Estadística Aplicada I

Actividad 6: Modelos de ANOVA complejos

Prof. Edlin Guerra Castro, Prof. María Muciño

Parte A. ANOVA dos factores ortogonales

Asumamos que se efectuó un estudio de impacto ambiental con un diseño BACI (**B**efore-**A**fter x **C**ontrol-**I**mpact), en el que se midió la DBO tres meses antes de operaciones de la planta y luego a los tres meses de iniciadas las operaciones. El modelo acá es que la puesta en operación de la planta promueve el incremento de la DBO del agua. No obstante, las mediciones se efectuaron en *dos* localidades: Loc1 (río arriba) y Loc2 (justo en las inmediaciones de la descarga de la planta). El modelo espacial implica que la DBO del agua en Loc2 debe ser mayor que en Loc1 debido a la cercanía con las descargas de la planta. Note, sin embargo, que esto debe ocurrir sólo después de iniciadas las operaciones de la planta, ya que si antes la Loc2 ya presentaba mayores niveles de DBO que Loc1, no se puede asumir que la planta incrementó la DBO del agua en Loc2. Inspeccione los datos (archivo **datos_6.csv**) y responda:

1. Identifica todas las fuentes de variación y define las hipótesis estadísticas de cada una. Para ello defina la naturaleza del factor (fijo o aleatorio)
2. Elabora dibujos de los posibles resultados combinando cambios solo temporales, solo espaciales, e interacciones. Discuta qué significaría cada una en el contexto del estudio. ¿Cuál de los posibles resultados es un indicador de impacto ambiental?
3. Efectúe un ANOVA bifactorial usando el paquete `GAD`. Este es uno de los pocos paquetes en R que calcula adecuadamente la descomposición de la variación considerando la naturaleza de los factores (use la función `as.fixed` o `as.random`). Primero, cargue el paquete y luego defina la naturaleza del factor. Luego ajuste un modelo lineal (llámelo mod1) que considere cada uno de los términos principales, así como la interacción de ambos, escribiendo alguno de los siguientes códigos.

```
library(GAD)
BA <- as.fixed(datos$BA)
CI <- as.fixed(datos$CI)
DBO <-datos$DBO
mod1<-lm(DBO~BA*CI)
```

4. Calcula los grados de libertad de los distintos términos en el modelo.
5. Obtén los valores de los MS estimados (EMS) de la tabla de ANOVA usando la función `estimates` aplicada a mod1. ¿Cuáles son los términos del modelo que deberán ser usados en el numerador y denominador para probar cada una de las Hipótesis propuestas?
6. Obtén los resultados del análisis usando las funciones `anova` y `gad` sobre el objeto mod1. Corrobore que los g.l. del numeral 4 fueron correctamente calculados.
7. Examina la tabla de ANOVA obtenida y responde a las siguientes preguntas:
    a) ¿Es significativa la interacción? ¿Cómo se interpreta este resultado?
    b) ¿Es significativo alguno de los términos principales en el modelo?¿qué significaría esto?
    c) ¿Con esta información es suficiente para sustentar el impacto ambiental, o faltan más procedimientos estadísticos? Si piensas que falta algo, ¿qué faltaría?
8. Compruebe cuál de las tablas ANOVA es correcta estimando usted mismo los grados de libertad, cuadrados medios esperados y razones de F según el protocolo de Underwood (1997) de la guía suministrada.
9. Elabore un gráfico de interacción **BACI** con promedios y desviaciones estándar usando `ggplot2` (vas a requerir estimar promedios y desviación estándar para cada grupo para generar este gráfico)
10. Como procedimiento de control de calidad de los resultados, verifica que se cumplieron los supuestos de normalidad y homogeneidad de las varianzas.

Parte B. ANOVA tres factores para detección de impacto ambiental

10. Imagina que Loc1 y Loc2 son en realidad localidades bien extensas, por lo que las muestras de agua (25 en cada muestreo) en realidad provienen de combinar cinco muestras de agua de cinco zonas aleatorias de cada localidad. El objetivo de ello era tener una mejor representatividad de cada localidad.
11. ¿Cuáles serían los pasos que se verían modificados en el procedimiento que acabas de aplicar?
12. ¿Cuántas fuentes de variación hay ahora? Lleva a cabo los cambios que has propuesto para el modelo y grábalo con el nombre mod2.

```
BA<-as.fixed(datos$BA)
CI<-as.fixed(datos$CI)
Zona<-as.random(datos$Zona)
DBO <- datos$DBO
mod2<-lm(DBO ~ BA + CI + BA*CI + Zona%in%(BA*CI))
```

13. Obtén los MS estimados y la tabla de ANOVA empleando las funciones `anova` y luego `gad`, y responde a las siguientes preguntas:
14. ¿Qué diferencias hay en las tablas de ANOVA generadas por cada función?
15. Compruebe cuál de las tablas ANOVA es correcta estimando usted mismo los grados de libertad, cuadrados medios esperados y razones de F según el protocolo de Underwood (1997) de la guía.
16. ¿Cuál es la interpretación estadística del nuevo resultado? ¿Cómo sería el gráfico ideal?
17. ¿Sería este resultado evidencia de impacto ambiental?
18. Como procedimiento de control de calidad de los resultados, verifica que se cumplieron los supuestos de normalidad y homogeneidad de las varianzas.

Parte C. ANOVA completamente jerárquico

Considere un estudio hecho por Anderson et al. (2005a) sobre la variabilidad espacial en ensamblajes de invertebrados colonizando rizoides del kelp *Ecklonia radiata* en el noreste de Nueva Zelanda. Se utilizó un diseño de muestreo jerárquico. Buzos recolectaron n = 5 rizoides individuales (separados por metros) dentro de cada una de 2 áreas (separadas por decenas de metros) dentro de cada uno de 2 sitios (separados por cientos de metros a kilómetros), dentro de cada una de 4 localidades (separadas por cientos de kilómetros) a lo largo de la costa. El diseño fue balanceado y totalmente anidado con tres factores:

- Factor A: Localidades (aleatorio con a = 4 niveles)
- Factor B: Sitios (aleatorio con b = 2 niveles, anidado en Localidades)
- Factor C: Áreas (aleatorio con c = 2 niveles, anidado en Sitios y Localidades)

Los datos de este ejemplo están localizados en el archivo **datos_7.csv**. Se registró un total de $p = 351$ especies de un total de $N = a \times b \times c \times n = 80$ rizoides. Para este ejemplo empezaremos concentrándonos en los cambios en la riqueza de especies por rizoide. El objetivo aquí es discriminar la riqueza de especies según el diseño experimental jerárquico de 3 factores. Para atender esta pregunta, aplicará las funciones del paquete GAD así como las funciones genéricas `lm` y `anova`. Responda:

a. ¿hay variabilidad significativa entre áreas, entre sitios y entre localidades en la riqueza de invertebrados?
b. Compruebe cuál de las tablas ANOVA es correcta estimando usted mismo los grados de libertad, cuadrados medios esperados y razones de F según el protocolo de Underwood (1997).
c. Si se detecta variabilidad significativa en cualquiera de esos niveles, sería entonces lógico (e interesante) estimar y comparar los tamaños de cada componente de variación, los cuales corresponden a estas diferentes escalas espaciales.
d. ¿Cómo puede representar estos componentes de variación?

e. Como procedimiento de control de calidad de los resultados, verifica que se cumplieron los supuestos de normalidad. Luego, analiza sobre la necesidad de evaluar la homogeneidad de las varianzas. En función de tu análisis justifica la decisión de evaluar o no este requisito.