TALLER 2

El archivo seguro.csv contiene datos referidos a pólizas de seguros adquiridas por 200 personas, de las cuales se registran los siguientes atributos:

- age: Edad de la persona
- sex: Sexo de la persona
- bmi: Índice de masa corporal
- children: Número de hijos menores de 12 años
- smoker: Indicador de hábito de fumar (0: No, 1: Sí)
- region: Región de residencia de la persona
- charges: Monto mensual que paga, en dólares

Copie cada una de las preguntas y desarróllelas en un formato Rmd, de modo que genere un archivo HTML.

0. Ejecute las siguientes tareas de preprocesamiento de datos, utilizando los códigos indicados en el recuadro inferior (requiere la carga previa de paquetes):
   a. Retire la columna region
   b. Renombre las columnas como edad, sexo, imc, nhijos, fuma, gastos
   c. Recodifique las categorías de la variable sexo como masculino y femenino, fuma: sí y no.

```
library(tidyverse)
datos <- read_csv("seguro.csv")

datos <- datos %>%
  select(-region) %>%
  rename(edad = 1,
         sexo = 2,
         imc  = 3,
         nhijos = 4,
         fuma = 5,
         gastos = 6) %>%
  mutate(sexo = recode(sexo, male = "masculino", female = "femenino"),
         fuma = recode(fuma, yes = "si", no = "no"))
```

Verifique que su objeto datos sea un tibble como se muestra a continuación:

```
> datos
# A tibble: 200 x 6
    edad sexo        imc nhijos fuma  gastos
   <dbl> <chr>     <dbl>  <dbl> <chr>  <dbl>
 1    19 femenino   27.9      0 si      7.59
 2    18 masculino  33.8      1 no     20.9
 3    28 masculino  33        3 no     18.5
 4    33 masculino  22.7      0 no      3.33
 5    32 masculino  28.9      0 no     12.6
 6    31 femenino   25.7      0 no      9.76
 7    46 femenino   33.4      1 no      9.26
 8    37 femenino   27.7      3 no     20.0
 9    37 masculino  29.8      2 no     19.5
10    60 femenino   25.8      0 no     19.7
# ... with 190 more rows
```

1. Se desea estudiar la **influencia lineal de las variables sobre los gastos**. Obtenga la estimación puntual de los coeficientes de regresión:
   a. Matricialmente (forme las matrices X e Y, y trabaje con ellas)
   b. Usando el modelo construido con el comando lm
2. Calcule, utilizando matrices, las sumas de cuadrados de regresión, de error y total.
3. Obtenga la matriz de varianzas – covarianzas estimadas para el vector de coeficientes de regresión estimados:
   a. Matricialmente (utilice la matriz X y obtenga el CME a partir de la pregunta anterior)
   b. Usando la función vcov
4. Muestre una elipse de 96% de confianza para las variables IMC y Número de hijos
5. Escriba el cuadro ANVA, y pruebe la hipótesis de significancia de la regresión con un $\alpha = 0.05$
   a. Utilice el criterio del pvalor
   b. Utilice el criterio de Fcalculado
6. Plantee y desarrolle una prueba de hipótesis para una de las variables predictoras cuantitativas
7. Plantee y desarrolle una prueba de hipótesis para una de las variables predictoras cualitativas
8. Plantee una situación donde el punto evaluado para una predicción NO corresponda a una extrapolación
9. Plantee una situación donde el punto evaluado para una predicción SÍ corresponda a una extrapolación
10. Verifique si se cumple el supuesto de normalidad de errores
11. Verifique si se cumple el supuesto de homocedasticidad de errores
12. Verifique si se cumple el supuesto de independencia de errores
13. Verifique si se cumple el supuesto de linealidad del modelo
14. En caso no se cumpla alguno(s) de los supuestos, proponga y de ser posible, ejecute una posible solución.