

Trabajo Práctico Final

Regresión No Paramétrica y Métodos de Regularización en Modelo Lineal

Este trabajo práctico es de carácter obligatorio, y la nota formará parte de la calificación final de la materia. Se debe entregar un informe .Rmd en formato html o pdf con la resolución y resultados del ejercicio, incluyendo todos los gráficos que crean pertinentes y el archivo .Rmd donde se realizaron los cálculos y se programó la implementación del análisis pedido. El trabajo se puede realizar en grupos de 2 ó 3 integrantes. Ver plazo e instrucciones de entrega en el campus.

En el archivo `body.xls` se encuentran datos de morfología corporal humana. El objetivo principal es estudiar la relación entre el peso y distintas variables morfológicas basándose en 507 individuos: 247 hombres y 260 mujeres. Las variables medidas son:

Medidas esqueléticas:

- BIAC: Diámetro biacromial
- BIIL: Ancho pélvico
- BITRO: Diámetro bitrocantérico
- CHEST1: Profundidad del tórax entre la columna y el esternón, a mitad de la espiración
- CHEST2: Diámetro del tórax, a mitad de la espiración
- ELBOW: Diámetro del codo, suma de dos codos
- WRIST: Diámetro de muñeca, suma de dos muñecas
- KNEE: Diámetro de rodilla, suma de dos rodillas
- ANKLE: Diámetro del tobillo, suma de dos tobillos

Medidas de circunferencia:

- SHOUL: Circunferencia de hombros sobre los músculos deltoides
- CHESTG: Circunferencia del pecho a mitad de la espiración
- WAISTG: Circunferencia de la cintura, promedio de la posición contraída y relajada
- Navel: Circunferencia abdominal
- HIP: Circunferencia de la cadera al nivel del diámetro bitrocantéreo

- GLUTE: Circunferencia del muslo por debajo del pliegue del glúteo, promedio de las circunferencias derecha e izquierda
- BICEP: Circunferencia del bíceps, flexionado, promedio de las circunferencias derecha e izquierda
- FOREA: Circunferencia del antebrazo, extendido, palma hacia arriba, promedio de circunferencias derecha e izquierda
- KNEEG: Circunferencia de rodilla sobre rótula, posición ligeramente flexionada, promedio de circunferencias derecha e izquierda
- CALF: Circunferencia máxima de pantorrilla, promedio de circunferencias derecha e izquierda
- ANKLEG: Perímetro mínimo del tobillo, promedio de los perímetros derecho e izquierdo
- WRISTG: Circunferencia mínima de muñeca, promedio de circunferencias derecha e izquierda

Otras medidas:

- AGE: Edad (años)
- WEIG: Peso (kg)
- HEIG: Altura (cm)
- GEN: Género (1 - masculino, 0 - femenino)

Lectura de Datos

- (a) Cargue los datos de `body.xls`. Verifique que las variables tengan el tipo correcto para ser tratadas y controle la presencia de valores faltantes. Puede usar el comando que se da abajo para dar nombre a las variables después de la lectura del archivo.

```
colnames(body)<- (c("BIAC", "BIIL", "BITRO", "CHEST1", "CHEST2", "ELBOW", "WRIST",
"KNEE", "ANKLE", "SHOUL", "CHESTG", "WAISTG", "NAVEL", "HIP", "GLUTE", "BICEP",
"FLOREA", "KNEEG", "CALF", "ANKLEG", "WRISTG", "AGE", "WEIG", "HEIG", "GEN"))
```

Etapla Exploratoria

- (b) Supongamos que en primera instancia los investigadores están interesados en algunas características distribucionales de las variables medidas. A modo de ejemplo consideremos la mediana de la variable WEIG en cada uno de los dos géneros. A partir de los datos, estime la mediana del peso de cada género y calcule por dos métodos bootstrap distintos un intervalo de nivel aproximado 0.95 para cada una de ellas. Interprete.

- (c) Los investigadores creen que las variables **WEIG** y **HEIG** presentan una asociación que puede variar de acuerdo al género. En una primera etapa exploratoria de los datos, realice un diagrama de dispersión de **HEIG** (eje x) vs. **WEIG** (eje y) discriminando por **GEN** ¿Qué sugiere este gráfico?
- (d) Supongamos que ahora interesa explorar esta relación entre las variables **WEIG** y **HEIG** en cada género. Utilice el comando `ksmooth` de R para dicho propósito. Con `ksmooth` ajuste para cada género una regresión no paramétrica usando el núcleo normal y utilice el argumento `bandwidth=10`. Agregue a cada gráfico del ítem anterior las dos regresiones obtenidas mediante el comando `lines` ¿Qué sugieren los gráficos obtenidos? ¿Qué tipo de relación sospecha para cada género?
- (e) Implemente un código que realice la búsqueda de la ventana óptima para `ksmooth` con núcleo normal para el parámetro `bandwidth`. Utilice el criterio de convalidación cruzada basado en *leave-one-out* y realice la búsqueda en una grilla de `bandwidth` entre 5 y 20 con paso 0.5.

Para cada género grafique la función objetivo y represente allí la ventana óptima hallada de acuerdo al criterio que está utilizando.

Nota: Para verificar la respuesta, explore la función `train` de la librería `caret`.

- (f) Para cada género realice el diagrama de dispersión de **HEIG** vs. **WEIG** usando todos los datos y superponga la estimación de la regresión no paramétrica que obtiene con la ventana óptima hallada. Asimismo, superponga la recta que obtiene utilizando el método de mínimos cuadrados. Compare y realice una conclusión final.

Regresión Lineal

- (g) Usando un mecanismo aleatorio se dividió la muestra en dos partes: entrenamiento y testeo. En el archivo `TrainTest.txt` los **TRUE**'s representan los datos en la muestra de entrenamiento y los **FALSE**'s los datos en la muestra de testeo. Utilizando los datos de entrenamiento, ajuste un modelo lineal para **WEIG** basado en todas las variables explicativas.
- ¿Cómo resulta el modelo ajustado? Evalúe la significación de cada coeficiente. ¿Cuáles dejaría en el modelo? Explique el criterio que utiliza.
 - Observe el valor del estadístico F y relacione con el ítem anterior. ¿Sospecha el efecto de algún fenómeno? En caso afirmativo, ¿cuál?
 - Explorar las correlaciones lineales entre las variables explicativas. Para ello, puede utilizar un *heatmap* de la matriz de correlaciones entre ellas. ¿Qué efecto produce la presencia de multicolinealidad en los errores de estimación de los coeficientes?
 - Calcule el error de predicción empírico del modelo ajustado en el grupo de testeo.
- (h) Elegir algunas covariables, las que el grupo considere apropiadas, tomando como criterios: 1) Conservar variables que hayan resultado significativas en el primer análisis, 2) Garantizar que en el conjunto de variables elegidas no haya problemas de multicolinealidad. Ajustar nuevamente el modelo y comparar los resultados. ¿Qué ocurre con R^2 ajustado? ¿Qué ocurre con la varianza residual y con la significatividad de los coeficientes?

- (i) Explore el comando `glmnet` de la librería homónima. Utilizando un modelo lineal con todas las variables, calcule el estimador regularizado usando la penalización LASSO con la muestra de entrenamiento.

Mediante el comando `coef` imprima la tabla de coeficientes para los distintos valores del parámetro de regularización λ e interprete la tabla de valores resultante en relación al gráfico que se obtiene con `plot` del objeto que devuelve `glmnet` usando como argumento adicional `xvar="lambda"`.

- (j) Usando el comando `cv.glmnet` identifique el λ óptimo y utilice el criterio de 1 desvío para elegir el parámetro de regularización.

- ¿Cómo resulta el modelo ajustado usando este valor de λ de un desvío standard? ¿Cómo se relaciona esta estimación de los coeficientes del modelo ajustado en el ítem (g)?
- Calcule el error de clasificación empírico en el grupo de testeo. Compare con el obtenido en el ítem (g)).

- (k) Realice una conclusión final.