Ejercicio 1 (4 puntos)

Diversos estudios han demostrado que los casos de enfermedad de Parkinson han aumentado en los últimos 30 años. Una medida mayoritariamente aceptada en la comunidad científica para valorar la gravedad de la enfermedad de Parkinson en una persona es su puntuación en la Escala Unificada de Enfermedad de Parkinson (UPDRS). Así, un método que pueda predecir la puntuación UPDRS de un paciente de Parkinson será eficaz para determinar la gravedad de la afección del paciente. En este ejercicio trataremos de pronosticar con técnicas de regresión la puntuación UPDRS de un paciente a partir de otros valores de parámetros independientes que se sabe que afectan a dicha puntuación UPDRS.
El conjunto de datos que vamos a estudiar está compuesto por un rango de mediciones de voz biomédicas de 42 personas con enfermedad de Parkinson en etapa temprana reclutadas para un ensayo de seis meses de un dispositivo de telemonitorización para el seguimiento remoto de la progresión de los síntomas. Las grabaciones fueron capturadas automáticamente en los hogares del paciente.
Estos datos se hallan en el Machine Learning Repository:

`http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring`

Las columnas del archivo contienen el número del sujeto, la edad, el sexo, el tiempo desde la fecha de reclutamiento, el UPDRS motor, el UPDRS total y 16 medidas de voz biomédicas. Cada fila corresponde a una de las 5875 gravaciones de voz de estos individuos.

El principal objetivo es predecir el UPDRS total a partir de las 16 medidas de voz.

Los datos se pueden cargar en una sesión de **R** con las siguientes instrucciones[1]:

```
import.data <-
"http://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons
                              /telemonitoring/parkinsons_updrs.data"
parkinson <- read.table(url(import.data), sep=",", skip=1)
names(parkinson) <- c("subject#","age","sex","test_time","motor_UPDRS","total_UPDRS",
  "Jitter(%)","Jitter(Abs)","Jitter:RAP","Jitter:PPQ5","Jitter:DDP",
  "Shimmer","Shimmer(dB)","Shimmer:APQ3","Shimmer:APQ5","Shimmer:APQ11",
  "Shimmer:DDA","NHR","HNR","RPDE","DFA","PPE")
set.seed(123)
```

Observemos que hemos fijado la semilla para que los cálculos pseudo-aleatorios sean fijos.
A continuación separamos la base de datos en dos grupos: el grupo `data.train` con el 80 % de las observaciones y el grupo `data.test` con el resto. La base de datos `data.train` servirá para calcular los diferentes modelos y la base de datos `data.test` para comprobar el ajuste a un grupo externo de datos.

Para evitar problemas es mejor que suprimamos de entrada las observaciones con valores perdidos, si los hay.

Utilizar la puntuación `total_UPDRS` como variable respuesta y las 16 medidas de voz como potenciales regresoras. Se trata de estudiar el mejor modelo por diferentes métodos. En cada caso se informará del número de variables (o componentes) que se utilitzan, el coeficiente R^2 ajustado (cuando se pueda) y el *root mean squared error* (RMSE) para el grupo de ajuste (*train*) y para el grupo de prueba (*test*).

Ajustar los siguientes modelos:

3. Regresión por componentes principales.

   Nota: Utilizar el mínimo de componentes razonable a la vista del gráfico de RMSE.

4. *Partial least squares.*

   Nota: Utilizar el mínimo de componentes razonable a la vista del gráfico de RMSE.

5. *Ridge regression.*

6. LASSO.

7. ¿Cree necesario repetir estos métodos tomando la variable `motor_UPDRS` como respuesta?

Nota: Para calcular el coeficiente de determinación R^2 en algunos métodos que no lo dan explícitamente, hay que hacerlo a partir de su definición:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Cuando el número de variables es grande, también podemos calcular el coeficiente R^2_{adj}

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}$$

donde p es el número de variables explicativas.

Ejercicio 2 (3 puntos)

Con la base de datos `data.train` del ejercicio anterior y el modelo OLS hacer un análisis de los residuos para detectar incumplimientos de las condiciones de un modelo lineal.

1. Investigar si los errores tienen varianza constante.

2. Estudiar la hipótesis de normalidad. Hacer un test de Jarque-Bera.

   Dibujar un gráfico de densidad estimada por kernel de los residuos (también se puede hacer un histograma) y añadir la densidad normal con los parámetros estimados por el modelo lineal. Comentar el resultado.

   Calcular la asimetría y la kurtosis de los residuos, compararla con la de la ley normal y decidir el tipo de distribución que tienen los residuos. ¿Qué transformación podemos hacer para que la distribución de los residuos sea "más" normal?

3. ¿Cuales son los puntos con influencia potencial (*leverage*)? Hacer un gráfico.

4. Con los residuos studentizados externamente (*jacknife residuals*), ¿cuales son los valores atípicos (outliers)? Calcular cuales son estadísticamente relevantes por el método de Bonferroni.

Ejercicio 3 (3 puntos)

Seguimos con la base de datos de los ejercicios anteriores. En esos ejercicios hemos visto que el modelo OLS presenta algunas dificultades. Para superarlas podemos estudiar otros métodos.

1. Eliminar de la base de datos los 3 puntos más influyentes y volver a calcular los modelos del ejercicio 1. Mostrar en una tabla los RMSE del grupo de prueba (*test*) para cada uno de los modelos con y sin los puntos influyentes.

   Nota: En este apartado sólo hay que mostrar la tabla.

2. Dado que el grupo de prueba (*test*) puede contener outliers, calcular un RMSE robusto para cada modelo utilizando la media recortada (*trimmed*) al 10 %. Añadir esta información a la tabla del apartado anterior.

3. Dados los problemas observados con los residuos del modelo OLS, podemos probar un método robusto como el de Huber o el *Least trimmed squares* (LTS).

4. Calcular el método PLS con la función `plsreg1()` del paquete `plsdepot`. Dar el RMSE para el grupo de prueba.