Taller de Análisis de datos - Problema 1

Jésica Charaf e Ignacio Spiousas

6 de noviembre de 2023

Problema 1-4

Los siguientes datos corresponden a un trabajo para determinar la composición de un conjunto de vasijas de vidrio de un yacimiento arqueológico. Como el análisis espectrométrico es más barato que el análisis químico, se procuró calibrar el primero para que reemplace al segundo. Con este objetivo se tomó una muestra de $n = 180$ vasijas, a las que se realizó una espectrometria de rayos X sobre 1920 frecuencias, y también un análisis de laboratorio para determinar el contenido de 13 compuestos químicos, a saber:

$$Na_2O,\ MgO,\ Al_2O_3,\ SiO_2,\ P_2O_5,\ SO_3,\ Cl,\ K_2O,\ CaO,\ MnO,\ Fe_2O_3,\ BaO\ y\ PbO$$

Cada fila del archivo `Vessel_X` es el espectro de una vasija, limitado a las frecuencias 100 a 400, pues las demás tienen valores casi nulos. O sea, para cada $i = 1, ..., n, x(i, j(j = 1, .., 301))$ es la energía correspondiente a la frecuencia j (en realidad la frecuencia es j+99, pero podemos dejar eso de lado).

Cada fila del archivo `Vessel_Y` tiene los contenidos de los 13 compuestos en esa vasija. Vamos a comparar distintos métodos para predecir el compuesto 4 (P_2O_5).

Para familiarizarse con los datos, grafique en función de la frecuencia las medias y varianzas de X, y también algunos espectros (o sea, $x(i, j)$ en función de j para algunos i). Aplique los métodos que le parecen adecuados para este problema, y encuentre el que muestra menor error de predicción.

Para el estimador que mejor funciona:

- Grafique los coeficientes (pendientes) en función de la frecuencia.
- Haga el clásico gráfico de residuos vs. ajustados.
- Si ve algo llamativo (outliers, residuos con estructura) tome las medidas correctivas que le parezcan adecuadas.

Resolución

Análisis exploratorio

Lo primero que vamos a hacer es a graficar el contenido de `Vessel_X.txt`, es decir, la energía por banda de frecuencia de cada una de las 150 vasijas. Esto puede verse en líneas continuas de colores en la Figura 1 junto con el promedio en línea sólida negra. En la figura pareciera indicarse que las diferencias entre vasijas ocurren a determinadas frecuencias (en las que la amplitud es distinta de cero y hay más diferencia entre las mediciones individuales) y, por lo tanto, resulta esperable que la información contenida en esas bandas de frecuencias sea la que más aporte a la determinación del contenido de P_2O_5 (aunque bajo condiciones particulares podría no ser el caso).

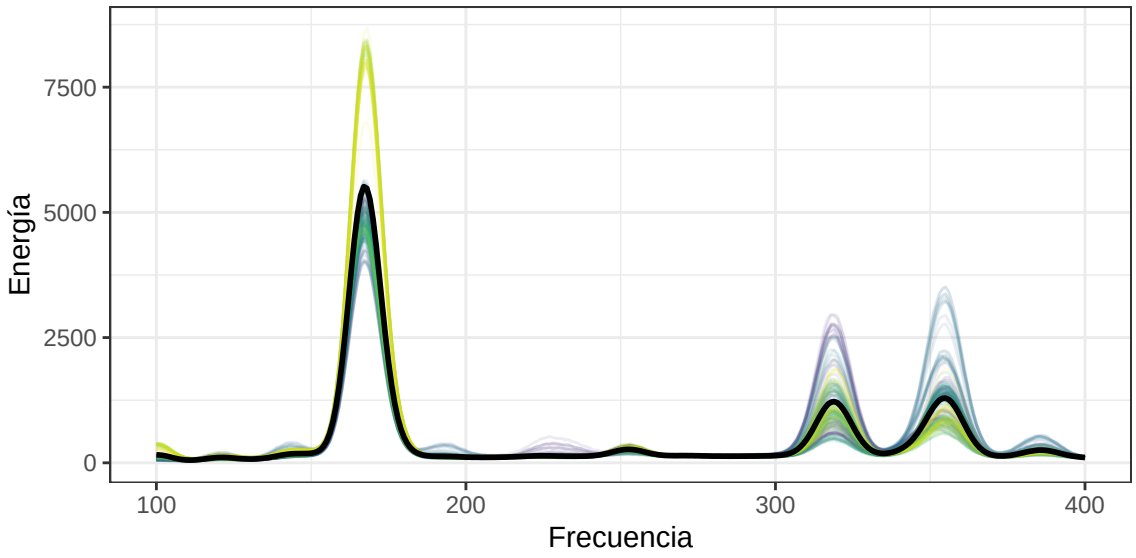

Figure 1: Energía en función de las frecuencias. Cada línea de color representa una vasija mientras que la línea negra representa al promedio de las 150 vasijas.

Para explorar esta ídea un poco más allá podemos ver en la Figura 2 el error estándar de la media en función de la banda de frecuencia. En esta figura vemos cuantificada la intuición que generamos en la Figura 1 de que, efectivamente la variabilidad en las mediciones se concentra en unas pocas bandas de frecuencia.

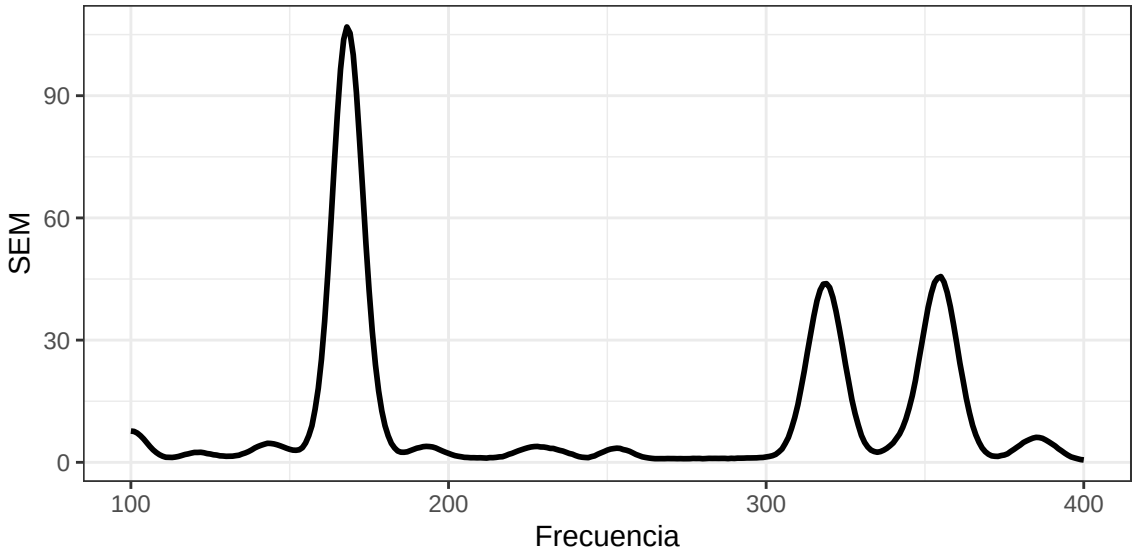

Figure 2: Error estándar en función de las frecuencias.

Luego, lo que queremos ver es como se distribuye la cantidad de P_20_5 en las muestras, para ver si esto tiene algún patrón. En la Figura 3 podemos ver el histograma y la densidad estimada para esta magnitud. En la misma se ve que no pareciera haber valores atípicos y que la distribución es unimodal y con una cola pesada a la izquierda (hacia valores más bajos). Esta asimetría podría llegar a influir en el supuesto no normalidad de los residuos del modelos a ajustar, más adelante lo evaluaremos.

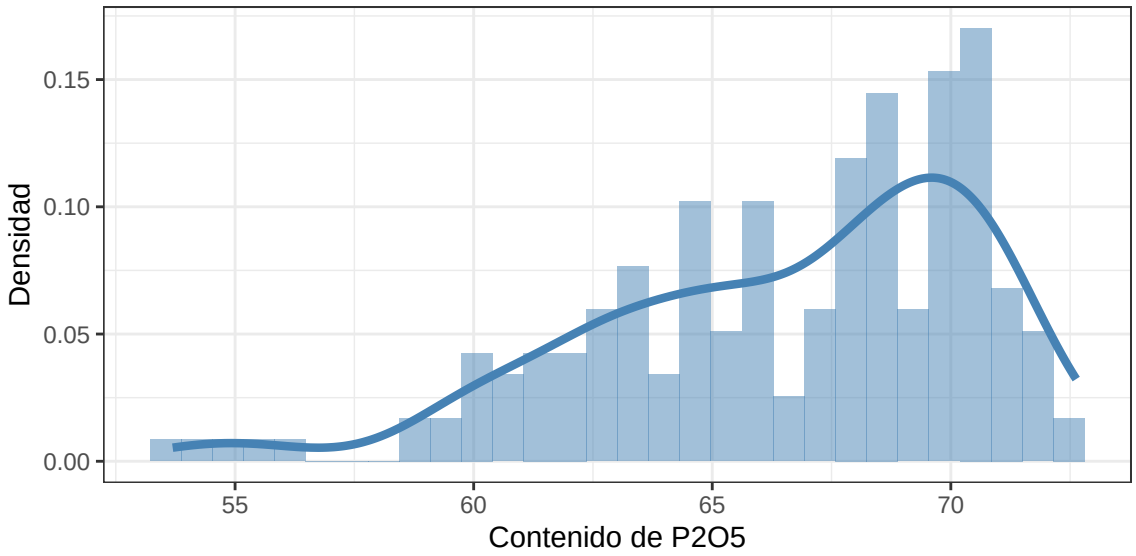

Figure 3: Histograma y estimación de la densidad de la cantidad de P205 en las muestras.

Finalmente, y a modo exploratorio, vamos a calcular el coeficiente de correlación entre la energía de cada banda de frecuencia y la cantidad de P_20_5. De esta forma queremos seguir indagando sobre qué bandas de frecuencia deberían ser más importantes en el modelo de predicción. En la Figura 4 puede verse el valor absoluto del coeficiente de correlación de Pearson en función de la banda de frecuencia. Retomaremos los resultados de esta figura luego de ajustar un modelo.

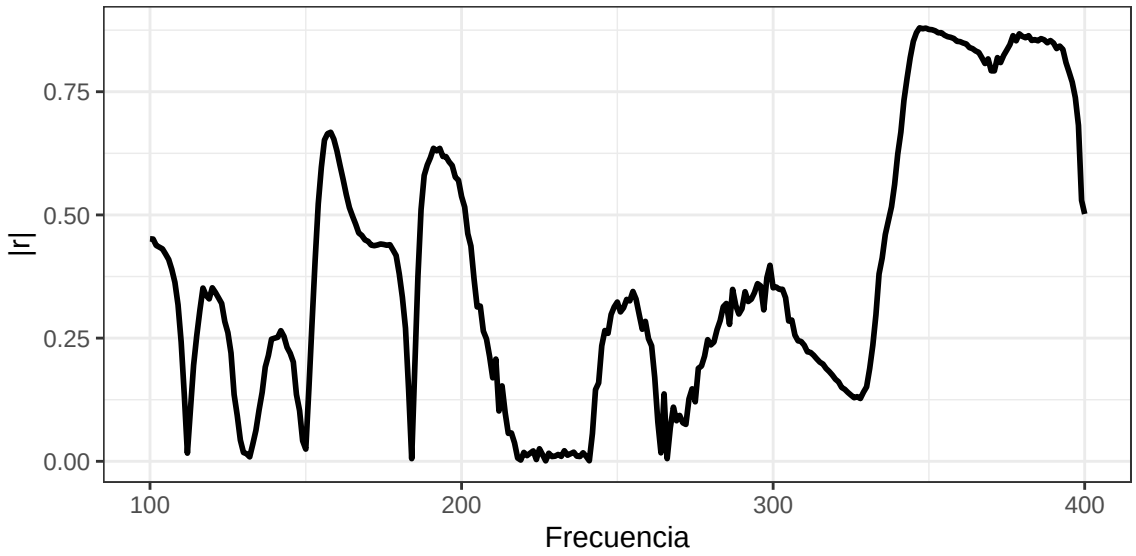

Figure 4: Energía en función de las frecuencias. Cada línea de color representa una vasija mientras que la línea negra representa al promedio de las 150 vasijas.

Modeloado utilizanod `glmnet`

Un modelo básico

Acá analizar el tema de que cuando hacemos Lasso se quead con las que esperábamos (hacer de nuevo la figura de la correlación pero coloreando las que se queda)

Separamos una porción de los datos para testeo

Ver el tema de la estratificación

Buscando los mejores parámetros λ y α

Hacerlo con k folds

El modelos final

Métricas y resultados del modelo final ajustado con todo el train