Practical One

Dr. Monique L. Mackenzie

January 2018

Introduction

In this practical, we will examine baseline and impact assessment data from one of the worlds largest off-shore wind farms (Horns Rev wind farm[1]). The data used here relates to the abundance and distribution of a large sea duck (Common Scoter[2]).

Research questions

We will address the following research questions in this practical by exploring the data and fitting models in `R`:

1. Do there appear to be differences in abundance before and after impact?
2. Of the covariates available, what are the best predictors of abundance?
3. Is there any evidence of re-distribution before and after impact?

Data description

As for the Nysted data, the data are collected along tracklines (transects) from the ocean surface using aerial survey methods. These tracks are followed from the air by plane and the number of animals at each location (on or near the track lines) are recorded and the recorded counts are inflated for the fact that not all animals at the surface are seen. This correction for the imperfect detection process was carried out using Distance Sampling[3]. The data has the following covariates:

```
> head(data)
     x.pos   y.pos year month day Depth TNO survey transect     area Nhat yearmonth impact
1 392441.6 6182144 2005    11  19 26.76 101      1        1 0.713176    0    200511      0
2 392405.3 6181709 2005    11  19 26.94 101      1        1 0.956000    0    200511      0
3 392397.7 6181209 2005    11  19 27.09 101      1        1 0.956000    0    200511      0
4 392409.9 6180709 2005    11  19 27.19 101      1        1 0.956000    0    200511      0
5 392433.2 6180210 2005    11  19 27.24 101      1        1 0.956000    0    200511      0
6 392445.2 6179710 2005    11  19 27.26 101      1        1 0.956000    0    200511      0
```

- `x.pos`: the x-coordinate of the animal(s)
- `y.pos`: the y-coordinate of the animal(s)
- `year`: the year of the survey
- `month`: the month of the survey
- `day`: the day of the survey
- `Depth`: the water depth (m)
- `TNO`: transect number (this is a unique identifier across surveys)
- `survey`: survey number
- `transect`: transect number (this is not unique across surveys)
- `area`: the area associated with each count
- `Nhat`: the estimated abundance at each location

[1] http://en.wikipedia.org/wiki/Horns_Rev
[2] http://en.wikipedia.org/wiki/Common_Scoter
[3] http://en.wikipedia.org/wiki/Distance_sampling

- `yearmonth`: an aggregate covariate for year and month
- `impact`: an impact related covariate (pre-impact=0, post-impact=1)

In this practical,however, we will only use `x.pos`, `y.pos`, `Depth`, `area`, `Nhat` and `impact`. The following sections contain some instructions and there are Moodle-based questions associated with each task which will form part of your practical-based assessment.

Exploratory Analysis

1. Load the data set (which can be found in Moodle) into R.
2. Plot histograms and boxplots for the estimated abundance per unit area pre and post impact.
3. Calculate confidence intervals for the estimated abundance per unit area before and after impact using three methods:
   a. Assume the data are Normal: i.e. use $\hat{\mu} \pm t_{0.025,df=n-1} \times se(\hat{\mu})$ where $\hat{\mu}$ represents the average estimated abundance per unit area and $se(\hat{\mu}) = \frac{s}{\sqrt{n}}$.
   b. Assume the data are Poisson. This still uses $\hat{\mu} \pm z_{0.025} \times se(\hat{\mu})$ but has $se(\hat{\mu}) = \sqrt{\frac{\hat{\mu}}{n}}$
   c. Rather than assume a distribution for the sample estimates, use a non-parametric bootstrap percentile-based interval. If you would like to revisit bootstrap resampling, watch the short video by Ben Lambert: `https://www.youtube.com/watch?v=5nM5e2_1OQ0`

Use the following code for the pre-impact state, but you will need to modify this for the post-impact state:

```
results<- matrix(0, nrow=1000, ncol=1)
for(j in 1:1000){

rowsToUse<-sort(sample(which(impact==0),
length(which(impact==0)), replace=T))
results[j,]<- mean(Nhat[rowsToUse]/area[rowsToUse])
}
cisBoot1<-quantile(results, probs=c(0.025,0.975))
```

4. Plot the data in x/y space (pooled together and separately for each impact code) using:

```
> par(mfrow=c(1,1))
> plot(x.pos, y.pos, xlab="X-coordinate", ylab="Y-coordinate",
main="Transect lines", pch=20, col="lightgrey")
> points(x.pos, y.pos, pch=20, col="blue", cex=log(Nhat+1))


> par(mfrow=c(1,2))
> plot(x.pos[impact==0], y.pos[impact==0], xlab="X-coordinate",
ylab="Y-coordinate", main="Pre impact", pch=20, col="lightgrey")
> points(x.pos[impact==0], y.pos[impact==0], pch=1, col="blue",
cex=log(Nhat[impact==0]+1))
> plot(x.pos[impact==1], y.pos[impact==1], xlab="X-coordinate",
ylab="Y-coordinate", main="Post impact", pch=20, col="lightgrey")
> points(x.pos[impact==1], y.pos[impact==1], pch=1,
col="blue", cex=log(Nhat[impact==0]+1))
```

Model Fitting

Fit Poisson-based GLMs with and without a dispersion parameter estimate (each with an offset), for models with estimated abundance per unit area as the response and:

1. the `impact` covariate alone
2. `impact`, `Depth`, `x.pos` and `y.pos` covariates
3. `impact`, `Depth`, `x.pos`, `y.pos` and interaction terms between `impact` and each of `x.pos` and `y.pos`

For example, models fitted to `impact` alone could be fitted using:

```
> glmFit1<- glm(Nhat ~ impact, data=dat, offset=log(area), family=poisson)
> Anova(glmFit1)

> glmFitOD1<- glm(Nhat ~ impact, data=dat, offset=log(area), family=quasipoisson)
> Anova(glmFitOD1, test="F")
```

Model Selection

1. Use the `dredge` function in the `MuMIn` library to carry out all possible subsets based selection to choose a model from the full model (including the covariates listed above and the interaction terms listed above). Note, you'll need to change your options for missing values, using:

```
> options(na.action = "na.fail")
```

2. For comparison, carry out stepwise (both directions) AIC-based selection for the Poisson-based model (not quasi-Poisson since these models do not have an AIC).

Results

Questions

1. Which of the following about the pre and post impact distribution of the estimated abundance per unit area is TRUE?
   - The data are heavily right-skewed both pre and post impact and the range of post-impact values was larger than the range seen pre-impact.
   - The data are heavily left-skewed both pre and post impact and the range of post-impact values was smaller than the range seen pre-impact.
   - The data are reasonably symmetrical both pre and post impact and the range of post-impact values was larger than the range seen pre-impact.
   - The data are heavily right-skewed both pre and post impact and the range of post-impact values was smaller than the range seen pre-impact.
   - The data are heavily left-skewed both pre and post impact and the range of post-impact values was larger than the range seen pre-impact.
2. What is the lower 95% confidence limit for the Normal-based pre-impact interval? Please quote your answer to 2 decimal places and round the second decimal place up if it is 5 or more, and down if it is 4 or less. 48.08
3. What is the upper 95% confidence limit for the Normal-based pre-impact interval? Please quote your answer to 2 decimal places and round the second decimal place up if it is 5 or more, and down if it is 4 or less. 55.04
4. What is the lower 95% confidence limit for the Poisson-based pre-impact interval? Please quote your answer to 2 decimal places and round the second decimal place up if it is 5 or more, and down if it is 4 or less. 51.44
5. What is the upper 95% confidence limit for the Poisson-based pre-impact interval? Please quote your answer to 2 decimal places and round the second decimal place up if it is 5 or more, and down if it is 4 or less. 51.67
6. What is your lower 95% confidence limit for the bootstrap-based pre-impact interval? Please quote your answer to 2 decimal places and round the second decimal place up if it is 5 or more, and down if it is 4 or less. 48.14
7. What is your upper 95% confidence limit for the bootstrap-based pre-impact interval? Please quote your answer to 2 decimal places and round the second decimal place up if it is 5 or more, and down if it is 4 or less. 54.81
8. Which of the following about the 95% confidence intervals you have created is FALSE?
   - The Poisson-based confidence intervals should always be used because the data are estimated abundances per unit area. count data 更适合用 p o i s s o n 建模
   - Either the Normal-based or bootstrap based confidence intervals can be used in this case because their results are similar.
   - The bootstrap based confidence intervals demonstrate the mean estimated abundance per unit area both pre and post impact has an approximately Normal distribution, despite the skewness in the parent population.
   - The Poisson-based confidence intervals would continue to be very different to the Normal based intervals even if the sample size was larger.
   - When there is a large discrepancy between the Normal-based and bootstrap based confidence intervals, it is almost always wise to use the bootstrap based intervals because they do not assume the distribution of the estimates is Normal.
9. Which of the following about the distribution of the birds in the survey area is FALSE?
   - The data appear to be more widely dispersed across the survey area post-impact compared with pre-impact.

- There are very few birds seen for low values of the X-coordinate pre impact, but many birds post-impact in the same area and so there may be some redistribution in the X-coordinate post impact.
- There are very few birds seen in the low values of the Y-coordinate surveyed pre impact, but many birds post-impact in the same area and so there may be some redistribution in the Y-coordinate post impact.
- There are very few birds seen for low values of the X-coordinate pre impact, but many post-impact in the same area and so using this graphic alone we can conclude there is redistribution in the x-range post impact.
- There is a large aggregation of birds associated with central values of the Y co-ordinates and low to mid values of the X-coordinate post impact.

10. Which of the following about the Poisson and Quasi-Poisson based models fitted with `impact` as the sole covariate is TRUE?

    - There is strong evidence for an increase in the average estimated abundance per unit area post-impact, compared with pre-impact, regardless of whether the dispersion parameter is estimated or assumed to be equal to one.
    - There is strong evidence for a decrease in the average estimated abundance per unit area post-impact, compared with pre-impact, regardless of whether the dispersion parameter is estimated or assumed to be equal to one.
    - There is strong evidence for an increase in the average estimated abundance per unit area post-impact, compared with pre-impact, when the dispersion parameter is assumed to be equal to one but evidence for a difference disappears once the dispersion parameter is estimated because it is so large.
    - The dispersion parameter for this model is estimated to be close to one.
    - There is strong evidence for a decrease in the average estimated abundance per unit area post-impact, compared with pre-impact, when the dispersion parameter is assumed to be equal to one but evidence for a difference disappears once the dispersion parameter is estimated because it is so large.

系数太大模型不可用

11. Which of the following about the Poisson and Quasi-Poisson based models fitted with `impact`, `Depth`, `x.pos` and `y.pos` as model covariates (but without any interactions) is TRUE?

    - The `y.pos` covariate is no longer statistically significant at the 5% level when the dispersion parameter is estimated.
    - The `x.pos` covariate is no longer statistically significant at the 5% level when the dispersion parameter is estimated.
    - The `x.pos` covariate is statistically significant at the 1% level when the dispersion parameter is estimated.
    - The dispersion parameter is estimated to be small and so the results (regarding the statistical significance of model predictors) are identical when it is assumed to be one or estimated as a part of the model.
    - The parameter estimates, and the fitted values, are noticeably different when the dispersion parameter is estimated because the dispersion parameter estimate is so large and they are adjusted by this value.

12. Which of the following about the Poisson and Quasi-Poisson based models fitted with `impact`, `Depth`, `x.pos`, `y.pos` and interaction terms between `impact` and each of `x.pos` and `y.pos` as model covariates is TRUE?

    - Based on interpreting p-values at the 5% level, there appears to be significant change in the relationship between the estimated abundance per unit area and the X-coordinate pre and post impact, but not in the Y-coordinate.
    - Based on interpreting p-values at the 5% level, there appears to be significant change in the relationship between the estimated abundance per unit area and both the X and Y-coordinates pre and post impact.

- Based on interpreting p-values at the 1% level, there appears to be significant change in the relationship between the estimated abundance per unit area and both the X and Y-coordinates pre and post impact.
- Based on interpreting p-values at the 5% level, there appears to be significant change in the relationship between the estimated abundance per unit area and the Y-coordinate pre and post impact, but not in the X-coordinate.
- The results are identical regardless of whether the dispersion parameter is estimated or assumed to be equal to one.

13. Based on the Quasi-Poisson based model fitted with `impact`, `Depth`, `x.pos`, `y.pos` and interaction terms between `impact` and each of `x.pos` and `y.pos` as model covariates, which of the following is FALSE?
    - There are three models with non-zero weights.
    - The model without the Y-coordinate and the impact/Y-coordinate interaction has the model with the second highest weight.
    - The full model (including all interactions) is the model with the highest weight.
    - A model averaging approach could be used here. This would involve averaging the predictions from the models with non-zero weight to give an average model prediction (weighted by their model weights) for each row in the data set.
    - The model with the third largest weight has the X-impact interaction term omitted from the full model.
14. Is the following statement TRUE or FALSE?

The 'best' model is the same if the all-possible subsets or stepwise selection is used.

15. Is the following statement TRUE or FALSE?

There is evidence of a change in the average estimated abundance per unit area, pre and post impact.

16. Is the following statement TRUE or FALSE?

Of the covariates (i.e. main terms) trialled, all appear to have a genuine (non-zero) relationship with the response.

17. Is the following statement TRUE or FALSE?

There appears to be evidence of animal re-distribution in both the X and Y coordinate directions after impact.