A4 - Análisis de la varianza y repaso del curso

Enunciado

Semestre 2021.2

Índice

Introducción

El conjunto de datos es `gpa.csv`. Este conjunto de datos contiene la nota media de estudiantes universitarios tras el primer semestre de clases (GPA: grade point average, en inglés), así como información sobre la nota de acceso, la cohorte de graduación en el instituto y algunas características de los estudiantes.

Este conjunto de datos surge de una encuesta realizada a una muestra representativa de estudiantes de una universidad de EEUU (por razones de confidencialidad el conjunto de datos no incluye el nombre de la universidad). Las variables incluidas en el conjunto de datos son:

- sat: nota de acceso (medida en escala de 400 a 1600 puntos)
- tothrs: horas totales cursadas en el semestre
- colgpa: nota media del estudiante al final del primer semestre (medida en escala de 0 a 4 puntos)
- athlete: indicador de si el estudiante practica algún deporte en la universidad
- hsize: numero total de estudiantes en la cohorte de graduados del bachillerato (en cientos)
- hsrank: ranking del estudiante, dado por la nota media del bachillerato, en su cohorte de graduados del bachillerato
- hsperc: ranking relativo del estudiante, en porcentaje (hsrank/hsize)
- female: indicador de si el estudiante es mujer
- white: indicador de si el estudiante es de raza blanca o no
- black: indicador de si el estudiante es de raza negra o no

El objetivo de esta actividad final es doble. En primer lugar, consolidar los conocimientos y competencias de preprocesado, análisis descriptivo, inferencia estadística y análisis de regresión. En segundo lugar, adquirir los conocimientos y competencias para llevar a cabo un análisis tipo ANOVA (análisis de la varianza).

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.

---

1. Lectura del fichero y preparación de los datos

Leed el fichero `gpa.csv` y guardad los datos en un objeto denominado `gpa`. A continuación, verificad el tipo de cada variable. ¿Qué variables son de tipo numérico? ¿Qué variables son de tipo cualitativo?

1.1. Preparación de los datos

La variable `tothrs` está clasificada como character. Para poder trabajar con ella hay que convertirla en numérica, eliminando el texto "h" de los datos.

1.2. Valores ausentes

- Comprobad cuántas observaciones tienen valores ausentes y sacad conclusiones sobre cómo de serio es el problema de valores ausentes en estos datos.
- Eliminad los valores ausentes del conjunto de datos. Denominad al nuevo conjunto de datos 'gpaclean'. Nota: En el resto de apartados se usará el nuevo conjunto de datos 'gpaclean'.

1.3. Equivalencia de la nota en letras

La variable `colgpa` contiene la nota numérica del estudiante. Cread una variable categórica denominada `gpaletter`, que indique la nota en letra de cada estudiante de la siguiente forma: A, de 3.50 a 4.00; B, de 2.50 a 3.49; C, de 1.50 a 2.49; D, de 0 a 1.49.

---

2. Estadística descriptiva y visualización

2.1. Análisis descriptivo

Realizad un análisis descriptivo numérico de los datos (resumid los valores de las variables numéricas y categóricas). Mostrad el número de observaciones y el número de variables.

2.2. Visualización

1. Mostrad con diversos diagramas de caja (boxplot) la distribución de la variable 'sat' según la variable 'female', según 'athlete', y según 'gpaletter'.
2. Cread una variable denominada 'excelente' que indique si el estudiante ha obtenido una A de nota media al final del semestre. Esta nueva variable debe codificarse como una variable dicotómica que toma el valor 1 cuando el estudiante ha obtenido una A, y el valor 0 en caso contrario. Realizad un gráfico que muestre el porcentaje de estudiantes excelentes.
3. Interpretad los gráficos brevemente.

---

3. Estadística inferencial

Utilizamos el conjunto de datos `gpaclean`.

3.1. Intervalo de confianza de la media poblacional de la variable `sat`

a) Calculad manualmente el intervalo de confianza al 95 % de la media poblacional de la variable `sat` de los estudiantes. Para ello, definid una función IC que reciba la variable, la confianza, y que devuelva un vector con los valores del intervalo de confianza. No se pueden utilizar funciones como `t.test` o `z.test` para el cálculo. Sí podéis usar otras funciones básicas de R como *mean*, *qnorm*, *qt*, *pnorm*, *pt*, etcétera.

A partir del resultado obtenido, explicad cómo se interpreta el intervalo de confianza.

b) Calculad los intervalos de confianza al 95 % de la media poblacional de la variable `sat`, en función de si los estudiantes son hombres o mujeres. ¿Qué conclusión se puede extraer de la comparación de los

dos intervalos, en relación a si existe solapamiento o no en los intervalos de confianza? Justificad la respuesta.

3.2. Contraste de hipótesis para la diferencia de medias de `colgpa`

Queremos analizar si la nota media del primer semestre es diferente para las mujeres que para los hombres utilizando un nivel de confianza del 95 %.

Nota: se deben realizar los cálculos manualmente. No se pueden usar funciones de `R` que calculen directamente el contraste como `t.test` o similar. Sí se puede usar `var.test` y funciones como `mean`, `sd`, `qnorm`, `pnorm`, `qt` y `pt`.

Seguid los pasos que se detallan a continuación.

3.2.1. Pregunta de investigación

Formulad la pregunta de investigación.

3.2.2. Escribid la hipótesis nula y la alternativa

3.2.3. Justificación del test a aplicar

3.2.4. Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95 %.

3.2.5. Interpretación del test

---

4. Modelo de regresión lineal

Estimad un modelo de regresión lineal múltiple que tenga como variables explicativas: `sat`, `female`, `tothrs`, `athlete`, y `hsperc`, y como variable dependiente `colgpa`.

4.1. Interpretación del modelo

Interpretad el modelo lineal ajustado:

– ¿Cuál es la calidad del ajuste?

– Explicad la contribución de las variables explicativas.

4.2. Predicción

Independientemente del R^2 obtenido en el apartado previo, aplicad el modelo de regresión para predecir la nota media de un estudiante hombre, atleta, con una nota de entrada de 800, un total de horas en el semestre de 60 y una posición relativa en el ranking del 60 %.

---

5. Regresión logística

5.1. Estimación del modelo

Estimad un modelo logístico para predecir la probabilidad de ser un estudiante excelente al final del primer semestre en la universidad en función de las variables: `female`, `athlete`, `sat`, `tothrs`, `black`, `white` y `hsperc`.

5.2. Interpretación del modelo estimado

Interpretad los resultados obtenidos. Concretamente, analizad la significatividad de las variables explicativas y explicad su contribución para predecir la probabilidad de ser un estudiante excelente.

5.3. Importancia de ser mujer

En el modelo anterior, interpretad los niveles de la variable `female` a partir del **odds ratio**. ¿En qué porcentaje se ve aumentada la probabilidad de ser un estudiante excelente si se es mujer? Proporcionad intervalos de confianza del 95 % de los odds ratio.

5.4. Predicción

¿Con que probabilidad una estudiante mujer, no atleta, con un sat de 1200 puntos, 50 horas cursadas, de raza negra y con un ranking relativo (`hsperc`) del 10 % será excelente?

---

6. Análisis de la varianza (ANOVA) de un factor

Vamos a realizar un ANOVA para contrastar si existen diferencias en la variable `colgpa` en función de la raza de los estudiantes. Seguid los pasos que se indican.

En primer lugar, a partir de las variables `black` y `white` cread una variable categórica denominada `race`, que indique la raza del estudiante en una de estas tres categorías: `black`, `white` y `other` (para estudiantes que no son de raza negra ni blanca).

6.1. Visualización gráfica

Mostrad gráficamente la distribución de `colgpa` según los valores de `race`.

6.2. Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

6.3. Modelo

Calculad el análisis de varianza, usando la función `aov` o `lm`. Interpretad el resultado del análisis, teniendo en cuenta los valores: Sum Sq, Mean SQ, F y Pr (> F).

6.4. Efectos de los niveles del factor

Proporcionad la estimación del efecto de los niveles del factor `race`. Calculad también la parte de la variabilidad de `colgpa` explicada por el efecto de los niveles.

6.5. Conclusión de los resultados del ANOVA

Sacad conclusiones del ANOVA realizado.

6.6. Normalidad de los residuos

Usad el gráfico Normal Q-Q y el test Shapiro-Wilk para evaluar la normalidad de los residuos. Podéis usar las funciones de R correspondientes para hacer el gráfico y el test.

6.7. Homocedasticidad de los residuos

El gráfico "Residuals vs Fitted" proporciona información sobre la homcedasticidad de los residuos. Mostrad e interpretad este gráfico.

---

7. ANOVA multifactorial

A continuación, se desea evaluar el efecto sobre `colgpa` de la raza del estudiante combinada con el factor género del estudiante (`female`). Seguid los pasos que se indican a continuación.

7.1. Análisis visual de los efectos principales y posibles interacciones

Representad la interacción de los dos factores `race` y `female` y comentad los gráficos resultantes.

7.2. Cálculo del modelo

Calculad el modelo ANOVA multifactorial. Podéis usar la función **aov**.

7.3. Interpretación de los resultados

Interpretad los resultados obtenidos.

7.4. Adecuación del modelo

Interpretad la adecuación del modelo ANOVA obtenido usando los gráficos de los residuos.

---

8. Conclusiones

Resumid las conclusiones principales del análisis (apartados 3 a 7). Para ello, podéis resumir las conclusiones de cada uno de los apartados.

Puntuación de la actividad

- Apartados 1 y 2 (10 %)
- Apartado 3 (10 %)
- Apartado 4 (10 %)
- Apartado 5 (15 %)
- Apartado 6 (20 %)
- Apartado 7 (15 %)
- Apartado 8 (10 %)
- Calidad del informe dinámico (10 %)