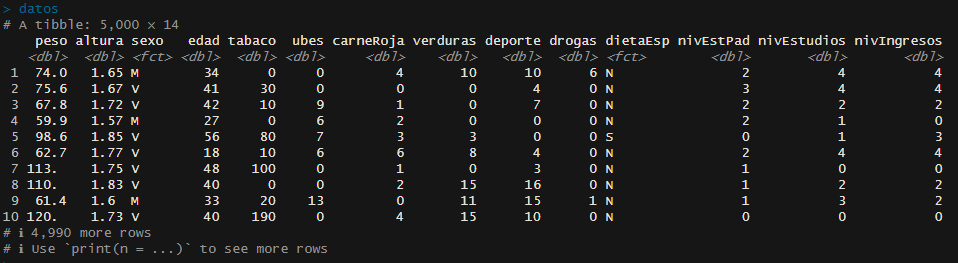
Trabajo de modelización estadística

## 1. Carga en memoria el fichero CSV como tibble, asegurándote de que las variables cualitativas sean leídas como factores

Texto

Descripción generada automáticamente

Para este apartado simplemente importamos la librería que vamos a usar y cargamos el archivo CSV como se pide en el enunciado. Así se vería:



## 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 12 variables exceptuando peso y altura.

Imagen que contiene Texto

Descripción generada automáticamente

Creamos una columna nueva aplicando la formula para aplicar IMC a cada fila. Así se vería:

Imagen que contiene interior, tabla, monitor, computadora

Descripción generada automáticamente

## 3. Elimina completamente las filas que tengan algún valor NA en una de sus columnas. Texto Descripción generada automáticamente

En este apartado primero comprobamos que columnas tienen valores NA. Si ejecutamos na veríamos lo siguiente:

Imagen que contiene Texto

Descripción generada automáticamente

Por último, seleccionamos todas las columnas de datos que en el dataframe na tengan un valor de 0, se vería así:

Pantalla de computadora

Descripción generada automáticamente con confianza media

## 4. Calcula las medias y desviaciones típicas (no cuasidesviación) de todas las variables numéricas.

Texto

Descripción generada automáticamente

Hacemos algo parecido al apartado anterior, guardamos en un dataframe que columnas son numéricas:

Imagen que contiene Texto

Descripción generada automáticamente

Y hacemos una selección igual que antes, solo que esta vez nos quedamos solo con los valores TRUE:

Pantalla de computadora

Descripción generada automáticamente

Ahora calculamos la media de todas las columnas de esta manera:  


Y nos daría:



Y calculamos la desviación típica:

Texto

Descripción generada automáticamente

Primero creamos la función para calcular la desviación típica, tras esto aplicamos la formula a todo el dataframe obteniendo:

Texto

Descripción generada automáticamente

## 5. Calcula los coeficientes de regresión y el coeficiente de determinación para las 12 regresiones lineales unidimensionales.

Captura de pantalla de computadora

Descripción generada automáticamente

Ahora creamos una función que calcule la regresión lineal unidimensional entre dos variables dado un dataframe y que guarde el coeficiente de regresión, el de determinación y la regresión lineal en una formula.

Tras esto aplicamos esta formula a todo el dataframe y obtenemos:

Texto

Descripción generada automáticamente

Veríamos algo así para cada columna del dataframe

## 6. Representa los gráficos de dispersión en el caso de variables numéricas y los boxplots en el caso 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.

En este caso tenemos que crear gráficos, en el caso de las variables numéricas tendrá la siguiente estructura:

Texto

Descripción generada automáticamente

Donde solo ira variando la parte de peso. Las gráficas se verían así:

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

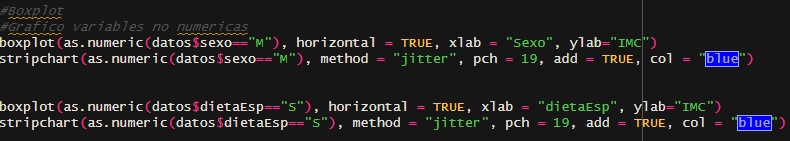
Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Rectángulo

Descripción generada automáticamente

Para las variables no numéricas tendremos el siguiente código:



Y las gráficas se verían así:

Gráfico, Histograma, Gráfico de cajas y bigotes

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente

## 7. Separa el conjunto original de datos en tres conjuntos de entrenamiento, test y validación en las proporciones 60%, 20% y 20%

Texto

Descripción generada automáticamente

Primero creamos una función, la cual dada un dataframe y dos porcentajes, nos separa de manera aleatoria en dataframe en 3 partes:

Imagen de la pantalla de un computador

Descripción generada automáticamente con confianza baja

## 8. Selecciona cuál de las 12 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.

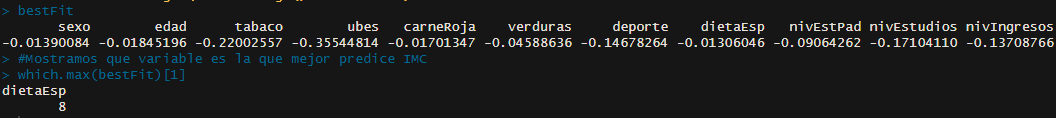
Captura de pantalla de computadora

Descripción generada automáticamente

Para este apartado, primero calculamos las regresiones lineales del dataSet de entrenamiento.

Tras esto creamos una función para calcular R2 ajustado tomando como parámetros un dataframe, un modelo linear y la variable a predecir.

Por último, con el modelo creado con el dataSet de entrenamiento encontramos que variable predice mejor la variable “IMC”. Para esto, le aplicamos a todas las regresiones lineales calculadas antes la función “R2ajustado” y vemos cual es el máximo:



El que nos de un R2 ajustado, será el que mejor prediga IMC para regresiones lineales simples. En este caso diestaEsp.

## 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.

Texto

Descripción generada automáticamente

Primero creamos una función que acepte regresiones lineales múltiples, de esta manera podemos pasarle una lista de variables y que las concatene para hacerlo.

Después una función a la que le pasaremos el conjunto de entrenamiento, el de test, la variable a predecir y las variables predictoras que calculara su regresión lineal y R2ajustado.

Texto

Descripción generada automáticamente

Ahora creamos una función a la que le pasaremos el conjunto de entrenamiento, el de test y el nombre de las variables predictoras. Esta ira haciendo bucles, añadiendo cada vez mas variables a la regresión lineal múltiple para encontrar el mejor R2 ajustado y ver que variable predice mejor IMC.



Por último, aplicamos esta función para ver cuál es la variable:



## 10.Evalúa el resultado en el conjunto de validación.

## Texto Descripción generada automáticamente

Para este apartado vamos a comprobar lo bueno que es nuestro modelo, para ello llamamos a la función de R2ajustado con dataset de validación y sus valores de IMC pero usando como modelo el que acabamos de calcular:

Texto

Descripción generada automáticamente

Como vemos, R2 ajustado sigue saliendo negativo aunque sigue siendo mejor que el calculado en el apartado 8.

## 11. Lee el dataframe de evaluación que te habrá llegado (eval.csv) y utiliza el modelo creado para añadirle una nueva columna con el valor de la variable IMC y, a continuación, otra columna con el valor de la variable Peso. Salva el resultado como evalX.csv para enviarlo como parte de la solución al trabajo.

Captura de pantalla de computadora

Descripción generada automáticamente

Por último cargamos el archivo “eval.csv”. Tras esto predecimos la columna IMC para los datos de “eval.csv” usando el modelo que acabamos de calcular:

Pantalla de computadora

Descripción generada automáticamente

Que como podemos ver, son unos valores que parecen correctos.

Ahora toca calcular peso con los valores IMC predichos:

Imagen de la pantalla de un computador

Descripción generada automáticamente con confianza baja

Que también se ven datos de peso correctos.

Por último, guardaríamos estos datos en un archivo CSV “evalX.csv”.

## 12.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 cómo plantearlo.

Hemos podido ver que en la relación entre variables para este caso es bastante pobre ya que al calcular R2 ajustado salían siempre números negativos, por lo que el modelo no seria algo apto para hacer uso del más allá de simplemente ver como crear un modelo y predecir datos.

El único problema encontrado es que al salir los R2 ajustados negativos pensaba que había hecho algo mal, tras repasarlo más, me di cuenta de que simplemente los datos no eran especialmente buenos para predecir.

Lo único que se me ocurre que se podría hacer es conseguir o mas datos o mejores datos para mejorar el modelo.