# A4 - Análisi de varianza y repaso del curso

Enunciado

Semestre 2021.1

# Índex

# 1 Introducción

El conjunto de datos `CensusIncomedata.txt` se inspira (ha sido modificado por motivos académicos) en un elemento de la base de datos disponible en la web Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Adult.

Este conjunto de datos contiene información de una muestra extraída a partir de un censo, en el que para cada persona, se registran los salarios aparte de información personal adicional. El conjunto de datos contiene 32.560 registros y 9 variables.

Las variables de esta muestra son:

- *Age:* Edad del individuo.
- *Workclass*: Categorización del individuo en base al perfil laboral.
- *Education_num*: Número de años de formación educativa del individuo.
- *Marital_status*: Estado civil del individuo.
- *Occupation*: Categorización del individuo en base a la tipología de trabajo.
- *Race*: Grupo racial al que pertenece el individuo.
- *Sex*: Género del individuo.
- *hours_per_week*: Horas por semana trabajadas por el individuo.
- *income*: Salario (anual) del individuo, en k€.

Estos datos nos ofrecen múltiples posibilidades para consolidar los conocimientos y competencias de manipulación de datos, preprocesamiento, análisis descriptivo e inferencia estadística, así como la regresión (lineal y logística) y el Análisis de Variancia (ANOVA).

Verás que, en relación a estos datos, pondremos el foco en el estudio de la probabilidad de no alcanzar cierto umbral de retribución económica en base a las características descritas en el conjunto de datos.

**Nota importante a tener en cuenta para entregar la actividad**:

- Es necesario entregar el archivo Rmd y el archivo de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se respetará la misma numeración de los apartados que el enunciado.
- No se pueden realizar listas completas del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden utilizar las funciones **head** y **tail** que sólo muestran unas líneas del archivo de datos.
- Se valora la precisión de los términos utilizados (es necesario utilizar de forma precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de realizar explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de forma clara y concisa.

# 2 Lectura del archivo y preparación de los datos

Leer el archivo `CensusIncomedada.txt` y guardar los datos en un objeto con identificador denominado *adult*. A continuación, verifica que los datos se han cargado correctamente.

### 2.1 Preparación de los datos

- Fíjate en los valores de las variables categóricas para identificar y proceder a quitar los *molestos* espacios en blanco al inicio de los valores.
- Corrige el error en el nombre de la séptima variable, ya que realmente nos queremos referimos al rol social o percepción individual del género propia del individuo. (https://en.wikipedia.org/wiki/*Sex_and_gender_distinct*
- ¿Qué podemos afirmar sobre la normalidad de la variable salario? Ayúdate de la inspección visual y el test conocido de Lilliefors.
- Genera una variable denominada 'Less50' que clasifique binariamente los salarios dado el límite de 50 k€. Como hemos dicho antes, focalizamos sobre tener un ingreso menor en esta cantidad ('Less50'), y por tanto, codificaremos la variable 'Less50' con el valor 1 cuando el salario sea inferior a 50k€, e igual a 0 en caso contrario.

### 2.2 Análisis visual

1. Muestra con varios diagramas de caja la distribución de la variable `income` según las variables `gender`, `race`, `workclass`, `marital_status` y `occupation`.
2. Interesa visualizar también las variables `age`, `hours_per_week` y `education_num`.
3. Interpreta los gráficos brevemente. Aprovecha que las últimas variables son continuas para interpretar su tendencia.

## 3 Estadística inferencial

### 3.1 Contrastes de hipótesis

Nos interesamos ahora por las potenciales diferencias en el salario de los individuos para diferentes grupos, en particular, las mujeres y los hombres, y los grupos raciales blanco y negro.

- ¿Cobran los hombres más que las mujeres? Responde a la pregunta con un nivel de confianza del 95%.
- ¿Cobra la gente blanca 6450€ más al año que la gente negra? Responde a la pregunta con un nivel de confianza del 95%.

**Nota:** Valora la conveniencia de crear funciones que le permitan no repetir cálculos.

Sigue la siguiente estructura de apartados:

#### 3.1.1 Hipótesis nula y alternativa (por el género y por el caso racial)

#### 3.1.2 Justificación del test a aplicar (por el género y por el caso racial)

#### 3.1.3 Aplicación, interpretación y comprobación del test (por el género y por el caso racial)

## 4 Modelo de regresión lineal

### 4.1 Estimación de modelos

- Estima un modelo de regresión lineal múltiple que tenga como variables explicativas: `age`, `education_num`, `hours_per_week` y `gender`, y como variable dependiente el `Income`.

- Genera un segundo modelo pero esta vez añadiendo la variable `race`.

### 4.2 Interpretación de los modelos

Interpreta los modelos lineales ajustados y valora la calidad del ajuste:

– Valora la significación de las variables explicativas.

– Explica la contribución de las variables explicativas en el modelo.

– ¿La inclusión de la variable `race` ha supuesto una mejora del segundo modelo respecto al primero?

### 4.3 Análisis de residuos

Por último, para profundizar en la calidad del ajuste deben analizarse los residuos que nos indicarán realmente cómo se ajusta nuestro modelo a los datos muestrales. Lo haremos sólo por el segundo de los modelos lineales obtenidos.

- La salida de 'summary()' presenta los principales estadísticos de la distribución de los residuos. Analiza los valores estimados de los estadísticos.
- Realiza ahora un análisis visual de los residuos. ¿Qué podemos decir sobre la bondad de la adecuación del modelo?

### 4.4 Predicción

De nuevo, sólo por el segundo modelo estimado, realiza la predicción del `income` esperado para las siguientes características: `age`=24, `education_num`= "4", `hours_per_week`="40", `gender`=" Female", `race`="Black". Proporciona, además, el intervalo de confianza del 95%.

## 5 Regressión logística

Utilizando las variables explicativas posibles, ajusta un modelo predictivo basado en la regresión logística para predecir **la probabilidad de tener un salario menor de 50 k€**. Por eso, usaremos la variable dicotómica `Less50` que ha creado en el primer apartado, que será nuestra variable dependiente del modelo.

Para poder estimar de forma más objetiva la precisión del modelo, separaremos el conjunto de datos en dos partes: el conjunto de entrenamiento (training) y el conjunto de prueba (test). Ajustaremos el modelo de regresión logística con el conjunto de entrenamiento, y evaluaremos la precisión con el conjunto de prueba.

Siga los pasos que se especifican a continuación.

– Generar los conjuntos de train y test

– Entrena el modelo

– Interprete el modelo entrenado

– Evalúe la calidad del modelo sobre los datos de test

– Predición

### 5.1 Generación de los conjuntos de entrenamiento y de test

Genere los conjuntos de datos para entrenar el modelo y para testarlo. Puedes fijar el tamaño de la muestra de entrenamiento a un 80% del original.

### 5.2 Modelo predictivo

Entrene el modelo con el conjunto que acaba de generar. Utilice, como valores de referencia, el valor mayoritario de cada variable. Por ejemplo, para `race`, utilizaremos `White`.

### 5.3 Interpretación

Interpreta el modelo ajustado. Concretamente, explica la contribución de las variables explicativas con coeficiente estadísticamente significativo para predecir el salario de los individuos.

### 5.4 Matriz de confusión

A continuación analiza la precisión del modelo, comparando la predicción del modelo contra el conjunto de prueba (`testing_set`). Asumiremos que la predicción del modelo es 1 (salario por debajo de 50k€) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 de lo contrario. Analice la matriz de confusión y las medidas de sensibilidad y especificidad.

**Nota**: Toma como categoría de interés que el salario esté por debajo de 50k€. Por tanto, `Less50` igual a 1 será el caso positivo en la matriz de confusión y 0 el caso negativo.

### 5.5 Predicción

Utiliza el modelo anterior para realizar predicciones. Haga el cálculo de la predicción manualmente, y use la función `predict` para validar.

– ¿Con qué probabilidad el salario de un individuo será menor a 50k€ para un hombre blanco de 20 años de edad, autónomo (self-employed), con 3 años de estudios, soltero, trabajando en el sector profesional, y ¿trabajando actualmente unas 25 horas semanales?

– ¿Con qué probabilidad el salario de un individuo será menor a 50k€ para un hombre negro de 60 años de edad, con trabajo gubernamental, con 15 años de estudios, casado, trabajando como 'white-collar', y ¿trabajando actualmente unas 35 horas semanales?

## 6 Análisis de la varianza (ANOVA) de un factor

### 6.1 Visualización

En este apartado, nos centraremos en analizar la existencia de diferencias significativas de `income` entre los diferentes grupos raciales. Tomaremos siempre un nivel de significación del 5%.

– Haga un análisis visual de esta dependencia.

### 6.2 Modelo ANOVA

Completa los siguientes apartados:

#### 6.2.1 Formula el modelo

Explica el modelo que se plantea en el ANOVA.

#### 6.2.2 Indica las hipótesis nula y alternativa

Escrivid las hipótesis nula y alternativa.

#### 6.2.3 Estima la signifcación del factor grupo racial

Calculad la variabilidad explicada per la variable `race` sobre la variable `income` mediante la función `anova()`.

#### 6.2.4 Estima los efectos de los niveles de factor

Interpretad los resultados del modelo generado en el apartado anterior.

#### 6.2.5 Realiza los contrastes dos-a-dos

Para los contrastes dos-a-dos, podeis usar, por ejemplo, la función `HSD.test()` del paquete `agricolae`.

#### 6.2.6 Adecuación del modelo

Mostrad la adecuación del modelo ANOVA en los dos siguientes sub-apartados.

**6.2.6.1 Homocedasticidad de los residuos** El gráfico "Residuals vs Fitted" proporciona información sobre la homocedasticidad de los residuos. Mostrad e interpretad este gráfico.

**6.2.6.2 Normalidad de los residuos** Se puede comprovar el supuesto de normalidad de los residuos con los gráficos usuales. Aplicad también el test de Kruskal-Wallis e interpretad los resultados.

# 7 ANOVA multifactorial

La modelización con ANOVA facilita la inclusión de múltiples factores. Estamos interesados en incluir el factor `occupation` para saber si existen diferencias en los ingresos entre los empleos, a la vez que estimar la existencia de interacción significativa entre ambos factores: grupo racial y empleo.

## 7.1 Estudio visual de la interacción.

- Calcula la tabla cruzada entre razas y empleos para saber cuántas observaciones hay por condición. ¿Se trata de un escenario balanceado? Valora los posibles inconvenientes de la modelización basada en anova en caso de un escenario no balanceado.
- Representa la interacción entre ambos factores y comenta los gráficos resultantes.

# 8 Conclusiones

Resuma las principales conclusiones del análisis. Para ello, puede resumir las conclusiones de cada uno de los apartados.

## Puntuación de la actividad

- Apartados 1 i 2 (15%)
- Apartado 3 (15%)
- Apartado 4 (15%)
- Apartado 5 (15%)
- Apartado 6 (15%)
- Apartado 7 (15%)
- Apartado 8 (5%)
- Calidad del informe dinámico (5%)