Segundo Parcial

21 de abril de 2023

Indicaciones generales

- Este es un examen **individual** con una duración de **180 minutos: de 7:00 a 10:00 a.m.**.
- Las respuestas deben estar totalmente justificadas.
- Se debe entregar un documento (word o PDF) con las respuestas a todos los ejercicios acompañado del archivo .R con el código. NO recibo parciales sólo con el código de R.
- ¡Suerte y ánimo!

1. (50 pts) Penguins es un conjunto de datos que contiene información acerca de tres especies diferentes de pingüinos respecto a diferentes variables, coleccionados a partir de tres islas en el archipiélago Palmer, en la Antártida (1=Torgesen, 2=Biscoe, 3=Dream). Las variables que se midieron fueron la longitud del pico (bill length), la profundidad del pico (bill depth) y la logintud de la aleta (ipper length) para tres especies diferentes de pingüinos (1=Adelie, 2=Gentoo, 3=Chinstrap). Cargue este conjunto de datos con load("penguins.RData").

   a) (20 pts) Haga una regresión lineal para predecir la variable bill length a partir de las otras.

   - ¿Qué variables son significativas para explicar la variable bill length? En caso de que no lo sean todas, ¿cuál es el modelo final en el que todas las variables son significativas?
   - Si quisiera predecir la variable bill lenght con una sola variable, ¿cuál debería elegir?

   b) (10 pts) ¿Hay diferencias físicas significativas entre los pingüinos de las diferentes islas?

   c) (20 pts) Haga un análisis de componentes principales.

   - ¿Con cuantas componentes principales se puede explicar más del 80 % de la variabilidad total de los datos?
   - ¿Qué variable tiene más influencia en la variabilidad explicada por la primera componente principal?

2. (50 pts) La base de datos MLBstats tiene información sobre los diferentes equipos de la liga de baseball de Estados Unidos. Las variables son: salario medio de cada equipo, porcentaje de victorias, porcentaje de bateo promedio, promedio de carreras, errores totales y si clasificaron o no a la postemporada (1=sí, 0=no). Cargue la base de datos asegurándose que la variable salary sea numérica.

   a) (40 pts) Haga un análisis de componentes principales con las variables estandarizadas y sin estandarizar

   - ¿Cuál de los dos análisis es más adecuado en este caso? ¿Por qué?
   - ¿Con cuántas componentes recomienda quedarse? ¿Por qué?

- Grafique los scores (valores de las observaciones en las componentes principales) en las dos primeras componentes. ¿Qué equipos son los más similares? ¿Y los más diferentes? Justifique su respuesta.

b) (10 pts) Haga una regresión lineal con para predecir el salario promedio dependiendo de las otras variables.

- ¿Qué porcentaje de la variabilidad de Y explica el modelo planteado?
- ¿Todas las variables son significativas?