Asignatura

Estadística y R para Ciencias de la Salud

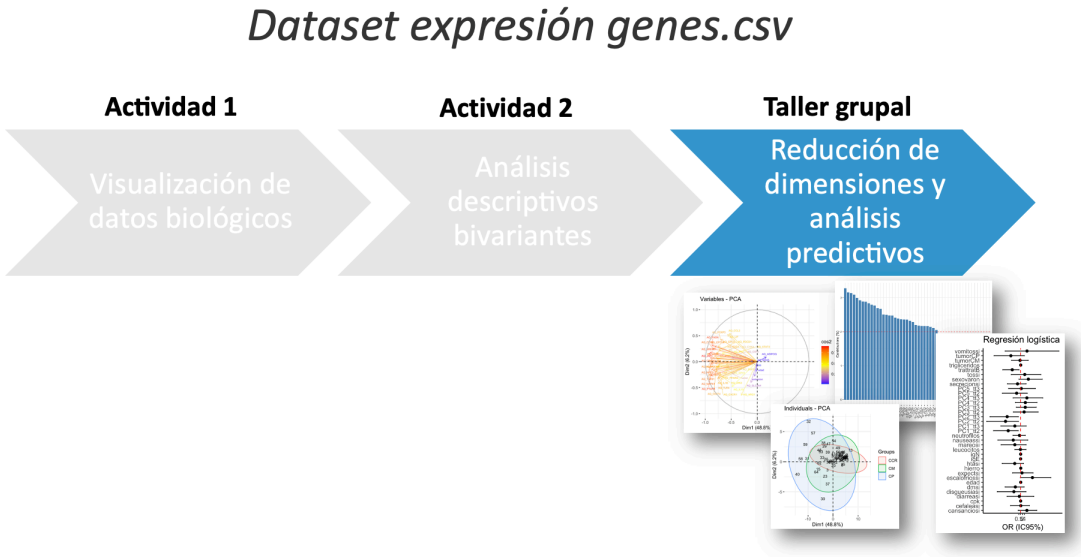
Datos del alumno

Fecha

Apellidos:

Nombre:

Actividad. Análisis de un caso práctico en R



Objetivos

Por medio de esta actividad aplicarás los conceptos aprendidos durante toda la materia (visualización gráfica, análisis descriptivos, contrastes de hipótesis, modelos inferenciales, PCA…). Utilizarás un *dataset* que contiene información de la expresión de 46 genes en 65 pacientes, cada uno con distintos tipos de tratamiento y características tumorales. Para ello realizarás las siguientes tareas:

1. Abrir la base de datos
2. PCA
3. Gráficos descriptivos
4. Tabla descriptiva
5. Modelo predictivo de regresión logística

Asignatura
Estadística y R para Ciencias de la Salud

Datos del alumno
Fecha
Apellidos:
Nombre:

Pautas de elaboración

Esta actividad consistirá en dos partes principales relacionado con datos de diferentes genes: 1) elaboración **de código** para que realices un caso real práctico en el documento HTML; y 2) **breve interpretación de los resultados finales** del *dataset* en el documento HTML.

A continuación verás: un apartado que explica el *dataset*, mientras que el siguiente son las preguntas que deberás de contestar.

***Dataset* de expresión de genes**: para la realización de esta parte de la actividad, deberás cargar el *dataset* de interés «Dataset expresión genes.csv», el cual se trata de una base de datos de 65 pacientes que contiene información de la expresión de 46 genes con diferentes funciones (para más información, ver el apartado de **Información de interés del dataset** después de la rúbrica). Además de estas variables, contiene otras variables de interés como el **tratamiento** (A o B) que siguen cada paciente, **tipo de tumor** que tienen (colorrectal, pulmón y mama) y la **extensión tumoral** (localizado, metastásico o regional). Por último, se recoge información de variables **bioquímicas**, **síntomas** y **otras variables** sociodemográficas.

Tras importarlo, deberás **responder a las siguientes cuestiones**, para realizar el PCA, gráficos descriptivos, tabla descriptiva y los modelos predictivos de regresión logística usando las librerías y comandos vistas en clase como `base`, `stats`, `factoextra`, `pheatmap`, `gtsummary`:

1. Abrir, explorar y preprocesar la base de datos:
    - Utiliza las funciones vistas para cargar, explorar y realizar la base de datos en formato CSV.

2. Aplicar un PCA:
   - Utiliza la librería correspondiente para realizar el PCA de los datos de expresión génica (como consejo, coger al menos aquellos componentes que explique un **70% de la varianza de los datos**). Para ello, tendrás que dejar bien claro cada uno de los pasos que se vieron en el temario y en la clase, creando tablas o figuras de cada uno de ellos (si se opta por tablas, **reflejarlas en una tabla modelo adjuntada al final de documento en la sección «Extensión y formato»).** Puedes apoyarte en el siguiente enlace para mejorar tus análisis: https://rpubs.com/Cristina_Gil/PCA

3. Crear gráficos descriptivos de los componentes principales:
   - Utiliza funciones vistas en el temario y clase para crear gráficos que representen visualmente los resultados del PCA, incluido gráficos que aporten información relevante a los resultados. Además, asegúrate de etiquetar adecuadamente los ejes y títulos de los gráficos para facilitar la interpretación. Puedes apoyarte en el siguiente enlace para mejorar tus análisis: https://rpubs.com/Cristina_Gil/PCA

4. Crear una tabla descriptiva con las variables más importantes:
   - Crea una tabla que incluya las estadísticas descriptivas de los valores sin transformar (media + desviación estándar si son paramétricas, mediana + rango intercuartílico (p25-p75) si no lo son) por terciles de cada componente del PCA (ver modelo de Tabla descriptiva adjuntada al final de documento en la sección **«Extensión y formato»**). Para calcular los terciles de un conjunto de datos, primero se determinan los puntos de corte que dividen el conjunto en tres partes iguales. Utilizando la función `quantile`, se calculan los valores en los que el 33.33% y el 66.67% de los datos se encuentran por debajo, lo que nos da los primeros y segundos terciles, respectivamente. Luego, para asignar a cada dato una categoría de tercil, se utiliza la función `cut`. Esto clasifica los datos en tres grupos según estos puntos de corte, etiquetándolos

como «t1», «t2» o «t3» para los primeros, segundos y terceros terciles. Así, cada dato en la columna PC1 se categoriza en uno de los tres terciles basándose en su valor relativo.

- Consejos:
  - Las tablas tienen que ser legibles, entendibles, ordenadas y limpias. Si hay decimales, lo normal es poner 1, a excepción de valores muy bajos que puede extenderse a los que se consideren para poder entenderse el número. Si el valor numérico es muy pequeño, puede optarse por usar el formato científico (por ejemplo: $2*10^{-6}$). Los valores P suele ponerse 3 decimales.
  - Lo más rápido es hacer las tablas descriptivas con la librería `gtsummary`. Para ello apóyate en lo visto en clase. Además, puedes ver aquí ejemplos y explicaciones: https://www.danieldsjoberg.com/gtsummary/. Si las generas con esta librería, no hace falta que generes las tablas modelo 2 y 3.
  - Si optas por `gtsummary`, tendrás que usar en primer lugar la función `tbl_summary` con `by`, `statistics`, `type` y `digits`; y `add_p` con `test` y `pvalue_fun`.
  - Si no optas por `gtsummary`, mi recomendación es que crees *datasets* independientes y que saques los descriptivos para reflejarlo en la tabla de anexos.

5. Implementar un modelo de regresión logística:
   - Utiliza la función vista en clase para construir el **modelo de regresión** logística, donde la **variable resultado es metástasis (sí/no)** y las variables predictoras son los terciles de los componentes principales obtenidos del PCA y otras variables de ajuste relevantes (pueden ser sociodemográficas o clínicas). Crea una tabla o gráfico con los datos de la regresión logística utilizando varios modelos de ajuste que sean lógicos y razonables. Importante, ten en cuenta los requisitos que había que hacer para la

identificación de variables confusoras. El formato de la tabla puedes guiarte tal y como se puede ver en la **Tabla de regresión logística** de los anexos.

- Utiliza las funciones específicas vistas en el temario para evaluar la calidad del modelo, además de las funciones específicas para sacar los parámetros (coeficientes exponenciados, IC 95 %, valores p) de cada variable introducida en el modelo.
- Basándote en los resultados obtenidos, elabora un informe de 1 página como máximo sobre: que conclusiones sacas del análisis del caso práctico en el HTML después de los análisis.

Extensión y formato

Para el informe grupal donde tienes que reflejar los hallazgos y el análisis, deberás entregar 3 archivos organizados de manera clara y concisa, destacando los puntos clave y las interpretaciones más relevantes. Los 3 archivos para entregar serán:

- Un único **fichero R Markdown** (.Rmd) con todo el código y texto Markdown que hayas generado. No existe un límite de extensión para este fichero. Hay que indicar los siguientes argumentos:
  - ***title:*** Resolución Actividad 1 máster Bioinformática UNIR (2023)
  - ***author:*** Nombre y 2 apellidos
  - ***date:*** yyyy-mm-dd
  - ***output:*** *html_document*

  Para crear código y generarlo en el fichero R Markdown poner siempre:

  ```
  ```{r}
  empezar aqui con el código
  ```
  ```

- Un **fichero HTML** generado a partir de dicho archivo R Markdown. Asegúrate de que aparecen todas las figuras, el texto correspondiente y todos los cuadros de código R que hayas introducido en el archivo .Rmd anterior.

Asignatura Datos del alumno
Fecha
Estadística y R para Ciencias de la Salud Apellidos:
Nombre:

- Un fichero Word únicamente con las tablas modelos (no hay extensión máxima para esta parte) y con el informe de 1 página como máximo. Asegúrate de que aparecen los títulos correspondientes de las tablas, las notas a pie de tabla, indicando en todo caso abreviaturas en orden alfabético, test utilizados para sacar las estadísticas, notas a pie de tabla que se desee reflejar para aclarar cualquier asunto relevante de la tabla, etc. Los valores numéricos se representan con 1 decimal, a excepción de los valores p que se representan con 3 decimales. Los datos descriptivos se representan en una misma celda la media y desviación estándar [por ejemplo: 3.4 (3)] al igual que la mediana y el RIQ [3.4 (1.3 a 9.8)]. El formato de Tablas que se debe de seguir es el siguiente:

Tabla PCA componentes y R^2. {Aquí poner título}

Componente	R^2
XX	XX
XX	XX
XX	XX

{Aquí poner pie de tabla}

Tabla PCA cargas. {Aquí poner título}

Variable	Componente 1	Componente "n"
Variable 1	XX	XX
Variable 2	XX	XX
...	XX	XX

{Aquí poner pie de tabla}

Asignatura Datos del alumno

Fecha

Estadística y R para Ciencias de la Salud Apellidos:

Nombre:

Tabla descriptiva. {Aquí poner título}

	Componente 1				Componente "n"			
	T1	T2	T3	Valor p	T1	T2	T3	Valor p
N	XXX	XXX	XXX		XXX	XXX	XXX	
Gen 1	XX (XX)	XX (XX)	XX (XX)	XXX	XX (XX)	XX (XX)	XX (XX)	XXX
Gen 2	XX (XX)	XX (XX)	XX (XX)	XXX	XX (XX)	XX (XX)	XX (XX)	XXX
...	XX (XX)	XX (XX)	XX (XX)	XXX	XX (XX)	XX (XX)	XX (XX)	XXX

{Aquí poner pie de tabla}

Tabla de regresión logística. {Aquí poner título}

	T1		T2		T3	
	OR (IC 95%)	P value	OR (IC 95%)	P value	OR (IC 95%)	P value
Terciles componente 1	1 (Ref.)	NA	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX
Terciles componente "n"	1 (Ref.)	NA	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX
Var ajuste 1	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX
Var ajuste 2	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX
...	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX	XX.X (XX a XX)	XXX

{Aquí poner pie de tabla}

Rúbrica

Título de la actividad	Descripción	Puntuación máxima (puntos)	Peso %
Análisis	Resolución de las cuestiones planteadas de forma correcta y ordenada	6	60%
Justificación	Justificación detallada en la respuesta a las preguntas planteadas y en el análisis de los gráficos generados	2	20%
Gráficos	Construcción correcta de los gráficos tal y como se solicitan en el enunciado de la actividad	1	10%
Originalidad	Generación de un informe R Markdown original, empleando temas o estilos CSS, índices, etc.	1	10%
		10	**100 %**