A4 - Anàlisi de la variància i repàs del curs

Enunciat

Semestre 2019.2

Índex

1 Introducció

Les dades que es faran servir per aquesta activitat correspon a les vendes de seients de cotxes infantils a 400 botigues diferents. Les variables són:

- Sales (Vendes unitàries, en milers, a cada ubicació)
- CompPrice (Preu cobrat pel competidor a cada ubicació)
- Income (Nivell d'ingressos comunitaris, en milers de dòlars)
- Advertising (Pressupost de publicitat local de l'empresa a cada ubicació, en milers de dòlars)
- Population (Mida de la població a la regió, en milers)
- Price (Preu per seients de cotxes a cada lloc)
- ShelveLoc (Un factor amb nivells Bad, Good i Medium que indica la qualitat de la ubicació dels prestatges dels seients del cotxe de cada lloc)
- Age (Edat mitjana de la població local)
- Education (Nivell educatiu a cada lloc)
- Urban (Un factor amb els nivells Yes i No per indicar si la botiga es troba en una ubicació urbana o rural)
- US (Un factor amb els nivells Yes i No per indicar si la botiga es troba als EUA o no)

Les dades de l'estudi estàn a l'arxiu `ChildCarSeats1.csv`

Nota: important a tenir en compte per a lliurar l'activitat:

Cal 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 l'execució de la mateixa (pas a pas). S'ha de respectar la mateixa numeració dels apartats que l'enunciat.

2 Estadística descriptiva i visualització

2.1 Tipus de dades

Comproveu el tipus de variable que correspon a cada una de les variables. Quines són tipus numèric? Quines són tipus factor? Hi ha dades faltants?

2.2 Resum de dades quantitatives

Realitzeu una taula de les dades quantitatives on apareixi la mitja, la mitjana, la desviació standard i l'amplitud interquartílica (IQR, en anglès). Comenteu els resultats.

2.3 Diagrama de caixa

Mostreu amb diversos diagrames de caixa la distribució de la variable `Sales` segons: `ShelveLoc`, `Urban` i `US`. Interpretar els gràfics breument.

2.4 Representació gràfica de les variables qualitatives.

Representeu gràficament les variables qualitatives.

3 Estadística inferencial

3.1 Interval de confiança de la variable `Price`

Calculeu l'interval de confiança al 95% de la variable `Price`. A partir del valor obtingut, expliqueu com s'interpreta el resultat de l'interval de confiança.

Nota: S'han de realitzar els càlculs manualment. No es poden fer servir funcions de `R` que calculin directament l'interval de confiança com `t.test` o similar. Si que podeu fer servir funcions com `qnorm`, `pnorm`, `qt` i `pt`

3.2 Test de comparació de dues mitjanes

Es pot acceptar que en les botigues d'EUA (variable `US`) la mitjana de vendes del seients de cotxes infantils (variable `Sales`) és superior a la mitjana de vendes en botigues fora d'EUA? Calculeu per a un nivell de confiança del 95%.

Nota: S'han de realitzar els càlculs manualment. No es poden fer servir funcions de `R` que calculin directament l'interval de confiança com `t.test` o similar. Si que podeu fer servir funcions com `qnorm`, `pnorm`, `qt` i `pt`

S'assumirà que la variable `Sales` té distribució normal.

Seguiu els passos que es detallen a continuació:

3.2.1 Escriure la hipòtesi nul · la i alternativa

3.2.2 Justificar quin mètode aplicareu

3.2.3 Realitzar els càlculs de l'estadístic de contrast, valor crític i p valor amb un nivell de confiança del 95%

Per comprovar-ho podeu usar la funció R.

3.3 Contrast no paramètric

En l'apartat anterior hem assumit la normalitat de la variable vendes (`Sales`). Ara apliqueu un test no paramètric per respondre la mateixa pregunta anterior. Podeu usar una funció `R` per resoldre el contrast.

3.3.1 Interpreteu el resultat

4 Regressió

4.1 Model de regressió

Apliqueu un model de regressió lineal múltiple que tingui com a variables explicatives: `Price`, `Advertising`, `Age`, `Population`, `ShelveLoc`, `US`, i `Urban`, i com a variable dependent les vendes, variable `Sales`.

Especifiqueu el nivell base (usant la funció `relevel`): per a la variable `ShelveLoc`, la categoria "Bad", per a la variable `US`, la categoria "Yes", i per a la variable `Urban`, la categoria "Yes".

4.2 Interpretar el model

Interpreteu el model ajustat. Expliqueu quina interpretació en feu de la contribució en el model de les variables regressores. Indiqueu com seria el model de regressió per una botiga fora d'EUA, no urbana i amb un `ShelveLoc` de tipus "Bad".

4.3 Predicció

Apliqueu el model de regressió per predir `Sales` d'una botiga fora d'EUA a una zona rural, amb `Price` de 131 dòlars, `Advertising` de 0 dòlars, `Population` de 139 milers de persones, `Age` de 40 anys i `ShelveLoc` de tipus ¨Bad¨.

Compareu el resultat amb el d'una botiga fora d'EUA a una zona rural, amb `Price` de 131 dòlars, `Advertising` de 9 mil de dòlars, `Population` de 139 milers de persones, `Age` de 40 anys i `ShelveLoc` de tipus ¨Good¨.

Expliqueu les diferències en funció dels coeficients del model de regressió.

5 Anàlisi de la variància (ANOVA)

5.1 Anova d'un factor

5.1.1 Vendes i qualitat de la localització dintre de l'expositor

Realitzeu un ANOVA per contrastar la significació de la variable `ShelveLoc` en la variable `Sales`.

5.1.1.1 Escriure la hipòtesis nul · la i alternativa

5.1.1.2 Model Calculeu l'anàlisi de variància, utilitzant la funció `aov` o `lm`. Interpreteu el resultat de l'anàlisi, tenint en compte els valors Sum Sq, Mean SQ, F i Pr (> F).

5.1.1.3 Càlculs

a) Mostreu gràficament la distribució de vendes, `Sales`, segons el factor `ShelveLoc` ordenat segons la qualitat: "Bad", "Medium" i "Good". Pots fer servir la funció `reorder`.

b) Per tal d'aprofundir en la comprensió del model ANOVA, calculeu manualment la suma de quadrats intra i la suma de quadrats entre grups. Els resultats han de coincidir amb el resultat del model ANOVA. Com a referència, pots obtenir les fórmules de López-Roldán i Fachelli (2015), pàgines 29-33.

c) També proporcioneu l'estimació dels efectes dels nivells del factor `ShelveLoc`. I l'estimació de la variància de l'error.

Comprovació amb la funció `lm()` o `aov()`.

5.1.2 Interpreteu els resultats

5.2 Adequació del model

Mostra visualment l'adequació del model ANOVA. Podeu fer servir plot sobre el model ANOVA resultant. En els apartats següents es demana la interpretació d'aquests gràfics.

5.2.1 Normalitat dels residus

Interpreteu la normalitat dels residus a partir de l'gràfic Normal Q-Q que es mostra en l'apartat anterior.

5.2.2 Homocedasticitat dels residus

Els gràfics "Residuals vs Fitted", "Scale-Location" i "Residuals vs Factor levels" donen informació sobre la homocedasticitat dels residus. Interpreteu aquests gràfics.

5.3 ANOVA no paramètric

Si la validació de les premisses de normalitat i homocedasticitat no es verifiquen es pot aplicar un test no paramètric, per exemple el test de Kruskal-Wallis.

Apliqueu un test de Kruskal-Wallis per contrastar si hi ha diferències entre les botigues segon on s'exposa (`ShelveLoc`) pel que fa a les vendes (`Sales`) només a les observacions amb `Advertising` major que el valor de la mediana.

5.3.1 Interpreteu els resultats

6 ANOVA multifactorial

A continuació, es vol avaluar l'efecte de més d'un factor sobre la variable `Sales` on el primer factor sempre serà `ShelveLoc`. Primer es realitzarà l'anàlisi on el segon factor és `US` i després, l'anàlisi on el segon factor és `Urban`.

6.1 Factors: `ShelveLoc` i `US`

6.1.1 Anàlisis visual dels efectes principals i possibles interaccions

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

1. Agrupeu el conjunt de dades per `ShelveLoc` i per `US`. Calcular la mitjana de vendes per a cada grup. Per realitzar aquest procés, es pot fer amb les funcions `group_by` i `summarise` de la llibreria `dplyr`.
2. Mostreu el conjunt de dades en forma de taula, on es mostri la mitjana de cada grup segons `ShelveLoc` i `US`.
3. Mostreu en un gràfic el valor mitjà de la variable `Sales` per a cada factor. Pots inspirar-te en els gràfics de López-Roldán i Fachelli (2015), p.38. Pots realitzar aquest tipus de gràfic usant la funció `ggplot` de la llibreria `ggplot2`.
4. Interpreteu el resultat sobre si només hi ha efectes principals o hi ha interacció entre els factors. Si hi ha interacció, explicar com s'observa aquesta interacció en el gràfic.

6.1.2 Adequació del model

Interpreteu l'adequació del model ANOVA obtingut usant els gràfics de residus.

6.2 Factors: `ShelveLoc` i `Urban`

6.2.1 Anàlisis visual dels efectes principals i possibles interaccions

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

1. Agrupeu el conjunt de dades per `ShelveLoc` i per `Urban`. Calcular la mitjana de vendes per a cada grup. Per realitzar aquest procés, es pot fer amb les funcions `group_by` i `summarise` de la llibreria `dplyr`.
2. Mostreu el conjunt de dades en forma de taula, on es mostri la mitjana de cada grup segons `ShelveLoc` i `Urban`.
3. Mostreu en un gràfic el valor mitjà de la variable `Sales` per a cada factor. Pots inspirar-te en els gràfics de López-Roldán i Fachelli (2015), p.38. Pots realitzar aquest tipus de gràfic usant la funció `ggplot` de la llibreria `ggplot2`.
4. Interpreteu el resultat sobre si només hi ha efectes principals o hi ha interacció entre els factors. Si hi ha interacció, explicar com s'observa aquesta interacció en el gràfic.

7 Comparacions múltiples

Prenent com a referència el model ANOVA multifactorial, amb els factors `ShelveLoc` i `US`, aplicar el test de comparació múltiple Scheffé. Interpreteu el resultat del test i indicar quins grups són diferents significativament entre si.

8 Conclusions

---

9 Comentaris importants sobre l'activitat

1. **No es pot inspeccionar ni corregir de manera manual** el fitxer de dades. Per exemple, **no** es poden realitzar instruccions d'aquest tipus:

   ```
   data[1,5] <- 32.5
   ```

   Aquest tipus de transformacions s'han de fer amb funcionalitats de cerca (buscar els registres que tenen errors o inconsistències) i després fer les correccions oportunes amb funcionalitats de R. Així el procediment de neteja és útil, independentment del fitxer de dades i de la posició i valors concrets de l'arxiu.

2. **No es poden fer llistats complets de les dades del fitxer a pantalla**, perquè generen arxius de sortida excessivament grans. Si es desitja validar el resultat d'una instrucció sobre les dades, es pot usar la funció **head** que mostra les primeres files de la taula de dades o **tail** que mostra les últimes.

Puntuació de l'activitat

- Apartat 2 (10%)
- Apartat 3 (10%)
- Apartat 4 (10%)
- Apartat 5 (20%)
- Apartat 6 (20%)
- Apartat 7 (10%)
- Apartat 8 (10%)
- Qualitat del informe dinàmic (10%)