

Métodos estadísticos para la computación

Trabajo de modelización estadística.

Curso 2020/21

Realiza las acciones que se indican con el Data Frame que has recibido. Debes entregar el trabajo en la tarea que se habilitará en el campus virtual para ello, dentro del plazo de tiempo que se especifica en dicha tarea.

Puedes realizar el código en un equipo con hasta tres personas más. En ese caso, deberás indicar en el documento los compañeros con los que has realizado el código R. En cualquier caso, debes ejecutarlo particularmente para el Data Frame que has recibido y escribir las conclusiones tú mismo para tus propios datos.

Debes entregar estos dos ficheros:

1. Fichero de texto en formato PDF que contenga para cada una de los apartados:
 - Código R realizado para ese apartado.
 - Resultado obtenido en la ejecución de R para ese apartado, copiado directamente de la consola de R.
 - Si lo que se pide son gráficos, los gráficos generados.
 - Una explicación de lo realizado en ese apartado.
2. Fichero script R que contenga todo el código R con suficientes líneas de comentarios explicando lo que se va haciendo. El fichero tiene que poder ser cargado y ejecutado en R studio.

Apartados a realizar:

1. Carga en memoria el fichero CSV como tibble, asegurándote de que las variables cualitativas sean leídas como factores.
2. Construye una nueva columna llamada IMC que sea igual al peso dividido por la altura al cuadrado. La variable explicada será IMC, las variables explicatorias serán el resto de 11 variables exceptuando peso y altura.
3. Elimina completamente las filas que tengan algún valor NA en una de sus columnas.
4. Calcula las medias y desviaciones típicas (no cuasidesviación) de todas las variables numéricas.
5. Calcula los coeficientes de regresión y el coeficiente de determinación para las 11 regresiones lineales unidimensionales.
6. Representa los gráficos de dispersión en el caso de variables numéricas y los box-plots en el caso de de variables cualitativas. En el caso de las variables numéricas (y sólo en ese caso) el gráfico debe tener sobreimpresa la recta de regresión simple correspondiente.
7. Separa el conjunto original de datos en tres conjuntos de entrenamiento, test y validación en las proporciones 60%, 20% y 20%.

8. Selecciona cuál de las 11 variables sería la que mejor explica la variable IMC de manera individual, entrenando con el conjunto de entrenamiento y testeando con el conjunto de test.
9. Selecciona un modelo óptimo lineal de regresión, entrenando en el conjunto de entrenamiento, testeando en el conjunto de test el coeficiente de determinación ajustado y utilizando una técnica progresiva de ir añadiendo la mejor variable.
10. Evalúa el resultado en el conjunto de validación.
11. Expresa tus conclusiones sobre el modelo creado. Incluyendo, al menos, respuestas a las siguientes cuestiones:
 - Que utilidad podría tener el modelo matemático que has obtenido.
 - Que se puede deducir a partir del modelo sobre la relación entre las variables.
 - Problemas que has encontrado en el desarrollo.
 - Qué te ha llamado la atención en el proceso.
 - Qué más podría hacerse y como plantearlo.

Descripción del Data Frame del trabajo:

Nombre: **Descripción:**

Variables para construir la variable dependiente:

1 peso Peso del individuo en Kg
2 altura Altura del individuo en metros

Datos básicos:

2 sexo Sexo
3 edad Edad

Hábitos de vida:

4 tabaco Media de consumo de cigarros por semana
5 ubes Media de consumo de unidades de bebida estándar (UBE) de alcohol por semana
6 carneRoja Media de veces que come carne roja por semana
7 verduras Media de veces que consume verduras a la semana
8 deporte Media veces práctica deporte por semana
9 drogas Media consumo de otras drogas por mes
10 dietaEsp Dieta especial por enfermedad (S/N)

Nivel socioeconómico:

11 nivEstPad Nivel estudio más alto padres(4=doct o master 3=grado, 2=Bach, 1=Secundaria, 0=menos)
12 nivEstudios Nivel de estudios (4=doct o master 3=grado, 2=Bach, 1=Secundaria, 0=menos)
13 nivIngresos Ingresos anuales en el hogar (0:<20k, 1:20k<i<=30k, 2:30k<i<40k, 3:40k<i<60k, 4:>60k)