Anàlisi estadística - Activitat 4

Semestre 2023.2

Índex

Introducció

Les dades utilitzades corresponen a iFood, la principal aplicació de lliurament d'aliments al Brasil. L'empresa ven aliments en diverses categories i cerca millorar el rendiment de les activitats de màrqueting. La nova campanya comercial, sisena, té com a objectiu vendre un nou gadget a la base de dades de clients. Es va dur a terme una campanya pilot a la que van participar 2.240 clients. Els clients van ser seleccionats a l'atzar i contactats per telèfon per a l'adquisició del gadget.

El conjunt de dades d'aquesta pràctica es denomina `màrqueting.csv`. Les variables que conté són:

1. ID: número identificatiu
2. Year_Birth: Any de naixement
3. Education: el nivell educatiu del client (factor amb 5 nivells)

4. Marital_Estatus: l'estat civil del client (factor amb 8 nivells)
5. Income: ingressos anuals del client
6. Kidhome: nombre de nens que habiten amb el client
7. Teenhome: nombre d'adolescents que habiten amb el client
8. Dt_Customer: data d'alta del client en l'empresa
9. Recency: nombre de dies des de l'última compra
10. MntWines: quantitat gastada en vi en els últims 2 anys
11. MntFruits: quantitat gastada en fruita en els últims 2 anys
12. MntMeatProducts: quantitat gastada en carn en els últims 2 anys
13. MntFishProducts: quantitat gastada en peix en els últims 2 anys
14. MntSweetProducts: quantitat gastada en dolços en els últims 2 anys
15. MntGoldProds: quantitat gastada en productes "*gold" en els últims 2 anys
16. NumDealsPurchases: nombre de compres fetes amb descompte
17. NumWebPurchases: nombre de compres fetes a través de la Web
18. NumCatalogPurchases: nombre de compres fetes usant el catàleg
19. NumStorePurchases: nombre de compres fetes directament en botigues
20. NumWebVisitsMonth: nombre de visites a la Web en l'últim mes
21. AcceptedCmp3: 1 si el client accepta l'oferta en la 3r campanya, 0 si no
22. AcceptedCmp4: 1 si el client accepta l'oferta en la 4t campanya, 0 si no
23. AcceptedCmp5: 1 si el client accepta l'oferta en la 5è campanya, 0 si no
24. AcceptedCmp1: 1 si el client accepta l'oferta en la 1r campanya, 0 si no
25. AcceptedCmp2: 1 si el client accepta l'oferta en la 2n campanya, 0 si no
26. Complain: 1 si el client formalitza una queixa en l'últim any
27. Z_CostContact: variable control (s'ha d'excloure de l'anàlisi)
28. Z_Revenue: variable control (s'ha d'excloure de l'anàlisi)
29. Response: 1 si el client accepta l'oferta en l'última campanya, 0 si no

El conjunt de dades original es troba disponible a: https://github.com/nailson/ifood-data-business-analyst-test/blob/master/ifood_df.csv

L'objectiu final és desenvolupar un model que permeti identificar als clients segons les seves característiques. En aquesta activitat s'analitzarà si els ingressos dels clients estan determinats pel nivell educatiu i altres característiques. Per a fer-ho, s'apliquen diferents tipus d'anàlisis, revisant el contrast d'hipòtesis de dues mostres, vist a l'activitat A2, i després realitzant anàlisis més complexes com ANOVA.

Notes importants a tenir en compte per al lliurament de l'activitat:

- És necessari lliurar el fitxer Rmd i el fitxer de sortida (PDF o html). El fitxer de sortida ha d'incloure el codi i el resultat de la seva execució (pas a pas). S'ha d'incloure un índex o taula de continguts. I s'ha de respectar la numeració dels apartats de l'enunciat.
- No realitzeu llistats dels conjunts de dades, ja que aquests poden ocupar diverses pàgines. Si voleu comprovar l'efecte d'una instrucció sobre un conjunt de dades podeu usar la funció **head** i **tail** que mostren les primeres o últimes files del conjunt de dades.

---

1 Preprocessament

1.1 Variables Income i Year_Birth

Carregueu el fitxer de dades "marketing.csv". Consulteu els tipus de dades de les variables i si és necessari, apliqueu les transformacions apropiades. Esbrineu possibles inconsistències en els valors de `Income` i `Year_Birth`. En cas que existeixin inconsistències, substituir els valors per valors perduts.

1.2 Valors absents

En aquest apartat, farem tractament per a valors absents. Adoptarem la següent estratègia:

- El tractament de la variable Income en cas de valors perduts, apliqueu imputació per veïns més pròxims, utilitzant la distància de Gower, considerant en el còmput dels veïns més pròxims les variables numèriques que representen la despesa en els diferents productes (variables de la 10 a la 15). Per a realitzar aquesta imputació, es pot usar la funció "kNN" de la llibreria VIM amb un nombre de veïns igual a 5.
- En cas de valors perduts de la variable Year_Birth, apliqueu imputació considerant el valor mitjà de la variable Year_Birth entre les persones enquestades en estat civil "Vidu/a"
- Elimineu les observacions amb valors absents per a la resta de variables del conjunt de dades. Denomineu al nou conjunt de dades markclean.
- Reviseu quantes observacions heu trobat valors absents i reflexioneu breument sobre com de preocupant és el problema de valors absents en aquestes dades.

2 Estadística descriptiva

2.1 Income

El coeficient de variació (`Desviació estàndar/valor absolut de la mitjana`) es fa servir per analitzar si la mitjana és representativa. Un coeficient de variació més petit a 1 es considera que la mitjana és un valor representatiu del conjunt de les dades. Calcula el coeficient de variació de la variable `Income`. A partir dels resultats obtinguts, penseu que l'ingrés mitjà és un valor representatiu de la distribució d'ingressos? A partir de mètodes visuals podem assumir que la variable té una distribució normal? Justifiqueu la resposta.

2.2 Education

Mostreu una taula amb els estadístics descriptius (mitjana, nombre d'observacions i desviació típica) dels ingressos segons el nivell educatiu. Mostreu un box plot dels ingressos segons el nivell educatiu. Què podem observar a partir d'aquest gràfic?

Nota: La taula de descriptius i el boxplot les categories del nivell educatiu han de mostrar-se ordenades de major a menor nivell educatiu. Per a calcular la mitjana o altres mesures per cada nivell educatiu, podeu utilitzar les funcions **summarize** i **group_by** de la llibreria **dplyr**.

3 Estadística inferencial

3.1 Contrast d'hipòtesi per a la diferència de mitjanes

Podem acceptar que els **ingressos mitjans de les persones sense estudis universitaris són inferiors als de les persones amb estudis universitaris?** Responeu a la pregunta utilitzant un nivell de confiança del 99%.

Nota: s'han de realitzar els càlculs manualment. No es poden usar funcions de `R` que calculin directament el contrast com a `t.test` o similar. Si que es poden usar funcions com '`mean`, `sd`, `qnorm`, `pnorm`, `qt` i `pt`.

Seguiu els passos que es detallen a continuació.

3.1.1 Escriviu la hipòtesi nul · la i l'alternativa

3.1.2 Justificació del test a aplicar

3.1.3 Càlculs

Realitzeu els càlculs de l'estadístic de contrast, valor crític i p valor a un nivell de confiança del 99%.

3.1.4 Interpretació del test

---

4 Model de regressió

4.1 Regressió lineal múltiple

Volem investigar quines variables expliquen els ingressos dels individus (Income). Estimeu un model de regressió lineal múltiple que tingui com a variables explicatives: Year_Birth, Kidhome, Teenhome, Education, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchases, NumCatalogPurchases, NumStorePurchases i NumWebVisitsMonth.

Interpreteu el model lineal ajustat. Com és la qualitat de l'ajust? Expliqueu breument la contribució de les variables explicatives en el model.

Nota: En la variable Education, la categoria de referència ha de ser **Basic**.

4.1.1 Multicolinealitat

Analitzeu possibles problemes de multicolinealitat (alta correlació entre variables explicatives) mitjançant la interpretació del factor d'inflació de la variància (vif). Es pot utilitzar la funció **vif** de la llibreria **car**.

4.2 Regressió logística

4.2.1 Model predictiu

Ajusteu un model predictiu basat en la regressió logística per a predir la probabilitat d'acceptar l'oferta en la sisena campanya en funció del nombre de compres amb descompte, el nombre de visites de l'últim mes a la web i si ha acceptat alguna oferta en campanyes prèvies. Mostreu el resultat del model i interpreteu el model en termes de: quines són les variables significatives i com és la qualitat del model.

4.2.2 Matriu de confusió

A continuació analitzeu la precisió del model, comparant la predicció del model sobre les mateixes dades del conjunt de dades. Assumirem que la predicció del model és 1 (Response) si la probabilitat del model de regressió logística és superior o igual a 0.5 i 0 en cas contrari. Calculeu la matriu de confusió. Interpreteu els resultats. Indiqueu els valors de sensibilitat i especificitat i interpreteu-los. Es pot utilitzar funció confusionMatrix de la llibreria **Caret**.

4.2.3 Predicció

Apliqueu el model de regressió logística per a predir la probabilitat que accepti última oferta tenint en compte que ha comprat 5 vegades amb descompte, ha visitat la web 10 vegades i ha acceptat totes les ofertes de campanyes prèvies. Feu els càlculs sense utilitzar la funció `predict`. Utilitzeu la funció `predict` per a comprovar el resultat.

---

5 ANOVA unifactorial

A continuació es realitzarà una anàlisi de variància, on es desitja comparar els ingressos per als diferents nivells educatius. L'anàlisi de variància consisteix a avaluar si la variabilitat d'una variable dependent pot explicar-se a partir d'una o diverses variables independents, denominades factors. En el supòsit que ens ocupa, ens interessa avaluar si la variabilitat de la variable `Income` pot explicar-se pel nivell educatiu.

Hi ha dues preguntes bàsiques a respondre:

- Existeixen diferències en els ingressos (Income) entre els diferents nivells educatius?
- Si existeixen diferències, entre quins nivells educatius es donen aquestes diferències?

5.1 Visualització gràfica

Per a completar el boxplot de l'apartat 2.2, mostreu gràficament la distribució de `Income` segons `Education` representant els valors mitjans per a cada categoria. Es pot utilitzar la funció ggline de la llibreria **ggpubr**.

5.2 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i l'alternativa.

5.3 Model

Calculeu l'anàlisi de variància, utilitzant la funció `aov` o `lm`. Interpreteu el resultat de l'anàlisi.

5.4 Efectes dels nivells del factor i força de relació

Proporcioneu l'estimació de l'efecte dels nivells del factor `Education`. Interpreteu els resultats.

Calculeu la part de la variabilitat dels ingressos explicada per l'efecte dels nivells (força de relació). És a dir, calculeu $\eta^2 = \frac{SSB}{SST}$ del model. Interpreteu els resultats.

5.5 Normalitat dels residus

Utilitzeu el gràfic Normal Q-Q i el test Shapiro-Wilk per a avaluar la normalitat dels residus. Podeu fer servir les funcions de R corresponents per a fer el gràfic i el test.

Homocedasticitat dels residus El gràfic "Residuals vs Fitted" proporciona informació sobre la homocedasticitat dels residus. Mostreu i interpreteu aquest gràfic.

6 Comparacions múltiples

Amb independència del resultat obtingut en l'apartat anterior, realitzeu un test de comparació múltiple entre els grups amb correcció de Bonferroni. Aquest test s'aplica quan el test ANOVA retorna rebutjar la hipòtesi nul · la d'igualtat de mitjanes. Per tant, procedirem com si el test ANOVA hagués donat com a resultat el rebuig de la hipòtesi nul · la.

Calculeu les comparacions entre grups amb la correcció Bonferroni. Podeu utilitzar la funció **pairwise.t.test**. Interpreteu els resultats.

7 ANOVA multifactorial

A continuació, es desitja avaluar l'efecte sobre `Income` del nivell educatiu combinat amb si accepta o no l'oferta de l'última campanya. Seguiu els passos que s'indiquen a continuació.

7.1 Anàlisi visual dels efectes principals i possibles interaccions

Dibuixeu en un gràfic la variable `Income` en funció de `Education` i en funció de `Response`. El gràfic ha de permetre avaluar si hi ha interacció entre els dos factors. Per això, es recomana seguir aquests passos:

1. Agrupeu el conjunt de dades per `Education` i per `Response`. Calculeu la mitjana d'ingressos per a cada grup. Mostreu el conjunt de dades en forma de taula (data frame), on es mostri la mitjana de cada grup segons `Education` i `Response`.

2. Mostreu en un gràfic el valor mitjà de la variable `Income` per a cada factor. Interpreteu el resultat sobre si només hi ha efectes principals o hi ha interacció entre els factors. Si hi ha interacció, expliqueu com s'observa aquesta interacció en el gràfic.

7.2 Càlcul del model

Analitzeu la interacció entre els factors `Education` i `Response` en relació a la variable `Income`. Podeu fer servir la funció **aov**.

7.3 Interpretació dels resultats

---

8 Resum executiu

Escriviu un resum executiu com si haguéssiu de comunicar a una audiència no tècnica. Per exemple, podria ser un equip de managers, als qui s'ha d'informar sobre les diferències en els ingressos dels clients, el seu nivell educatiu i la propensió que tenen a acceptar l'oferta de l'última campanya.

Puntuació dels apartats

- Pregunta 1: 10%
- Preguntes 2 i 3: 10%
- Pregunta 4: 10%
- Pregunta 5: 30%
- Pregunta 6: 10%
- Pregunta 7: 20%
- Pregunta 8: 10%

```
knitr::knit_exit()
```