Análisis estadístico - Actividad 4

Solución

Semestre 2023.2

Índice

Introducción

Los datos utilizados corresponden a iFood, la principal aplicación de entrega de alimentos en Brasil. La empresa vende alimentos en diversas categorías y busca mejorar el rendimiento de las actividades de marketing. La nueva campaña comercial, sexta, tiene como objetivo vender un nuevo gadget a la base de datos de clientes. Se llevó a cabo una campaña piloto en la que participaron 2.240 clientes. Los clientes fueron seleccionados al azar y contactados por teléfono para la adquisición del gadget.

El conjunto de datos de está practica se denomina `marketing.csv`. Las variables que contiene son:

1. ID: número identificativo

2. Year_Birth: Año de nacimiento
3. Education: el nivel educativo del cliente (factor con 5 niveles)
4. Marital_Status: el estado civil del cliente (factor con 8 niveles)
5. Income: ingresos anuales del cliente
6. Kidhome: número de niños que habitan con el cliente
7. Teenhome: número de adolescentes que habitan con el cliente
8. Dt_Customer: fecha de alta del cliente en la empresa
9. Recency: número de días desde la última compra
10. MntWines: cantidad gastada en vino en los últimos 2 años
11. MntFruits: cantidad gastada en fruta en los últimos 2 años
12. MntMeatProducts: cantidad gastada en carne en los últimos 2 años
13. MntFishProducts: cantidad gastada en pescado en los últimos 2 años
14. MntSweetProducts: cantidad gastada en dulces en los últimos 2 años
15. MntGoldProds: cantidad gastada en productos "gold" en los últimos 2 años
16. NumDealsPurchases: número de compras hechas con descuento
17. NumWebPurchases: número de compras hechas a través de la Web
18. NumCatalogPurchases: número de compras hechas usando el catálogo
19. NumStorePurchases: número de compras hechas directamente en tiendas
20. NumWebVisitsMonth: número de visitas a la Web en el último mes
21. AcceptedCmp3: 1 si el cliente acepta la oferta en la 3º campaña, 0 si no
22. AcceptedCmp4: 1 si el cliente acepta la oferta en la 4º campaña, 0 si no
23. AcceptedCmp5: 1 si el cliente acepta la oferta en la 5º campaña, 0 si no
24. AcceptedCmp1: 1 si el cliente acepta la oferta en la 1º campaña, 0 si no
25. AcceptedCmp2: 1 si el cliente acepta la oferta en la 2º campaña, 0 si no
26. Complain: 1 si el cliente formaliza una queja en el último año
27. Z_CostContact: variable control (se debe excluir del análisis)
28. Z_Revenue: variable control (se debe excluir del análisis)
29. Response: 1 si el cliente acepta la oferta en la última campaña, 0 si no

El conjunto de datos original se encuentra disponible en: https://github.com/nailson/ifood-data-business-analyst-test/blob/master/ifood_df.csv

El objetivo final es desarrollar un modelo que permita identificar a los clientes según sus características. En esta actividad se analizará si los ingresos de los clientes están determinados por el nivel educativo y otras características. Para hacerlo, se aplican diferentes tipos de análisis, revisando el contraste de hipótesis de dos muestras, vistos en la actividad A2, y después realizando análisis más complejos como ANOVA.

Notas importantes a tener en cuenta para la entrega de la actividad:

- Es necesario entregar el fichero Rmd y el fichero de salida (PDF o html). El fichero de salida tiene que incluir el código y el resultado de su ejecución (paso a paso). Se tiene que incluir un índice o tabla de contenidos. Y se tiene que respetar la numeración de los apartados del enunciado.
- No realicéis listados de los conjuntos de datos, puesto que estos pueden ocupar varias páginas. Si queréis comprobar el efecto de una instrucción sobre un conjunto de datos podéis usar la función **head** y **tail** que muestran las primeras o últimas filas del conjunto de datos.

---

1 Preprocesado

1.1 Variables Income y Year_Birth

Cargad el fichero de datos "marketing.csv". Consultad los tipos de datos de las variables y si es necesario, aplicad las transformaciones apropiadas. Averiguad posibles inconsistencias en los valores de `Income` y `Year_Birth`. En caso de que existan inconsistencias, sustituir los valores por valores perdidos.

1.2 Valores ausentes

En este apartado, realizaremos tratamiento para valores ausentes. Adoptaremos la siguiente estrategia:

- En caso de valores perdidos de la variable Income, aplicad imputación por vecinos más cercanos, utilizando la distancia de Gower, considerando en el cómputo de los vecinos más cercanos las variables numéricas que representan el gasto en los distintos productos (variables de la 10 a la 15). Para realizar esta imputación, se puede usar la función "kNN" de la librería VIM con un número de vecinos igual a 5.
- En caso de valores perdidos de la variable Year_Birth, aplicad imputación considerando el valor mediano de la variable Year_Birth entre las personas encuestadas en estado civil "Viudo/a"
- Eliminad las observaciones con valores ausentes para el resto de variables del conjunto de datos. Denominad al nuevo conjunto de datos markclean.
- Comprobad cuántas observaciones tienen valores ausentes y reflexionad brevemente sobre cómo de preocupante es el problema de valores ausentes en estos datos.

2 Estadística descriptiva

2.1 Income

El coeficiente de variación (`Desviación estándar/valor absoluto de la media`) se utiliza para analizar si la media es representativa. Un coeficiente de variación más pequeño a 1 se considera que la media es un valor representativo del conjunto de datos.

Calcula el coeficiente de variación de la variable `Income`. En base a los resultados obtenidos, ¿pensáis que el ingreso medio es un valor representativo de la distribución de ingresos? A partir de métodos visuales ¿podemos asumir que la variable tiene una distribución normal? Justificad la respuesta.

2.2 Education

Mostrad una tabla con los estadísticos descriptivos (media, número de observaciones y desviación típica) de los ingresos según el nivel educativo. Mostrar un box plot de los ingresos según el nivel educativo. ¿Que podemos decir?

Nota: La tabla de descriptivos y el boxplot las categorías del nivel educativo deben mostrarse ordenadas de mayor a menor nivel educativo. Para calcular la media u otras medidas por cada nivel educativo, podéis utilizar las funciones **summarize** y **group_by** de la librería **dplyr**.

3 Estadística inferencial

3.1 Contraste de hipótesis para la diferencia de medias

¿Podemos aceptar que los **ingresos medios de las personas sin estudios universitarios son inferiores a los de las personas con estudios universitarios?** Responded a la pregunta utilizando un nivel de confianza del 99%.

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

Seguid los pasos que se detallan a continuación.

3.1.1 Escribid la hipótesis nula y la alternativa

3.1.2 Justificación del test a aplicar

3.1.3 Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y p valor a un nivel de confianza del 99%.

3.1.4 Interpretación del test

---

4 Modelo de regresión

4.1 Regresión lineal múltiple

Queremos investigar qué variables explican los ingresos de los individuos (Income). Estimad un modelo de regresión lineal múltiple que tenga como variables explicativas: Year_Birth, Kidhome, Teenhome, Education, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchases, NumCatalogPurchases, NumStorePurchases y NumWebVisitsMonth.

Interpretad el modelo lineal ajustado. ¿Como es la calidad del ajuste? Explicad brevemente la contribución de las variables explicativas en el modelo.

Nota: En la variable Education, la categoría de referencia debe ser **Basic**.

4.1.1 Multicolinealidad

Analizad posibles problemas de multicolinealidad (alta correlación entre variables explicativas) mediante la interpretación del factor de inflación de la varianza (vif). Se puede utilizar la función **vif** de la libreria **car**.

4.2 Regresión logística

4.2.1 Modelo predictivo

Ajustad un modelo predictivo basado en la regresión logística para predecir la probabilidad de aceptar la oferta en la sexta campaña en función del número de compras con descuento, el número de vísitas en último mes a la web y si ha aceptado alguna oferta en campañas previas. Mostrad el resultado del modelo e interpretad el modelo en términos de: cuáles son las variables significativas y como es la calidad del modelo.

4.2.2 Matriz de confusión

A continuación analizad la precisión del modelo, comparando la predicción del modelo sobre los mismos datos del conjunto de datos. Asumiremos que la predicción del modelo es 1 (Response) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 en caso contrario. Calculad la matriz de confusión. Interpretad los resultados. Indicad los valores de sensibilidad y especificidad e interpretadlos. Se puede utilizar función confusionMatrix de la libreria **Caret**.

4.2.3 Predicción

Aplicad el modelo de regresión logística para predecir la probabilidad que acepte última oferta teniendo en cuenta que ha comprado 5 veces con descuento, ha visitado la web 10 veces y ha aceptado todas las ofertas de campañas previas. Haced los cálculos sin usar función `predict`. Utilizad la función `predict` para comprobar resultado.

---

5 ANOVA unifactorial

A continuación se realizará un análisis de varianza, donde se desea comparar los ingresos para los distintos niveles educativos. El análisis de varianza consiste a evaluar si la variabilidad de una variable dependiente puede explicarse a partir de una o varias variables independientes, denominadas factores. En el supuesto que nos ocupa, nos interesa evaluar si la variabilidad de la variable `Income` puede explicarse por el nivel educativo.

Hay dos preguntas básicas a responder:

- ¿Existen diferencias en los ingresos (Income) entre los diferentes niveles educativos?
- Si existen diferencias, ¿entre qué niveles educativos se dan estas diferencias?

5.1 Visualización gráfica

Para completar el boxplot del apartado 2.2, mostrad gráficamente la distribución de `Income` según `Nivel educativo` representando los valores medios para cada categoría. Se puede utilizar la función ggline de la libreria **ggpubr**.

5.2 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

5.3 Modelo

Calculad el análisis de varianza, usando la función `aov` o `lm`. Interpretad el resultado del análisis.

5.4 Efectos de los niveles del factor y fuerza de relación

Proporcionad la estimación del efecto de los niveles del factor `Education`. Interpretad los resultados.

Calculad la parte de la variabilidad de ingresos explicada por el efecto de los niveles (fuerza de relación). Es decir, calculad $\eta^2 = \frac{SSB}{SST}$ del modelo. Interpretad los resultados.

5.5 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.

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

6 Comparaciones múltiples

Independientemente del resultado obtenido en el apartado anterior, realizad un test de comparación múltiple entre los grupos con corrección de Bonferroni. Este test se aplica cuando el test ANOVA devuelve rechazar la hipótesis nula de igualdad de medias. Por lo tanto, procederemos como si el test ANOVA hubiera dado como resultado el rechazo de la hipótesis nula.

Calculad las comparaciones entre grupos con la corrección Bonferroni. Podéis utilizar la función **pairwise.t.test**. Interpretad los resultados.

7 ANOVA multifactorial

A continuación, se desea evaluar el efecto sobre `Income` del nivel educativo combinado con si acepta o no la oferta de la última campaña. Seguid los pasos que se indican a continuación.

7.1 Análisis visual de los efectos principales y posibles interacciones

Dibujad en un gráfico la variable `Income` en función de `Education` y en función de `Response`. El gráfico tiene que permitir evaluar si hay interacción entre los dos factores. Por eso, se recomienda seguir estos pasos:

1. Agrupad el conjunto de datos por `Education` y por `Response`. Calculad la media de ingresos para cada grupo. Mostrad el conjunto de datos en forma de tabla (data frame), donde se muestre la media de cada grupo según `Education` y `Response`.
2. Mostrad en un gráfico el valor medio de la variable `Income` para cada factor. Interpretad el resultado sobre si solo hay efectos principales o hay interacción entre los factores. Si hay interacción, explicad cómo se observa esta interacción en el gráfico.

7.2 Cálculo del modelo

Con el análisis ANOVA multifactorial se comprobará si existe interacción entre los factores `Education` y `Response` en relación a la variable `Income`. Podéis usar la función **aov**.

7.3 Interpretación de los resultados

---

8 Resumen ejecutivo

Escribid un resumen ejecutivo como si tuvierais que comunicar a una audiencia no técnica. Por ejemplo, podría ser un equipo de managers, a los cuales se los tiene que informar sobre las diferencias en los ingresos de los clientes, su nivel educativo y la propensión que tiene a aceptar la oferta de la última campaña.

Puntuación de los apartados

- Pregunta 1: 10%
- Preguntas 2 y 3: 10%
- Pregunta 4: 10%
- Pregunta 5: 30%
- Pregunta 6: 10%
- Pregunta 7: 20%
- Pregunta 8: 10%