TP. 1 MOEA

Mariana Vargas V.
prof. Mónica Balzarini

June 13, 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.** El modelo se ve como el anterior, sólo que esta vez modelamos 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*. La estructura de covarianzas es de simetría compuesta. Este modelo es prácticamente equivalente al anterior (observar que las estimaciones son casi iguales), la diferencia reside en que en este caso estamos introduciendo todos los efectos aleatorios en la matriz R, cuya diagonal consistirá de estimaciones de $\sigma^2 + \sigma_p^2$ y estimaciones de σ_p^2 en los demás lugares.

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

  Notar que $G = 0$. Lo que estamos diciendo es que una observación de p está correlacionada con la misma observación en años distintos.

- **Tercer modelo.** En este modelo 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, la interacción entre ellos, y la interacción bloque-tratamiento son significativos.

- **Cuarto modelo.** Podemos decir que hemos seleccionado la estructura de covarianzas autorregresiva de orden 1, por lo que podemos concentrarnos

en ajustar los parámetros. Ajustamos entonces un modelo polinomial de orden dos en el factor año, que ahora consideramos continuo.

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

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. Además observamos que los criterios de verosimilitud indican que este modelo ajusta mejor que el anterior.

- **Quinto modelo.** En este modelo no incluimos intercepto, lo que significa que estamos modelando las medias directamente.

$$y_{i,j,k} = a_i + t_j + b_k + a_i^2 + (a^2t)_{i,j} + (at)_{i,j} + (tb)_{j,k} + \epsilon_{i,j,k}$$

Cada componente representa la media de esa sub-muestra. El resultado es un modelo equivalente al anterior, es decir, que da las mismas estimaciones, con la diferencia de que no estamos modelando desvíos. Para cada tratamiento se ajustó un modelo polinomial en el año que describe el comportamiento de las alturas promedio de los árboles con respecto al paso de los años.

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. El modelo que mejor ajusta es el cuarto: el que asume 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 $\sigma^2\rho^t$.

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. Cada tratamiento tiene un estimador diferente para la interacción con el año y con el año al cuadrado, esto quiere decir que a medida que pasan los años cada tratamiento impone una "corrección" sobre la tasa anual general, que en todos los casos es negativa excepto en tres casos, en dos de los cuales es cero. El resultado es que en general las curvas son parábolas en las que puede verse que el crecimiento de los árboles, en la mayoría de los casos, es cada vez menor. Si lo que nos interesa es elegir la plantación que al término de estos siete años en los que los datos fueron observados, ha producido la mayor altura promedio, basta

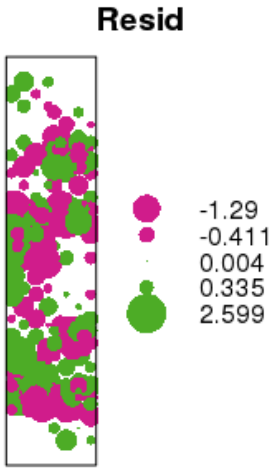

Figure 1: Gráfico de burbujas para ver la correlación espacial.

con hacer los cálculos de la altura promedio alcanzada al cabo de siete años sin tener en cuenta los bloques, que no fueron significativos:

- Tratamiento 1: $3.36m$
- Tratamiento 2: $3.35m$
- Tratamiento 3: $3.35m$
- Tratamiento 4: $3.492m$
- Tratamiento 5: $3.87m$

El último tratamiento sólo depende de su intercepto ya que las interacciones con el año y el año al cuadrado tienen coeficiento cero. Concluimos que la plantación con mayor tasa de crecimiento en siete años fue aquella que combinó cerezo con alisos en baja proporción. Este resultado podría ser diferente para más o menos años transcurridos.

4 Correlación espacial (problemas 5 y 6)

Ajustamos en R un modelo de correlación espacial usando el comando `gls` de la librería `nlme` para el año 7. Elegimos una estructura gaussiana. El modelo ajustado arrojó que la correlación entre dos árboles separados por una distancia de d unidades (metros) es

$$corr = (1 - n) * e^{-(d/r)^2}$$

en donde $n = 0.1$ (nugget effect) y $r = 1.272792$. Por lo tanto existe correlación espacial entre los árboles. Podemos confirmar esto mediante el gráfico de burbujas de la figura 1, que indica que árboles cercanos tienen residuos similares.