# TRABAJO R

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

Para cargar el archivo lo primero que haré será cargar la librería readr para cargar el archivo como tibble, en esta importación además asignare a cada variable su tipo de variable.

**Imagen que contiene Diagrama

Descripción generada automáticamente**

**Resultado:**

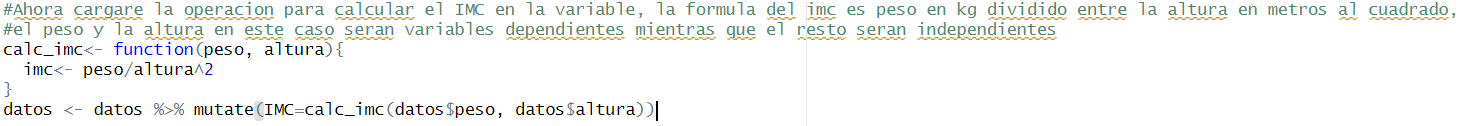
**Imagen que contiene Texto

Descripción generada automáticamente**

**Se han añadido las variables correctamente**

**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.**

**Para crear la columna de IMC creo una función con la formula del IMC que es peso en kg dividido entre altura al cuadrado y realizo un mutate para añadir la columna a la tabla.**

****

**Resultado:**

**Tabla

Descripción generada automáticamente**

**Puedo ver que la columna IMC se ha añadido correctamente.**

## **3. Elimina completamente las filas que tengan algún valor NA en una de sus columnas.**

Para eliminar las filas con datos no disponibles uso na.omit

****

**Resultado:**

**Tabla

Descripción generada automáticamente**

**Puedo ver que la salida no tiene valores na.**

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

Primero para quedarme con las variables numéricas aprovechare la función is.numeric con la función keep, luego aplico map con la función mean para la media, en el caso de la desviación típica aplico con el map una función sin nombre que calcula la desviación típica directamente con la corrección de Bessel aplicada

Texto

Descripción generada automáticamente con confianza baja

**Resultado:**

**Texto

Descripción generada automáticamente con confianza baja**

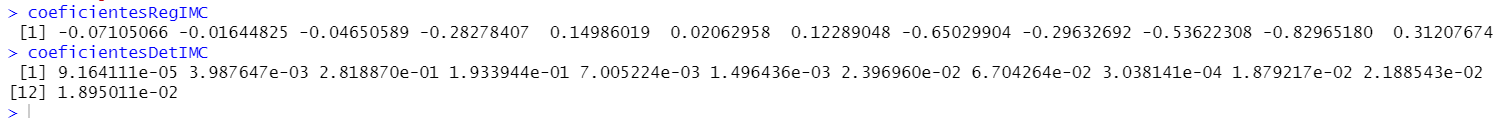
**Los resultados obtenidos me parecen bastante acordes por lo que puedo deducir que mi calculo está bien, en la media podemos apreciar valores que cuadran perfectamente y en la desviación apreciamos una dispersión coherente respecto a la media para cada variable.**

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

Para hallar los coeficientes he tenido que primero aislar las variables independientes, luego crear una función que me calculase cada coeficiente con la entrada del dataset, la variable dependiente y la variable independiente, y finalmente con un map he llamado a estas funciones creadas por mí.

## **Texto Descripción generada automáticamente**

Resultados:

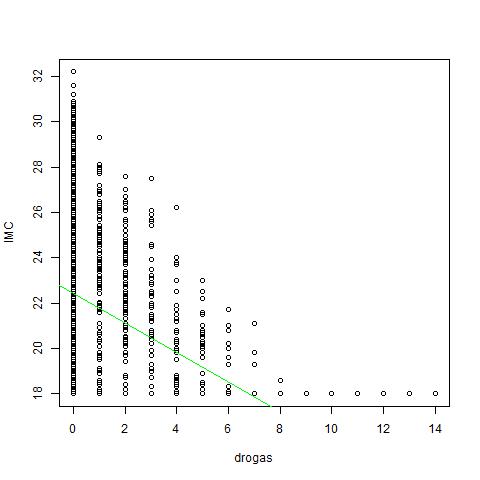
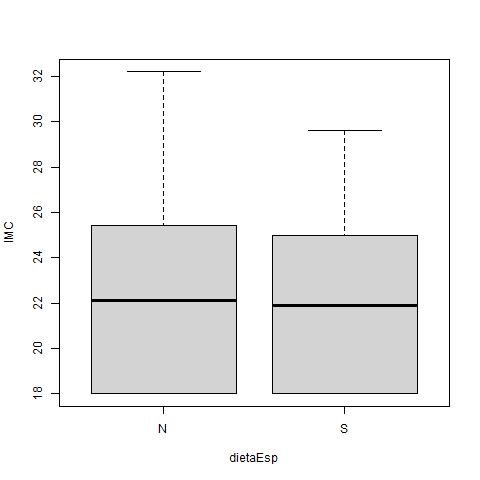
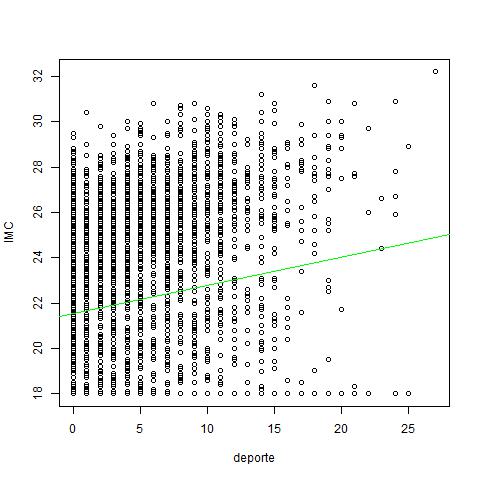
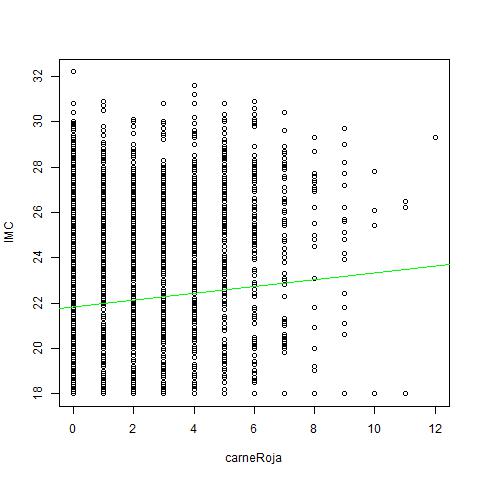
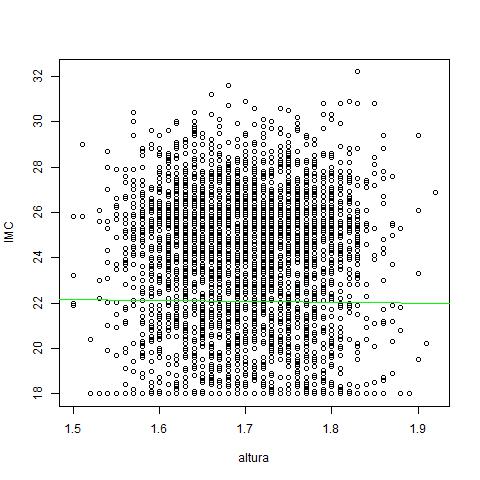
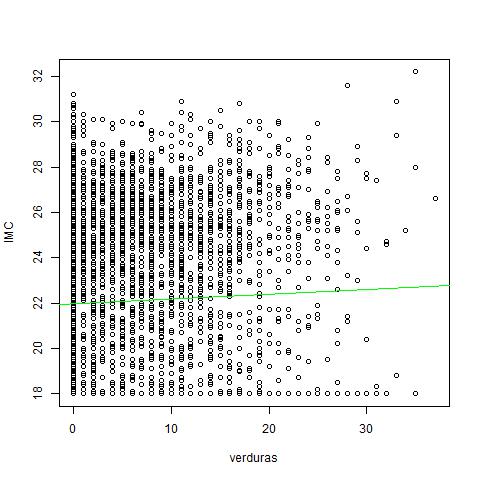
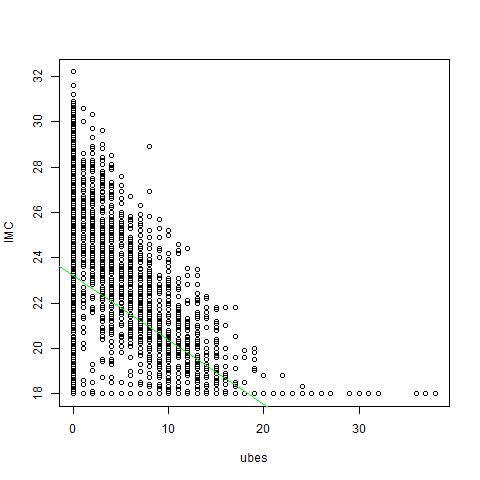
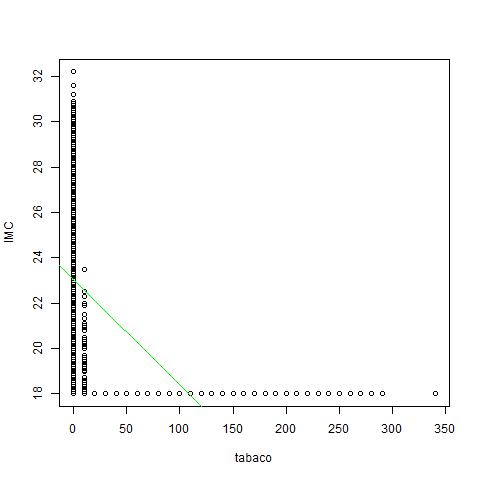
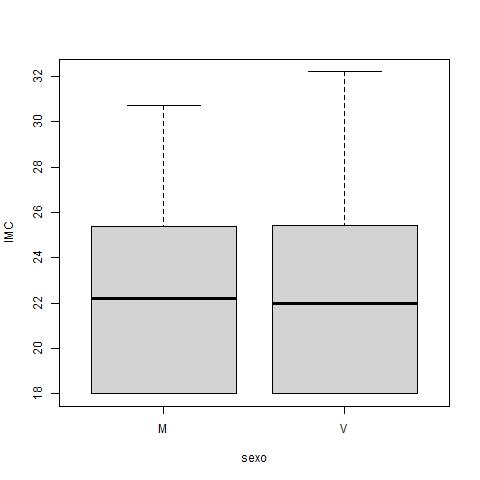
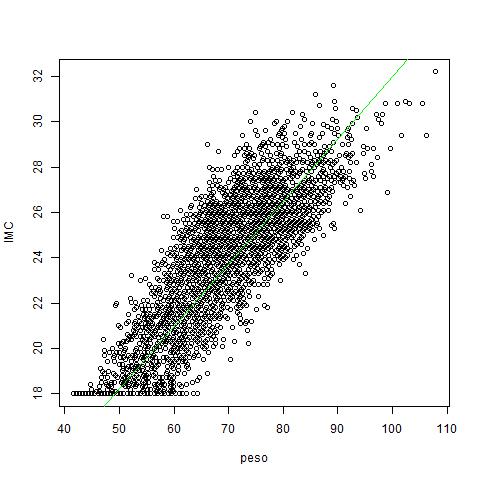
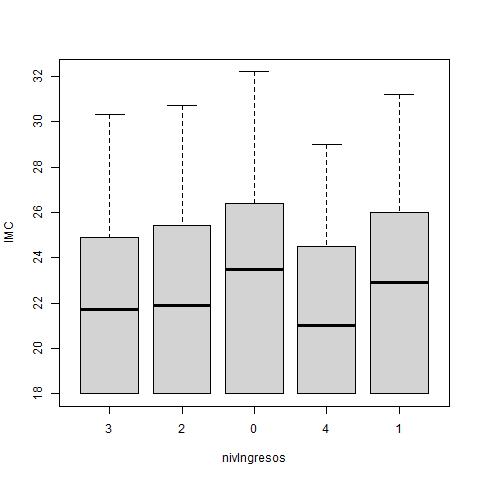
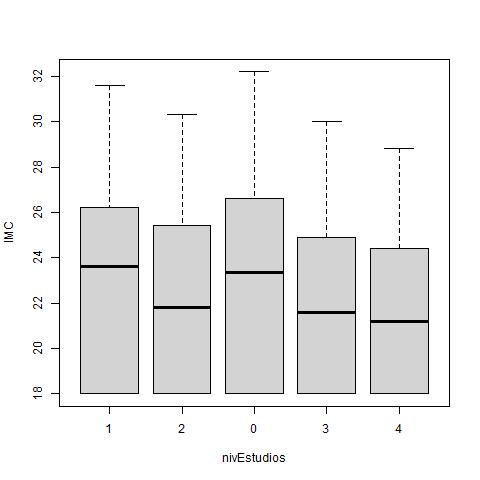
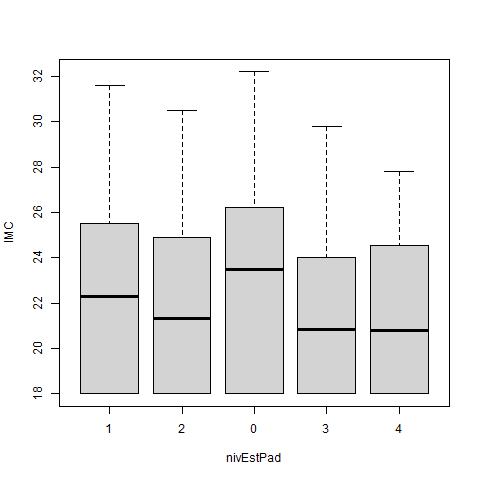
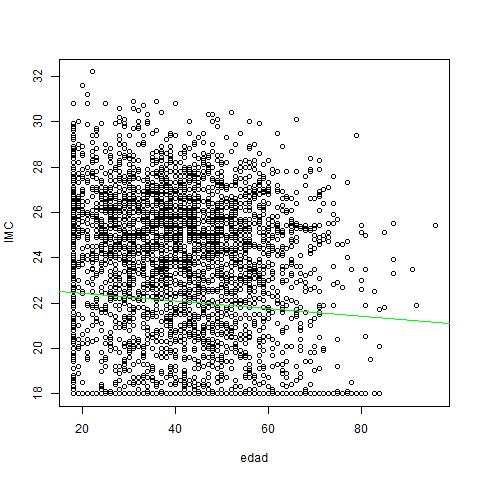


La salida aparenta ser correcta , hay valores negativos que indican un decrecimiento de la regresión y viceversa.

## **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.**

Para llevar a cabo la tarea creo una función que hace el ajuste linear entre las variables, otra que dibuja los modelos, finalmente solo tengo que aplicarlas al dataset y R genera en la carpeta que le he especificado las gráficas abajo mostradas.

**Interfaz de usuario gráfica, Texto

Descripción generada automáticamente**

Los resultados obtenidos representan la variable IMC en función del resto de variables, en estas imágenes ya podemos ir haciéndonos a la idea de que manera variables como las drogas, el tabaco, el deporte pueden afectar al IMC de una persona.

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

Para separar en tres conjuntos creo una función cuya entrada es el dataset, y los tamaños de los conjuntos.

**Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente**

**Resultado:**

**Tabla

Descripción generada automáticamente**

**Se aprecian que los sets se han creado con éxito.**

## **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.**

**Tengo que buscar cual de las 12 variables es un mejor reflejo del IMC.** Para ello, defino una función que calcula el coeficiente de determinación ajustado R2 para un modelo de regresión lineal y otra función que evalúa el modelo de regresión lineal con el conjunto de entrenamiento y prueba. Luego, aplico la función a cada variable predictora y selecciono la variable que genera el coeficiente de determinación ajustado R2 más alto.

**Calendario

Descripción generada automáticamente con confianza baja**

**Resultado:**

****

**El resultado recibido es que la variable tabaco es la que mejor explica el IMC de una persona, esto quiere decir que es la que mas probabilidades tiene de predecir la realidad de la variable independiente.**

## **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.**

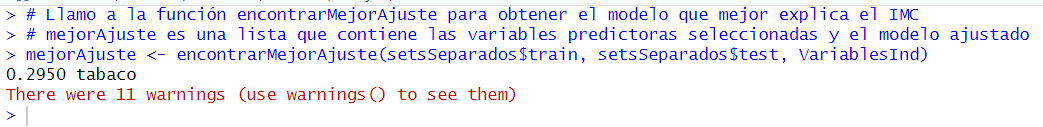
Para encontrar un modelo optimo lineal de regresión uso la función “encontrarMejorAjuste”

para encontrar la combinación de variables predictoras que mejor explican el IMC en el conjunto de datos de entrenamiento y devuelve un modelo ajustado utilizando esas variables.

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente

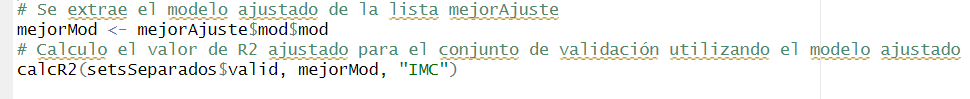
Resultado:



El mejor ajuste es 0.2950 en tabaco

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

Para evaluarlo se extrae el modelo ajustado de la lista devuelta y se utiliza para calcular el coeficiente de determinación ajustado R2 en el conjunto de datos de validación utilizando la función "calcR2". Finalmente, el valor del coeficiente de determinación ajustado R2 se utiliza para evaluar la calidad del modelo ajustado.



Resultado:

Texto

Descripción generada automáticamente

**Como el dato es coherente respecto al del apartado A comprobamos que el resultado es válido.**

## **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.**

Para llevar a cabo esta tarea solo tengo que leer el dataframe, usar predict.lm con el mejor modelo sacado en el ejercicio anterior y finalmente usar la formula del IMC para con la altura y esta columna deducida sacar el valor de peso.

Texto, Carta

Descripción generada automáticamente

## **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.**

Este modelo podría ser muy útil para saber la predisposición a la obesidad o a la anorexia de una persona según sus hábitos y estilo de vida.

## **• Que se puede deducir a partir del modelo sobre la relación entre las variables.**

Se puede deducir que existe una relación entre los hábitos y las características de una persona y su IMC.

## **• Problemas que has encontrado en el desarrollo.**

Los problemas que he encontrado ha sido al manejar vectores y dataframes en R ya que cambia la manera de aplicar algunas funciones y eso me ha producido muchos errores que me ha costado solucionar.

## **• Qué te ha llamado la atención en el proceso.**

Me llama mucho la atención el poder predecir desde una base algo que a primera vista parece imposible de saber, pero que con un análisis de los datos se descubren relaciones entre variables que tienen mucho poder y como esto puede ser aplicado a ámbitos desde marketing, pasando por afluencia de personas o estimaciones de ventas, hasta a la salud.

## **• Qué más podría hacerse y cómo plantearlo.**

Se me ocurre que se podría tener una lista con las variables que mas influyen y tratar de controlar a las personas que reúnen estas debido a que sufren una probabilidad mucho mas alta de tener problemas de salud relacionados con su peso salvando así incluso vidas. Fuera del ámbito de salud hay muchas otras relaciones que también podrían ser interesantes como la relación entre las personas que consumen drogas y su nivel de estudios, la relación entre la gente que hace deporte y el nivel de ingresos o el nivel de ingresos y los estudios de los padres.