TP. 1 MOEA

Mariana Vargas V.
prof. Mónica Balzarini

June 15, 2017

Abstract

Se nos plantea el problema de decidir qué tipo de plantación, entre varios tipos, es más eficiente en términos de la altura promedio alcanzada por árboles de cerezo. En este trabajo desarrollamos el análisis estadístico de los datos y arribamos a una conclusión respecto de qué tratamiento se adecúa mejor a nuestros objetivos.

1 Descripción de los modelos y salidas (problema 1)

Hemos analizado cinco modelos que describiremos a continuación.

- **Primer modelo**. Este modelo puede escribirse como

$$y_{i,j,k} = \mu + a_i + t_j + b_k + (at)_{i,j} + (tb)_{j,k} + \epsilon_{i,j,k} \tag{1}$$

en donde

  - $y_{i,j,k}$ es la altura promedio de los árboles en el bloque k, bajo el tratamiento j, en el año i, $k = 1, 2, 3, 4$, $j = 1, ..., 5$, $i = 1, ..., 7$.
  - a_i es el efecto del i-ésimo año y es un efecto fijo,
  - t_j es el efecto, también fijo, del j-ésimo tratamiento,
  - b_k es el efecto del bloque k, también fijo,
  - $(at)_{i,j}$ es el efecto de la interacción entre el año y el tratamiento, que también es considerado efecto fijo,
  - $(tb)_{j,k}$ es la interacción entre el tratamiento y el bloque, considerado efecto aleatorio.
  - $\epsilon_{i,j,k}$ es el residuo.

  Recordemos que un efecto aleatorio representa valores tomados aleatoriamente de una distribución de niveles de un factor aleatorio, es decir que en este modelo asumimos que la interacción entre el bloque y el tratamiento son una muestra de una población de este factor con distribución normal de media cero y varianzas y covarianzas a estimar. La estructura de

varianzas y covarianzas es tal que

$$G = \begin{bmatrix} \sigma_p^2 & \sigma_p^2 & \dots & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 & \dots & \sigma_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p^2 & \sigma_p^2 & \dots & \sigma_p^2 \end{bmatrix}$$

y

$$R = \sigma^2 I$$

en donde p es el factor *bloque* $*$ *tratamiento*. Las salidas sugieren que tanto el tratamiento como el año son significativos, no así la interacción entre ellos. Los bloques tampoco lo son, aunque el efecto aleatorio de interacción entre bloques y tratamiento, si.

- **Segundo modelo.** La ecuación de este modelo es

$$y_{i,j,k} = \mu + a_i + t_j + b_k + (at)_{i,j} + \epsilon_{i,j,k} \tag{2}$$

La principal diferencia con el anterior es que esta vez consideramos los datos como datos longitudinales: con el comando `repeated` le decimos a SAS que el nivel *tratamiento*$*$*bloque* (que llamaremos p por conveniencia) se repite a través del factor *anio*. Es decir, las instancias de esta interacción serán nuestros sujetos. Debido a que elegimos una estructura de covarianzas de simetría compuesta el resultado es prácticamente equivalente al anterior (observar que las estimaciones son casi iguales). Lo que ha cambiado es cómo hemos introducido las varianzas. En este caso la matriz R será

$$R = \begin{bmatrix} \sigma^2 + \sigma_p^2 & \sigma_p^2 & \dots & \sigma_p^2 \\ \sigma_p^2 & \sigma^2 + \sigma_p^2 & \dots & \sigma_p^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p^2 & \sigma_p^2 & \dots & \sigma^2 + \sigma_p^2 \end{bmatrix}$$

y $G = 0$. Esto significa que la covarianza de una observación de p en años distintos es aproximadamente σ_p^2, que dos observariones distintas tienen correlación 0, y que la varianza de las observaciones es constante.

- **Tercer modelo.** Este modelo tiene la misma ecuación que el anterior, aunque se propone otra estructura de covarianzas: la autorregresiva de orden 1. Esto quiere decir que la matriz de varianzas y covarianzas será

$$R = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n-1} \\ \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n-1} & \sigma^2\rho^{n-2} & \dots & \sigma^2 \end{bmatrix}$$

de modo que se estimarán los parámetros σ y ρ. Los criterios de verosimilitud AIC, AICC, y BIC indican que este modelo es el mejor de los que hasta ahora mencionamos. Los p-valores indican que el tratamiento, el año, y la interacción entre ellos son significativos.

- **Cuarto modelo.** Seleccionamos la estrucutra de covarianzas autorregresiva de orden 1, y nos concentramos en ajustar los parámetros del modelo de medias. En este caso tratamos al año como una variable continua e incluimos un término cuadrático, por lo tanto estimamos también coeficientes de regresión. La ecuación es:

$$y_{i,j,k} = \mu + c_1 a + t_j + b_k + c_2 a^2 + c_2(a^2 t_j) + c_1(a t_j) + (tb)_{j,k} + \epsilon_{j,k}$$

  En la ecuación, a es la covariable año y c_1, c_2 son coeficientes de regresión a estimar. Observar que en este modelo estamos incluyendo el intercepto, lo que significa que estamos ajustando una estructura de medias que tiene la forma de una media general más los desvíos. El año, el año al cuadrado, el tratamiento, y la interacción entre estos resultaron significativos. Esto nos dice que los datos tienen un comportamiento cuadrático con respecto al año. Además observamos que los criterios de verosimilitud indican que este modelo ajusta mejor que el anterior.

- **Quinto modelo.** En este modelo no incluimos intercepto, es decir, modelamos las medias directamente.

$$y_{j,k} = c_1 a + t_j + b_k + c_2 a^2 + c_2(a^2 t_j) + c_1(a t_j) + (tb)_{j,k} + \epsilon_{j,k}$$

  Cada componente representa la media de esa sub-muestra. El resultado es un modelo equivalente al anterior, con las mismas estimaciones, con la diferencia de que no estamos modelando desvíos.

2 Modelo elegido (problema 2)

Lo que observamos es una interacción significativa entre el año y el tratamiento, lo que implica que las alturas de los árboles se comportan de distinta manera a través de los años para distintos tratamientos. Si hiciécemos un gráfico de promedio de altura vs. años veríamos que las curvas no sería paralelas, sino que más bien tenderían a cruzarse en algún momento. Además notamos que la interacción entre el tratamiento y término cuadrático del año es significativa con coeficientes negativos, lo que sugiere que hay una tendencia cuadrática negativa en la evolución de las alturas con respecto al paso de los años. Los modelos que mejor ajustan son el cuarto y el quinto, que asumen una estructura de covarianzas autorregresiva de orden uno y el factor *anio* como una variable continua para la que se ajustan tendencias polinomiales. Lo que asumimos es que las varianzas de los residuos, σ^2, son constantes, y que la covarianza de los residuos de dos observaciones separadas por t años es de $\sigma^2\rho^t$. Ambos modelos son equivalentes. Podríamos seleccionar el cuarto para entender los factores como desvíos de una media general.

3 Resultados estadísticos (problemas 3 y 4)

Podemos concluir que el crecimiento de los árboles está relacionado con el tiempo que transcurre (en años), que, al ser una variable continua, nos provee de una tasa de crecimiento anual que en este caso es del 68% aproximadamente. Los coeficientes de interacción entre el año y el tratamiento, y el año al cuadrado y

el tratamiento imponen una "corrección" sobre esta tasa anual general. En la mayoría de los casos los coeficientes son negativos, excepto en tres, dos de los cuales tienen coeficiente cero. Esto nos dice que el crecimiento de los árboles se va desacelerando a medida que pasan los años.

Para decidir qué tratamiento fue mejor al cabo de siete años debemos estudiar las medias de las alturas del último año, en particular si la diferencia entre ellas resulta significativa. Las medias del séptimo año son:

	Tratamiento	Altura Promedio
1	1	3.40
2	2	3.38
3	3	3.38
4	4	3.52
5	5	3.90

Usamos el test LSD de Fisher para estudiar la diferencia de medias. Ajustamos un modelo lineal sobre los datos del año 7, realizamos un anova y corremos el comando `LSD.test` del paquete `agricolae`. El resultado es:

	trt	means	M
1	5	3.90	a
2	4	3.52	b
3	1	3.40	b
4	3	3.38	b
5	2	3.38	b

Notar que la media del tratamiento 5 es significativamente diferente a las demás, por lo que seleccionamos la plantación de cerezos con alisos en baja densidad.

4 Correlación espacial (problemas 5 y 6)

Ajustamos en R un modelo de correlación espacial usando el comando `gls` de la librería `nlme`. Usamos una estructura gaussiana y una exponencial. Al comparar los modelos con el comando `anova` obtuvimos:

Model	AIC	BIC	logLik
sin correlación	2403.314	2514.337	-1181.657
correlación gaussiana	2360.601	2471.624	-1160.300
correlación exponencial	2334.517	2445.541	-1147.259

Concluimos que el modelo con correlación espacial exponencial es el que mejor ajusta los datos. Nos concentramos ahora en el modelo con y sin correlación espacial para los datos de los cerezos de siete años. Usamos correlación gaussiana y comparamos contra el modelo sin correlación:

```
l_nocorr = gls(Altura ~ Tratamiento + Bloque,
               data = df_Anio7,
               method = "ML",
               na.action = na.omit)
```

```
l_corr = update(l_nocorr,
                correlation = corGaus(form = ~Fila + Columna,
                                      nugget = FALSE))

anova(l_nocorr, l_corr)
```

El resultado de esta comparación arroja:

```
          Model df      AIC      BIC    logLik   Test  L.Ratio p-value
l_nocorr      1  9 513.6581 546.1104 -247.8291
l_corr        2 10 480.7371 516.7952 -230.3686 1 vs 2 34.92102  <.0001
```

indicando que el modelo con correlación espacial gaussiana ajusta mejor (notar que el test de cociente de verosimilitud es significativo). Para verificar qué plantación elegiríamos usamos el test LSD de Fisher nuevamente a los fines de estudiar las medias:

```
anova(l_corr) # Esto nos dice que los gl de los residuos es 264.
mean_sq_errors_7 = mean(l_corr$residuals^2); mean_sq_errors_7
lsd_corr_esp_7 = LSD.test(df_Anio7$Altura,
                          df_Anio7$Tratamiento,
                          DFerror = 264,
                          MSerror = mean_sq_errors_7,
                          console = TRUE)
```

El resultado es el mismo que antes:

```
Groups, Treatments and means
a   5   3.895833
b   4   3.52
b   1   3.400677
b   3   3.384783
b   2   3.381029
```

con lo que concluimos que la plantación de cerezos con alisos en baja densidad es la más adecuada alcanzando un promedio de 3.895 m.