Modelación Estadística

Actividada 4: Modelo lineal y ANOVA

Dr. Edlin Guerra

PARTE 2: Aplicación del Análisis de Varianza (ANOVA) desde R

Continuamos con el caso de la planta procesadora de celulosa en Uruguay. Una vez iniciada las operacioens de la planta, se requiere evaluar el comportamiento de la Demanda Bioquímica de Oxígeno (DBO) a lo largo de los ríos que captan el agua contaminada. Con este propósito, se ha recomendado que la DBO se mida en cuatro lugares: (1) río arriba (zona carente de contaminación por los efluentes de la planta), (2) en los alrededores de descarga de la planta, (3) en la desembocadura del rió Uruguay sobre el río de la Plata, y (4) en la desembocadura del Río de la Plata sobre el océano Atlántico (Figura 1).

Figure 1: **Mapa del área afectada. Localidades muestreadas marcadas con estrellas**

1. Importa los datos en R, y verifica sus características y estructura. ¿Cuántas dimensiones tiene la tabla que importaste? ¿En qué difiere esta de aquellas usada en las pruebas de t para dos muestras?

2. Lleva a cabo una breve exploración gráfica y numérica de los datos que te permita responder a las siguientes preguntas. Para esto usa las funciones `boxplot` y `aggregate`. Este último para calcular las medias y desviaciones estándar de cada localidad con las funciones `mean` y `sd`, respectivamente. Note que no podrá estimar ambos parámteros simultáneamente, por lo que deberá calcularlos separadamente, y luego combinar resultados para tener una sola tabla. Recomiendo que para ambas funciones (`boxplot` y `aggregate`) use el método *S3 method for class 'formula'*.

```
#Completa los argumentos para la construcción de el gráfico de cajas
boxplot()

#Completa los argumentos para la estimación de promedio y desviación estándar
promedio <- aggregate()
promedio

desv.est <- aggregate()
desv.est

# Combinar ambos resultados en una tabla
tabla1 <- data.frame(
    "localidades" = promedio$localidades,
    "DBO" = c(NA, NA, NA, NA),
    "Desv.Est" = c(NA, NA, NA, NA)
  )

tabla1[,2]<-promedio[,2]
tabla1[,3]<-desv.est[,2]
tabla1
```

Responde a las siguientes preguntas: a) ¿Son similares o diferentes los valores promedios de las 4 localidades? b) ¿Son similares o diferentes las dispersiones de las 4 localidades? c) ¿Cómo es la distribución de la variable de respuesta? d) ¿Es esta distribución similar entre los distintos niveles?

3. Aplica un **ANOVA** a los datos. Para ello se requiere primero obtener un modelo lineal usando la función `lm`. Esta función ajusta un modelo lineal de la variable de respuesta en función de la variable explicativa. Como en este caso la variable explicativa es un factor (categórico), es conveniente hacerlo explícito. Puedes ajustar el modelo usando la *localidad* como variable explicativa. Copia el siguiente comando y analiza la respuesta que R devuelve (PISTA: la primera linea de la respuesta es el modelo).

```r
#Especificar que localidades son una variable expicativa (factor)
datos$localidades <- as.factor(datos$localidades)

#Preguntamos si localidades son reconocidas como factor en R
is.factor(datos$localidades)

#Modelo lineal
lm(DBO ~ localidades, data = datos)
```

Responde a las siguientes preguntas: a) ¿Reconoces alguno de los valores bajo el título de *Coefficients*? ¿Qué crees que son éstos valores? b) ¿qué representa el primer coeficiente generado por `lm`?

4. Guarda el modelo que acabas de ajustar bajo un objeto con el nombre *mod1*, y aplica la función `anova` a dicho objeto.

Responde a las siguientes preguntas:

a) ¿Qué hace la función `anova`?
b) ¿Qué es la *Sum Sq* correspondiente al factor *localidades* y a los residuales?¿qué es *Df*?
c) ¿Cuánto vale la *Sum Sq* total?
d) ¿Corresponden los valores de *Sum Sq* y *Df* que aparecen en la consola con aquéllos calculados en la primera parte de la actividad?
e) ¿Qué es la *Mean Sq*?
f) ¿Qué representa el valor de F de la tabla? ¿Es un valor grande o pequeño? ¿Cómo lo sabes?
g) ¿Qué representa el valor de probabilidad? ¿Es un valor grande o pequeño? ¿Cómo lo sabes?
h) A partir de este resultado, concluye si tienes evidencias suficientes para rechazar la Ho que formulaste antes.
i) ¿Cuál es la probabilidad de equivocarte en esta aseveración?

5. Utilizando la función `qf` obtén el valor crítico de F bajo la hipótesis nula. Los argumentos de la función están en el 'help'. Busca valores de los grados de libertad para el numerador y el denominador en la tabla anterior, y considera un valor de $\alpha = 0.05$. ¿Qué representa este valor?

```r
#Completa los argumentos de la función

qf(p = , df1 =, df2 =, lower.tail=F)
```

6. Intenta predecir lo que sucedería con el valor crítico de F bajo las siguientes situaciones. Después modifica el comando que escribiste en el inciso 5 para corroborar tus predicciones.

a) si se aumenta el valor de $\alpha = 0.10$ (uno en diez chances de equivocarme).
b) si se disminuye el valor de alfa a $\alpha = 0.001$ (uno en mil chances de equivocarme).
c) si aumentas el número de réplicas en este experimento a $n = 30$ réplicas por cada nivel del factor, manteniendo $\alpha = 0.05$.

7. Aplica la función `summary` al modelo lineal que ajustaste, y responde a las siguientes preguntas:

```
#Completa los argumentos de la función
summary()
```

a) ¿Reconoces algún valor ya obtenido o calculado en el resultado que R devuelve?
b) ¿Qué crees que sea el valor dado en 'Residual Standard Error'?

8. Aplica la función `fitted` al modelo lineal que ajustaste. ¿Qué hace la función `fitted`? ¿Qué pasa si aplicas la función `predict` al modelo lineal? ¿Cuántos hay?

```
#Completa los argumentos de la función
fitted()
predict()
```

9. Para obtener una visualización prolija del modelo con los datos observados, copia los siguientes códigos del paquete `ggplot2`. Estos códigos representarán los valores por localidad, los promedios y desviaciones estándar. Explora cada uno y trata de identificar qué se va ganando a medida que agregas capas.

```
library(ggplot2)

#Figura básica
fig1 <- ggplot(datos, aes(y = DBO, x = localidades)) +
  geom_point()
fig1

#Figura básica con los promedios
fig1.1 <- fig1 +
  geom_point(data = tabla1,
             aes(x = localidades, y = DBO, col = localidades),
             size = 3)

fig1.1

#Figura básica con promedios y barras de desviación estándar
fig1.2 <- fig1.1 +
```

```
  geom_point(data = tabla1,
             aes(x = localidades, y = DBO, col = localidades),
             size = 3) +
  geom_errorbar(
    data = tabla1,
    aes(
      x = localidades,
      ymin = DBO - Desv.Est,
      ymax = DBO + Desv.Est
    ),
    width = 0.2
  )
fig1.2

#figura básica con promedios, barras de desviación estándar
#y cambios en la estética de la figura

fig1.3 <- fig1.2 +
  theme_bw() +
  ylab(expression(paste("DBO ", "(mg ", O[2], "/l/d)"))) +
  xlab("Localidades")

#figura sólo con promedios, barras de desviación estándar
# y cambios en la estética de la figura
fig1.4 <- ggplot(data = tabla1, aes(y = DBO, x = localidades)) +
  geom_errorbar(
    data = tabla1,
    aes(
      x = localidades,
      ymin = DBO - Desv.Est,
      ymax = DBO + Desv.Est
    ),
    width = 0.2
  ) +
  geom_point(aes(col = localidades), size = 3) +
  theme_bw() +
  ylab(expression(paste("DBO ", "(mg ", O[2], "/l/d)"))) +
  xlab("Localidades")
```

Responde a las siguientes preguntas:

a) ¿Qué representan los puntos de color?

b) ¿Qué representan los puntos negros?
c) ¿Qué representan las barras?
d) Desde el punto de vista gráfico ¿qué se gana al pasar de fig1 a fig1.1,luego a fig1.2, a fig1.3 y fig1.4?
e) En el contexto del seguimiento ambiental ¿qué sugiere el resultado?

PARTE 3: Comparaciones múltiples luego del Análisis de Varianza (ANOVA)

Continuamos con el caso de la planta procesadora de celulosa. Ya hicimos la primera evaluación estadística a los valores de DBO, y conseguimos evidencias estadísticas para indicar que la DBO cambia a lo largo del río. Ahora se requiere evaluar el patrón de diferencia, ya que solo algunas combinaciones de diferencias pueden o no reflejar impacto ambiental. Es importante destacar que el resultado del ANOVA que realizamos no implica automáticamente que podemos hablar de un impacto ambiental, se debe detallar más en los resultados para poder señalar esto.

1. Identifica todas las posibles hipótesis alternativas y qué significaría cada una en el contexto del estudio.

2. En caso no tengas desplegado las gráficas y resultados del ANOVA, repite el ANOVA, tablas y gráficos.

3. Para aplicar comparaciones pareadas entre todos los pares de medias de un factor, se puede usar la función `pairwise.t.test`. Esta función toma en cuenta un valor de error común ("pooled"), y realiza todas las comparaciones posibles (como si no hubieras aplicado un ANOVA). Conviene utilizar un método de ajuste a alfa por el alto número de pruebas aplicadas ("family wise error rate"). Resuelto esto, responde:

   a) ¿Cuántas pruebas de t habría que aplicar?
   b) ¿Qué representan los números de esta matriz triangular?

4. Para aplicar el procedimiento Tukey's HSD (Honest Smallest Difference) es necesario que el objeto sea el resultado de una función `aov` (una función distinta de aplicar un ANOVA e R). Toma los mismos datos de DBO analizados, y aplica la función `aov` en sustitución de la función `lm` usada antes. Al objeto resultante aplícale la función `TukeyHSD`.

   a) ¿Qué representan los datos de la columna con el nombre diff? Usa la función 'aggregate' recien aprendida para ayudarte en los cálculos.
   b) ¿Qué crees que sean los que están bajo lwr y upr?
   c) Aplica la función `plot` al objeto que resultó de aplicar la de Tukey. Estudia e interpreta el gráfico que produce.

5. Para aplicar el procedimiento SNK de la librería `GAD` es necesario establecer cuales son factores fijos y cuales aleatorios. Primero tienes que instalar el paquete `GAD`, y luego llamarlo para hacerla disponible en esta sesión de R. Luego tienes que volver explícito que el factor dietas es un factor fijo. Ajustas el modelo, y después aplicas el procedimiento. Identifica estos pasos con las líneas de código a continuación e interpreta la salida.

```
library(GAD)
datos$localidades <- as.fixed(datos$localidades)
mod.lm<-lm(datos$DBO~datos$localidades)
snk.test(mod.lm, term="datos$localidades")
```

7. Compara los resultados de los 3 métodos usados.

a) ¿Cuál produce un mayor número de resultados significativos?
b) ¿Cuál involucra menor número de pruebas?
c) ¿Cuál prefieres en este caso? ¿Por qué?