessing significance in such over-dispersed models. The t-ratios are a reasonable guide, and only elevation is non-significant. How does this accord with our preliminary investigations?: for all but one island elevation is well-predicted by log(area).

Given the separate biological mechanisms which are expected for the presence of native and non-native species (the principal agents for the latter being man and the animals he introduced) it would make sense to develop models for the two separately and then sum their predictions.

Many of you fitted over-elaborate models. ‘Interactions’ between continuous variates are products, and models with 16 parameters for 18 observations are not throwing any light on biogeography.

Question (b)

explaining the proportion of native species. Almost all of you fitted a Poisson regression with offset. This makes little sense, as the total has just been treated as random and the maximum possible value is total. However, if we condition on the total we get a binomial model, but not one with a logit link. A simple logistic regression is all that was needed.