Universidad Nacional de Ingeniería
Escuela Profesional de Matemática
Ciclo 2021-II

[**CM2H2 : Estadística Inferencial**]
[**Tema: Selección de un modelo multilineal**]

Proyecto 2

El objetivo del presente proyecto es seleccionar el mejor modelo multilineal. Para esto usaremos la siguiente base de datos datahelado.txt

1. Extraiga los datos de *datahelado.txt* en un data frame utilizando `read.table` no olvidar especificar bien la separación con el parámetro `sep`, `row.names` y `header`.

2. Indique las variables y su tipo(cualitativas y cuantitativas).

3. Utilice `plot(data)` (sin parámetros) y `summary(data)`. Explique lo obtenido.

4. Divida los datos en dos partes el 70 % llamado dataAprendizaje y 30 % llamada dataTest. Enseguida, solamente utilizaremos dataAprendizaje para la regresión multilineal.

5. Realice una regresión multilineal con todas **las variables cuantitativas** para predecir el consumo dada por **cons**, utilice la función `lm()` de R, dicho modelo lo denotaremos por *modelo1*.

6. Indique cuantos modelos multilineales se pueden generar con las variables.

7. Explique sin muchos detalles un algoritmo para determinar el mejor modelo entre todos los anteriores.

   Criterio - BIC(Bayesian Information Criterion)
   El BIC es un escalar que nos indica que tan bueno es nuestro modelo, cuanto más pequeño es el BIC entonces el modelo multilineal es mejor. Definimos BIC por la siguiente fórmula:

   $$BIC = n\ln\left(\frac{SSE}{n} + k\ln(n)\right)$$

   donde k es el número de parámetros y n cantidad de los datos.

8. Crea un función en R para obtener el BIC de la forma `BIC_ML(Y,X)`.

   Hemos mencionado que cuando menor es el BIC entonces el modelo dado es mejor. Sin embargo, necesitamos una forma de búsqueda entre los modelos posibles. Para esto utilizamos un procedimiento llamado **stepwise model selection** esto se puede realizar de tres formas en **backward** , **forward** y en **both**. En el caso **backward** realiza lo siguiente: Inicia con un modelo completo o inicial, enseguida quita una variable independiente obtiene el BIC en cada caso, y se queda con el modelo que tiene menor BIC y descarta los otros, y esto lo repite con el nuevo modelo obtenido, hasta que ya no disminuya el BIC. El caso **forward** es similar aunque inicia con un modelo constante o inicial e incrementa uno en uno la variable hasta obtener el modelo con menor BIC. El caso **both** en cada paso se decide si se agrega o quita una variable independiente, a diferencia de los modelos anteriores se puede volver a agregar o quitar las variables que han sido agregadas o quitadas.

9. En R tenemos el comando `MASS::stepAIC(modelo_lineal, direction  = "tipo", k = log(n))` donde modelo_lineal es nuestro modelo, tipo es backward, forward, both, y n es el cantidad de los datos(número de filas). Modo de uso:

```
# Modelo sin variables independientes
modZero <- lm(Y ~ 1, data = dframe )
# Modelo con todas las variables independientes
modFull <- lm(Y ~ . , data = dframe)
# Modelo intermedio con algunas variables
modInt <- lm(Y ~ X1+X2+X3, data = dframe)
# Busqueda backward
MASS::stepAC(modFull, direction = "backward", k= log(n))
# Busqueda forward
```

```
MASS::stepAC(modZero, direction = "forward",
scope = list( lower = modZero, upper = modFull),
k= log(n))
# Busqueda both
MASS::stepAC(modInt, direction = "both",
scope = list( lower = modZero, upper = modFull),
k= log(n))
```

Utilice una búsqueda **backward** en nuestro *modelo1.* Comente el resultado en la consola e interprete.

10. Utilice las otras búsquedas **forward** y **both**. Elija el mejor modelo al cual lo llamaremos modelo2. Esto puede ser realizado con una asignación de la forma

```
modelo2 <- MASS::stepAC(..., direction = " ... ",
scope = list( lower = modZero, upper = modFull),
k= log(n))
```

11. Enseguida, podemos predecir los valores utilizando nuestros datos, no utilizados, **dataTest**. Utilizando `predict` obtenga el valor predicho en ambos modelos y guárdelos en vectores `pred1` y `pred2`.

12. Denotamos por `N` la talla de datosTest y por $\mathbf{Y}$ la variable **cons** de datosTest y por $\mathbf{Y}^1$ los datos `pred1`, y $\mathbf{Y}^2$ los datos `pred2`. Finalmente, calcule e interprete:

$$MSE_1 = \frac{1}{N}\sum_{i=1}^{N}(Y_j - Y_j^1)^2, \qquad MSE_2 = \frac{1}{N}\sum_{i=1}^{N}(Y_j - Y_j^2)^2,$$

Los profesores .
UNI, 23 de noviembre de 2021.