MT5764: Advanced Data Analysis
Practical One

Rachel Sippy

Due Tuesday 1st February 2022

Introduction

In this practical, we will examine impact assessment data from an offshore wind farm Horns Rev. The data used here relates to the abundance and distribution of a large sea duck, common scoter (*Melanitta nigra*).

Note: Don't forget to answer the questions at the end of this practical (and every practical in the future) on Moodle.

Research questions

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

1. Are there any differences in average abundance before and after impact?
2. Of the covariates available, what are the best predictors of abundance?
3. Is there any evidence of redistribution before and after impact?

Data description

As for Nysted, 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. The observed counts are adjusted 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. The data has the following covariates:

```
# Libraries
library(tidyverse) # ggplot()

# Read dataset
df <- read.csv("HornsRev.csv")

# Set Impact as factor
df$Impact <- as.factor(df$Impact)
```

```
dim(df)
```

```
[1] 27854    11
```

```
head(df)
```

```
      XPos    YPos Year Month Day Depth TransectID     Area Nhat YearMonth
1 392441.6 6182144 2005    11  19 26.76        101 0.713176    0    200511
2 392405.3 6181709 2005    11  19 26.94        101 0.956000    0    200511
3 392397.7 6181209 2005    11  19 27.09        101 0.956000    0    200511
4 392409.9 6180709 2005    11  19 27.19        101 0.956000    0    200511
5 392433.2 6180210 2005    11  19 27.24        101 0.956000    0    200511
6 392445.2 6179710 2005    11  19 27.26        101 0.956000    0    200511
  Impact
1      0
2      0
3      0
4      0
5      0
6      0
```

- `Nhat` (response variable): *estimated* duck abundance using distance sampling.
- `Impact`: before (0)/after (1) wind farm was built.
- `Area`: observation area for each count (km^2).
- `XPos/YPos`: spatial location (UTM projection; metres).
- `Depth`: sea depth at spatial location (metres).
- `Day/Month/Year`: collection date.
- `TransectID`: unique transect ID.

Exploratory analysis

1. Load the `HornsRev.csv` dataset (which can be found on Moodle) into R.
2. Plot histograms and boxplots for the estimated abundance **per unit area** pre- and post-impact.
3. Compute and plot the 98% confidence intervals (i.e. $\alpha = 0.02$) for the mean estimated abundance per unit area $\hat{\mu}$ before and after impact using four methods:

   a. Assume response is normally distributed.

   $$\hat{\mu} \pm t_{\alpha/2,n-1} \times se(\hat{\mu})$$
   $$se(\hat{\mu}) = \frac{s}{\sqrt{n}}$$

   b. Use the Normal approximation for large samples, assuming response is Poisson.

   $$\hat{\mu} \pm z_{\alpha/2} \times se(\hat{\mu})$$
   $$se(\hat{\mu}) = \sqrt{\frac{\hat{\mu}}{n}}$$

   c. Use the Normal approximation for large samples, without assuming response comes from a specific distribution. Recall that the central limit theorem tells us that for large sample sizes, the sampling distribution of the mean tends to a Normal distribution *regardless* of the population distribution.

   $$\hat{\mu} \pm z_{\alpha/2} \times se(\hat{\mu})$$
   $$se(\hat{\mu}) = \frac{s}{\sqrt{n}}$$

d. Use **bootstrapping**, a non-parametric resampling/simulation method for estimating the sampling distribution of any quantity of interest. Once the sampling distribution is generated, the empirical quantiles are used to approximate the confidence intervals. If you would like to revisit bootstrap resampling, watch this short video by Ben Lambert.

Use the following bootstrap code to get you started:

```
set.seed(145) # set this to reproduce results
NBOOT <- 1000 # no. of bootstrap samples
alpha <- 0.02 # alpha level of confidence

# For pre-imact (Impact=0)
dfPre <- subset(df, Impact==0)
muHat <- sapply(seq(NBOOT),
                function(x) mean(sample(x=dfPre$Nhat/dfPre$Area,
                                        size=nrow(dfPre),
                                        replace=TRUE)))
CI <- quantile(muHat, c(alpha/2,  1-(alpha/2)))
```

4. Plot the abundance per unit area spatially (i.e `XPos/YPos` space), scaling the size of each data point by their count. Show also the transect points.

    a. Pool all the data together.
    b. Split by impact.

Model fitting

Fit Poisson-based GLMs with (`quasi-poisson`) and without (`poisson`) a dispersion parameter estimate (each with an offset), produce a summary of the results and use the `Anova` function from the `car` package (i.e. add `library(car)` to your script) to get a handle on the importance of each predictor. Remember to use the right statistical test depending on whether the dispersion parameter ϕ is known or not. Consider the following covariates for your models:

1. `Impact` only.
2. `Impact`, `Depth`, `XPos` and `YPos`.
3. `Impact`, `Depth`, `XPos`, `YPos` and interaction terms between `Impact` and each of `XPos` and `YPos`.

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 98% confidence limit for the average abundance per unit area when assuming the data to be normally distributed for the **pre**-impact case. Please quote your answer to 2 decimal places.

3. What is the upper 98% confidence limit for the average abundance per unit area when assuming the data to be normally distributed for the **pre**-impact case. Please quote your answer to 2 decimal places.

4. What is the lower 98% confidence limit for the average abundance per unit area when using the Normal approximation for large samples, assuming response is Poisson, for the **pre**-impact case. Please quote your answer to 2 decimal places.

5. What is the upper 98% confidence limit for the average abundance per unit area when using the Normal approximation for large samples, assuming response is Poisson, for the **pre**-impact case. Please quote your answer to 2 decimal places.

6. What is the lower 98% confidence limit for the average abundance per unit area when using the Normal approximation for large samples, for an unknown population distribution, for the **post**-impact case. Please quote your answer to 2 decimal places.

7. What is the upper 98% confidence limit for the average abundance per unit area when using the Normal approximation for large samples, for an unknown population distribution, for the **post**-impact case. Please quote your answer to 2 decimal places.

8. What is the lower 98% confidence limit for the average abundance per unit area when using bootstrapping for the **post**-impact case (remember to `set.seed(145)` to reproduce the results). Please quote your answer to 2 decimal places.

9. What is the upper 98% confidence limit for the average abundance per unit area when using bootstrapping for the **post**-impact case (remember to `set.seed(145)` to reproduce the results). Please quote your answer to 2 decimal places.

10. Which of the following about the 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.
    - Assuming data is normally distributed, using the normal approximation for an unknown population distribution or bootstrap based confidence intervals can be used in this case because their results are similar.
    - The bootstrap based confidence intervals demonstrate that the sampling distribution for 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 other intervals even if the sample size was larger.
- When there is a large discrepancy between the normal approximation for an unknown population and bootstrap based confidence intervals, it is almost always wise to use the bootstrap based intervals because they do not assume the sampling distribution to be Normal.

11. 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.

12. Which of the following about the Poisson and quasi-Poisson based models fitted with `Impact` as the sole covariate is TRUE?

    a. 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.
    b. 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.
    c. 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.
    d. The dispersion parameter for this model is estimated to be close to one.
    e. 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.

13. Which of the following about the Poisson and quasi-Poisson based models fitted with `Impact`, `Depth`, `XPos` and `YPos` as model covariates (but without any interactions) is TRUE?

    - The `YPos` covariate is no longer statistically significant at the 5% level when the dispersion parameter is estimated.
    - The `XPos` covariate is no longer statistically significant at the 5% level when the dispersion parameter is estimated.
    - The `XPos` 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.

14. Which of the following about the quasi-Poisson based model fitted with `Impact`, `Depth`, `XPos`, `YPos` and interaction terms between `Impact` and each of `XPos` and `YPos` as model covariates is TRUE?

    - Based on interpreting p-values at the 1% level, there appears to be a 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 a change in the model-based conclusions about the relationship between the estimated abundance per unit area in the Y-coordinate but not the X-coordinate pre and post impact.
    - Based on interpreting p-values at the 1% level, there appears to be a 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 p-value based conclusions (at the 5% level) are identical regardless of whether the dispersion parameter is estimated or assumed to be equal to one.

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.