A4 - Anàlisi de variància i repàs del curs

Enunciat

Semestre 2020.2

Índex

1 Introducció

El conjunt de dades trainCLEAN.csv s'inspira (ha estat modificat per motius acadèmics) en la base de dades disponible en la plataforma Kaggle: https://www.kaggle.com/c/actuarial-loss-estimation.

Aquest conjunt de dades conté informació d'una mostra d'indemnitzacions atorgades per una companyia d'assegurances pel temps que ha estat de baixa laboral el treballador. El conjunt de dades conté 54,000 registres i 15 variables.

Les principals variables que s'usaran en aquesta activitat són:

- ClaimNumber: Identificador de la pòlissa.
- DateTimeOfAccident: Data de l'accident.
- DateReported: Data que es comunica a la companyia i aquesta obre un expedient del sinistre (obertura).
- Age: Edat del treballador.
- Gender: Sexe.
- MaritalStatus: Estat civil, (M)arried, (S)ingle, (U)nknown.
- DependentChildren: Nombre de fills dependents.
- DependentsOther: Nombre de dependents excloent fills
- WeeklyWages: Salari setmanal (en EUR).
- PartTimeFullTime: Jornada laboral, Part time (P) o Full time(F).
- HoursWorkedPerWeek: Nombre d'hores per setmana.
- DaysWorkedPerWeek: Nombre de dies per setmana.
- ClaimDescription: Descripció sinistres.
- InitialIncurredClaimCost: Estimació inicial del cost realitzat per la companyia.
- UltimateIncurredClaimCost: Cost total pagat per sinistre.

Aquestes dades ens ofereixen múltiples possibilitats per a consolidar els coneixements i competències de manipulació de dades, preprocessament, anàlisi descriptiva i inferència estadística.

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

- És necessari lliurar l'arxiu Rmd i el fitxer de sortida (PDF o html). L'arxiu de sortida ha d'incloure: el codi i el resultat de l'execució del codi (pas a pas).
- S'ha de respectar la mateixa numeració dels apartats que l'enunciat.
- No es poden realitzar llistats complets del conjunt de dades en la solució. Això generaria un document amb centenars de pàgines i dificulta la revisió del text. Per a comprovar les funcionalitats del codi sobre les dades, es poden usar les funcions **head** i **tail** que només mostren unes línies del fitxer de dades.
- Es valora la precisió dels termes utilitzats (cal fer servir de manera precisa la terminologia de l'estadística).
- Es valora també la concisió en la resposta. No es tracta de fer explicacions molt llargues o documents molt extensos. Cal explicar el resultat i argumentar la resposta a partir dels resultats obtinguts de manera clara i concisa.

2 Lectura del fitxer i preparació de les dades

Llegiu el fitxer `trainCLEAN.csv` i guardeu les dades en un objecte amb identificador denominat *claim*. A continuació, verifiqueu que les dades s'han carregat correctament.

2.1 Preparació de les dades

Canviem el nom de les variables a castellà. En concret, es demana que es denominin de la següent forma: `Id`, `Ocurrencia`, `Apertura`, `Edad`, `Sexo`, `Estado`, `Dependientes`, `OtrosDepend`, `Salario`, `Jornada`, `CosteInicio`, `CosteFinal`, `HorasSemana`, `DiasSemana` y `Descripcion`.

- Les variables 'Ocurrencia' i 'Apertura' estan classificades com a factor. Per a poder treballar amb elles cal convertir-les en dates
- Crear una variable denominada 'tiempo' que comptabilitzi en dies el temps que triga a obrir-se un sinistre per la companyia des de la seva ocurrència.

2.2 Classificació del temps

La variable `tiempo` indica la durada d'obertura del sinistre de la següent forma: "Molt ràpid" si s'obertura en 15 dies o menys, "Ràpid" si s'obertura entre 16 i 30 dies, "Lent" si s'obertura entre 31 i 89 dies, i "Molt lent" si triga 90 dies o més en obrir-se el sinistre. Creeu una variable categòrica denominada `Clasificacion`, que classifiqui el sinistre segons aquestes categories.

2.3 Valors absents

- Analitzeu el nombre de categories diferents en les variables 'Descripcion', 'Sexo' i 'Estado'. Quantes descripcions diferents hi ha dels sinistres?
- Representeu els observacions amb la categoria "U" (U=unknown) en les variables 'Sexo' i 'Estado' com missings.
- Comproveu la proporció d'observacions que tenen valors absents i traieu conclusions sobre com de seriós és el problema de valors absents en aquestes dades.
- Elimineu els valors absents del conjunt de dades. Denominem al conjunt de dades claimNet.

2.4 Salut mental

La companyia està preocupada per les baixes per salut mental. Per aquest motiu, vol monitorar les baixes que incloguin les paraules `Stress`, `Anxiety`, `Harassment` o `Depression`. Es demana:

- Crear la variable dicotòmica denominada 'RiesgoSM' si la variable 'Descripcion' inclou alguna de aquestes paraules.

2.5 Anàlisi visual

1. Mostreu amb diversos diagrames de caixa la distribució de la variable 'CosteFinal' en escala logarítmica segons la variable 'Sexo', segons 'Estado', segons 'Clasificacion' i segons 'RiesgoSM'.
2. Interpreteu els gràfics breument.

2.6 Comprovació de normalitat

Podem assumir que la variable `CosteFinal` té una distribució normal? Heu de justificar la resposta a partir de mètodes visuals i contrastos.

- Realitzeu inspecció visual de normalitat.
- Realitzeu contrast de normalitat de Lilliefors (p. ex. amb funció lillie.test de la llibreria nortest).
- Realitzeu inspecció visual i contrast de normalitat a la variable 'CosteFinal' en escala logarítmica.

---

3 Estadística inferencial

Utilitzem el conjunt de dades claimNet.

3.1 Interval de confiança de la mitjana poblacional de la variable `CosteFinal`

- Calculeu manualment l'interval de confiança al 95% de la mitjana poblacional de la variable 'CosteFinal' en escala normal (No es poden utilitzar funcions com a t.test o z.test per al càlcul).
- A partir del resultat obtingut, expliqueu com s'interpreta l'interval de confiança.

3.2 Contrast d'hipòtesi per a la diferència de mitjanes

Podem acceptar que la indemnització a les dones supera en més de 1000 EUR la dels homes?

Responeu a la pregunta utilitzant un nivell de confiança del 95%.

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

Seguiu els passos que es detallen a continuació.

3.2.1 Escriviu la hipòtesi nul · la i l'alternativa

3.2.2 Justificació del test a aplicar

3.2.3 Càlculs

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

3.2.4 Interpretació del test

---

4 Model de regresió lineal

Estimeu un model de regressió lineal múltiple que tingui com a variables explicatives: `Edad`, `Sexo`, `Estado`, `Dependientes`, `OtrosDepend`, `Salario`, `Jornada`, `HorasSemana`, `DiasSemana`, `Clasificacion`, `RiesgoSM`, `CosteInicio` i com variable depenent el `CosteFinal` en escala logarítmica (Nota: es recomana transformar també a escala logarítmica la variable explicativa `CosteInicio`)

4.1 Interpretació del model

Interpreteu el model lineal ajustat:

– Quina és la qualitat de l'ajust?

– Expliqueu la contribució de les variables explicatives en el model.

4.2 Anàlisi residus

Finalment, per a aprofundir en la qualitat de l'ajust s'han d'analitzar els residus que ens indicaran realment com s'ajusta el nostre model a les dades mostrals.

- La sortida de 'summary()' presenta els principals estadístics de la distribució dels residus. Analitzeu els valors estimats dels estadístics.
- Realitzeu una anàlisi visual dels residus

4.3 Predicció

Heu de predir el cost esperat per a les següents característiques: Edad=24, Sexo= "F", Estado="S", Dependientes=1, OtrosDepend=0, Salario=500, Jornada="F",HorasSemana=40,DiasSemana=5, Clasificacion="Lent", RiesgoSM="TRUE" y "CosteInicio"=10000.

(Nota: Heu de tenir en compte que el valor esperat d'una variable aleatòria que el seu logaritme es distribueix segons una normal, i.e. distribució lognormal, és $exp(mu + var/2)$ on mu i var són la mitjana i la variància de la transformació logarítmica).

---

5 Regressió logística

5.1 Model predictiu

Utilitzant les mateixes característiques com a variables explicatives, ajusteu un model predictiu basat en la regressió logística per a predir la probabilitat que la companyia quantifiqui inicialment el cost del sinistre de manera insuficient.

Per a això, creeu una variable `Deficit` que indiqui si la valoració inicial del cost del sinistre (`CosteInicio)` és inferior a la indemnització finalment pagada per la companyia (`CosteFinal`). La variable `Deficit` ha de codificar-se com una variable dicotòmica, que pren el valor 0 quan la valoració inicial ha estat suficient i 1 quan la valoració inicial ha estat insuficient.

La variable `Deficit` serà la variable dependent del model. Analitzeu la qualitat del model i les variables que són rellevants.

5.2 Interpretació

Interpreteu el model ajustat. Concretament, expliqueu la contribució de les variables explicatives amb coeficient estadísticament significatiu per a predir si la valoració inicial és insuficient per a cobrir el cost del sinistre.

5.3 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 (valoració inicial del cost insuficient) si la probabilitat del model de regressió logística és superior o igual a 0.5 i 0 en cas contrari. Analitzeu la matriu de confusió i les mesures de 'sensitivity' i 'specificity'.

Nota: Preneu com a categoria d'interès que hi hagi dèficit en la valoració inicial del cost. Per tant, dèficit igual a 1 serà el cas positiu en la matriu de confusió i 0 el cas negatiu.

5.4 Predicció

Amb que probabilitat la valoració inicial del sinistre serà insuficient per a un home de 20 anys d'edat, solter, sense fills ni altres dependents, amb un salari setmanal de 300 EUR, jornada partida, amb 30 hores setmanals i cinc dies a la setmana, una classificació del temps fins a l'obertura del sinistre de "Molt lent", una baixa que no és per salut mental i una valoració inicial de 10000 EUR?

---

6 Anàlisi de la variància (ANOVA) de un factor

Realitzarem un ANOVA per a contrastar si existeixen diferències en la variable `CosteFinal` en escala logarítmica en funció de la classificació del sinistre en relació al temps transcorregut fins a l'obertura. Seguiu

els passos que s'indiquen.

6.1 Hipòtesi nul · la i alternativa

Escriviu la hipòtesi nul · la i l'alternativa.

6.2 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 y Pr (> F).

6.3 Efectes dels nivells del factor

Calculeu la variabilitat explicada per la variable `Clasificacion` sobre la variable `CosteFinal` mitjançant la mètrica eta squared. Interpreteu els resultats.

6.4 Contrast dos-a-dos

Com els factors han resultat significatius cal fer contrastos de comparacions múltiples. Es pot utilitzar la prova de Tukey-Kramer que compara dos-a-dos les diferents categories de la variable. (Nota: per exemple, amb la funció `HSD.test()` del paquet **agricolae**).

6.5 Adequació del model

Mostreu l'adequació del model ANOVA. Es demana el següent:

- Anàlisi visual de normalitat dels residus. Podeu utilitzar la funció plot sobre el model ANOVA calculat.
- Anàlisi visual de homocedasticitat dels residus. Podeu utilitzar plot sobre el model ANOVA calculat.
- Contrast de normalitat i homocedasticitat.

6.5.1 Normalitat dels residus

L'anàlisi visual de la normalitat dels residus es pot fer a partir del gràfic Normal Q-Q. Mostreu i interpreteu aquest gràfic.

6.5.2 Homocedasticitat dels residus

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

6.5.3 Contrast de normalitat

Es pot comprovar el supòsit de normalitat dels residus amb les proves estadístiques de Shapiro-Wilk o Lilliefors, entre altres. El supòsit de homocedasticitat es pot comprovar a partir de la prova de Bartleet.

---

7 ANOVA multifactorial

A continuació, es desitja avaluar l'efecte sobre `CosteFinal` en escala logarítmica segons `Sexo` combinat amb el factor `RiesgoSM`. Seguiu els passos que s'indiquen a continuació.

7.1 Anàlisi dels efectes principals i possibles interaccions

Dibuixeu en un gràfic la variable `CosteFinal` en escala logarítmica en funció de `Sexo` i en funció de `RiesgoSM`.

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 `Sexo` i per `RiesgoSM`. Calculeu el nombre de casos disponibles de cada combinació de factors.
2. Calculeu la mitjana de cost (en log) per a cada grup.
3. Mostreu en un gràfic el valor mitjà de la variable `CosteFinal` en escala logarítmica per a cada factor.
4. 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

- Calculeu el model incloent la interacció entre els factors.
- Mesureu l'efecte dels factors sobre la variabilitat explicada del Cost final (en escala logarítmica).
- Analitzeu dos-a-dos les diferències de mitjanes entre els diferents factors.
- Adequació del model. Realitzar anàlisi visual de normalitat i homocedasticitat.

7.3 Interpretació dels resultats

---

8 Conclusions

Resumiu les conclusions principals de l'anàlisi. Per fer-ho, podeu resumir les conclusions de cadascun dels apartats.

Puntuació de l'actividad

- Apartats 1 i 2 (15%)
- Apartat 3 (15%)
- Apartat 4 (15%)
- Apartat 5 (15%)
- Apartat 6 (15%)
- Apartat 7 (15%)
- Qualitat de l'informe dinàmic (10%)