Estadística Avançada-Activitat 4: Anàlisi de la variança i repàs del curs

Enunciat

19 de maig 2019

Índex

Introducció

Es realitza un estudi transversal en 952 pacients amb afectació a l'artèria coronària (Coronary Artery Disease, CAD). L'objectiu consisteix en investigar la interacció entre gènere i d'altres característiques socioeconòmiques sobre la qualitat del son en els pacients de CAD.

Es consideren les característiques socioeconòmiques següents: nivell d'educació (**educ_level**), ingressos mensuals (**monthly_income**), estat civil (**marital_status**) i lloc de residència (**residence**). També es consideren les variables: ingressos mensuals en dòlars a l'inici de l'estudi (**income_monthly_c**) i les hores de son diàries promig al llarg d'un mes des de l'inici de la investigació (**sleep_duration_avg**). També es registra el sexe del pacient (**sex**).

La qualitat del son es mesura usant l'índex de qualitat del son de Pittsburgh (PSQI). PSQI mesura l'autovaloració de la qualitat del son durant el darrer mes en les set àrees següents: 1) qualitat subjetiva del son (autopercepció de la qualitat general del son), 2) latència del son, 3) durada del son, 4) eficiència del son habitual, 5) trastorns del son, 6) ús de medicació per a dormir, i 7) disfunció diürna (problemes experimentats durant el dia degut al son desordenat). Cada component s'escala de 0 a 6, on el 0 representa condició normal, i 1 - 6 condicions anormals. La suma de puntuacions d'aquests set components produeix una puntuació total (**psqi**).

Les dades estan en el fitxer adjunt a l'activitat: "RawDataUnbalanced.csv".

Per a realitzar aquesta anàlisi, se seguiran els passos següents:

- S'inicia l'estudi amb una anàlisi descriptiva, juntament amb gràfics que ofereixin una primera idea de les variables que influeixen en la qualitat del son.
- En els apartats 2-3, s'analitza si la qualitat del son dels pacients està influïda per certes variables socioeconòmiques. S'aplicaran tests d'hipòtesis de dues mostres i anàlisi de regressió, revisant els conceptes que s'han tractat al llarg del curs.
- A continuació (apartats 4 i 5), es realiza una anàlisi ANOVA unifactorial, tenint en compte el nivell d'educació com a factor que pot influir en la qualitat del son.
- A l'apartat 6, s'aplica l'anàlisi de variança tenint en compte dos factors (sexe i nivell educatiu; sexe i ingressos mensuals (categòric)). S'estudiaran els efectes principals i les interaccions entre factors.
- Finalment, s'aplica un test de comparacions múltiples per a investigar quins grups de pacients tenen una qualitat del son significativament diferent de la resta.

Nota important a tenir en compte per a lliurar 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 de respectar la numeració dels apartats de l'enunciat.
- No realitzeu llistats dels conjunts de dades, donat que aquests poden ocupar varies pàgines. Si voleu comprovar l'efecte d'una instrucció sobre les dades, podeu usar la funció head que mostra les 10 primeres files del conjunt de dades.

1 Anàlisi descriptiva i visualització

Realitzar una primera anàlisi descriptiva de les dades de la mostra

1.1 Distribució de la qualitat del son

Mostrar en un diagrama de caixa la distribució de la qualitat del son (variable **psqi**) de la mostra.

1.2 Diagrames de caixa

Mostrar en varis diagrames de caixa la distribució de la variable **psqi** segons el sexe, segons el nivell educatiu, segons el nivell d'ingressos, segons l'estat civil i segons la residència.

1.3 Variables categòriques

Mostreu en diagrames circulars o en diagrames de barres la distribució de valors de les variables categòriques de la mostra.

1.4 Interpretació

Interpreteu els gràfics breument.

2 Estadística inferencial

2.1 Interval de confiança de la qualitat del son

Calcular l'interval de confiança al 95% de la variable **psqi** dels pacients. 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 podeu usar funcions d'R que calculin directament l'interval de confiança com **t.test** o similar. Sí que podeu usar funcions com **qnorm** , **pnorm** , **qt** i **pt**

2.2 Test de dues mostres: de la qualitat del son en funció del tipus de residència

Es pot afirmar que la qualitat del son (variable **psqi**) en zones rurals és millor que la qualitat del son en zones urbanes? Calcular-ho per un nivell de confiança del 95% i 97%.

Nota: S'han de realitzar els càlculs manualment. No podeu usar funcions d'R que calculin directament l'interval de confiança com **t.test** o similar. Sí que podeu usar funcions com **qnorm**, **pnorm**, **qt** i **pt** Assumiu que la variable **psqi** té una distribució normal.

Seguiu els passos que s'indiquen a continuació:

2.2.1 Escriure la hipòtesi nul · la i alternativa

2.2.2 Justifiqueu quin mètode aplicareu

2.2.3 Realitzeu els càlculs de l'estadístic de contrast, valor crític i valor p, al 95% i 97% de nivell de confiança

2.2.4 Interpreteu el resultat i doneu resposta a la pregunta plantejada

3 Regressió

3.1 Model de regressió

Apliqueu un model de regressió lineal múltiple que usi com a variables explicatives: les hores promig de son, els ingressos mensuals, el sexe, el nivell d'educació i el tipus de residència, i com a variable dependent la qualitat del son.

Especifiqueu en el nivell base (en el `relevel`): per a la variable sexe, la categoria "M", per al nivell d'educació, la categoria "Primary", i per la variable tipus de residència, la categoria "Rural".

3.2 Interpretació

Interpreteu el model de regressió resultant, indicant quins regressors són significatius i expliqueu si existeixen diferències degut al nivell educatiu i si hi són, en quin sentit. De manera anàloga, repetiu la mateixa discussió per a la resta de variables regressores en el model.

3.3 Predicció

Apliqueu el model de regressió per a predir **psqi** d'una dona, amb ingressos mensuals de 3600 dòlars, que dorm en promig 6 hores, de nivell d'estudis primaris i residència de tipus rural.

Compareu el resultat amb el d'una dona, amb els mateixos valors, excepte que la residència és urbana.

Compareu també el resultat de la primera dona amb el un d'home amb els mateixos paràmetres i residència tipus rural.

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

3.4 Intervals de predicció

Calculeu els intervals de predicció dels tres pacients al 98%. Per a fer-ho, podeu afegir a la funció `predict()` l'argument `interval = prediction`. Per a especificar el nivell de confiança podeu usar l'argument `level` (que per defecte és 0.95). Interpreteu els resultats. També obtingueu l'interval de confiança al 98% per la resposta mitjana. Compareu els dos intervals i interpreteu el resultat.

3.5 Interpretació de la predicció

Interpreteu els resultats obtinguts a l'apartat anterior. Concretament, consideraríeu que el model és precís, tenint en compte els valors d'R^2 i p-value obtinguts?

3.6 Ajust del model

Analitzeu l'adequació del model a partir de l'anàlisi gràfic de residus. A tal efecte, es pot usar la instrucció `plot()` passant com a paràmetre el model de regressió lineal. Per saber com interpretar els gràfics de residus, podeu consultar l'enllaç següent: http://data.library.virginia.edu/diagnostic-plots/

4 Anàlisi de variança unifactorial

A continuació, ens preguntem si la qualitat del son (variable **psqi**) està influïda pel nivell d'educació. Donat que aquesta variable té tres nivells, s'aplicarà anàlisi de variança.

4.1 Hipòtesis nul · la i alternativa

Escriure la hipòtesis nul · la i l'alternativa per a l'anàlisi ANOVA.

4.2 Model

Calculeu l'anàlisi de variança, usant la funció **aov**. Interpreteu el resultat de l'anàlisi, tenint en compte els valors Sum Sq, Mean SQ, F i Pr (> F).

4.3 Càlculs

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

5 Adequació del model ANOVA

Valideu l'adequació del model ANOVA. Podeu consultar López-Roldán i Fachelli (2015), gràfic III.8.6, pàgina 25. Seguiu els passos que s'indiquen a continuació.

5.1 Visualització de l'adequació del model

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

5.2 Normalitat dels residus

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

5.3 Homocedasticitat dels residus

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

Podeu consultar informació complementària sobre com interpretar aquests gràfics en els enllaços següents:

https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/anova/how-to/one-way-anova/interpret-the-results/all-statistics-and-graphs/#normal-probability-plot-of-the-residuals

https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/modeling-statistics/anova/how-to/two-way-anova/interpret-the-results/all-statistics-and-graphs/ Veure: **Residuals versus fits**

5.4 ANOVA no paramètric

5.4.1 Kruskal-Wallis

Si les condicions d'aplicació ANOVA no es compleixen, llavors se sol aplicar un test no paramètric. El test Kruskal-Wallis és l'equivalent no paramètric d'ANOVA. Apliqueu el test usant la funció **kruskal.test()** i interpreteu el resultat, amb independència de que es compleixin o no les condicions d'aplicació d'ANOVA.

Podeu consultar l'enllaç següent:

https://www.sheffield.ac.uk/polopoly_fs/1.714570!/file/stcp-karadimitriou-KW.pdf

5.4.2 Interpretació

Expliqueu en què es diferencia el càlcul d'ANOVA del càlcul del test Kruskal-Wallis. Si el test Kruskal-Wallis no té les mateixes restriccions d'aplicació que el test ANOVA, per què creieu que es continua aplicant ANOVA, si les condicions ho permeten, enlloc de Kruskal-Wallis?

6 ANOVA multifactorial

A continuació, es vol avaluar l'efecte de més d'un factor sobre la variable **psqi**. En primer lloc, es realitzarà l'anàlisi combinat de sexe i nivell educatiu sobre la qualitat del son. En segon lloc, es realitzarà l'anàlisi combinat de sexe i ingressos mensuals sobre la qualitat del son.

6.1 Factors: sexe i nivell educatiu

6.1.1 Anàlisi visual dels efectes principals i possibles interaccions

Dibuixeu en un gràfic la variable **psqi** en funció del sexe i en funció del nivell educatiu. El gràfic ha de permetre avaluar si hi ha interacció entre els dos factors. Per això, es recomana que se segueixin els passos següents:

1. Agrupeu el conjunt de dades por sexe i per nivell educatiu. Calculeu la mitjana de la qualitat del son per a cada grup. Per a 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 el sexe i el nivell d'estudis.

3. Mostreu en un gràfic el valor mig de la variable **psqi** per a cada sexe i educació. Podeu inspirar-vos en els gràfics de López-Roldán i Fachelli (2015), p.38. Podeu realitzar aquest tipus de gràfic usant la funció ggplot de la llibreria **ggplot2**.
4. Interpreteu el resultat sobre si existeixen efectes principals o hi ha interacció entre els factors. Si hi ha interacció, expliqueu com s'observa aquesta interacció en el gràfic.

6.1.2 ANOVA multifactorial

Realitzeu l'anàlisi ANOVA amb els factors sexe i nivell educatiu i la seva interacció Interpreteu el resultat del model i expliqueu si els factors (i la interacció entre factors) són significatius per a modelar la variable **psqi**.

6.1.3 Adequació del model

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

6.2 Factors: sexe i ingressos mensuals

6.2.1 Anàlisi visual dels efectes principals i possibles interaccions

Realitzeu l'anàlisi visual de la variable **psqi** en funció del sexe i en funció dels ingressos. El gràfic ha de permetre avaluar si hi ha ha interacció entre factors.

6.2.2 ANOVA multifactorial

Realitzeu l'anàlisi ANOVA amb els factors sexe i ingressos mensuals, i la interacció. Interpreteu el resultat del model i expliqueu si els factors (i la interacció entre factors) són significatius per a modelar la variable **psqi**.

7 Comparacions múltiples

Prenent com a referència el model ANOVA multifactorial, amb els factores sexe i nivell educatiu, apliqueu el test de comparació múltiple Scheffé. Interpreteu el resultat del test i indiqueu quins grups són diferents significativament entre si.

8 Conclusions

Escriure les conclusions finals de l'estudi en relació als objectius de la investigació.

9 Puntuació dels apartats

- Apartat 1: 5%
- Apartat 2: 5%
- Apartat 3: 10%
- Apartat 4: 20%
- Apartat 5: 20%
- Apartat 6: 20%
- Apartat 7: 10%

- Qualitat de l'informe dinàmic: 10%

10 Referències

López-Roldán, P., i Fachelli, S. (2015). Capítulo III.8 Análisis de varianza. A Metodología de la investigació social quantitativa. Barcelona: UAB.