A4 - Análisis de varianza y repaso del curso

Enunciado

Semestre 2020.2

Índex

Introducción

El conjunto de datos trainCLEAN.csv se inspira (ha sido modificado por motivos académicos) en la base de datos disponible en la plataforma Kaggle: https://www.kaggle.com/c/actuarial-loss-estimation.

Este conjunto de datos contiene información de una muestra de indemnizaciones otorgadas por una compañía de seguros por el tiempo que ha estado de baja laboral el trabajador. El conjunto de datos contiene 54,000 registros y 15 variables.

Las principales variables que se usarán en esta actividad son:

- ClaimNumber: Identificador de la póliza.
- DateTimeOfAccident: Fecha del accidente.
- DateReported: Fecha que se comunica a la compañía y ésta abre un expediente del siniestro (apertura).
- Age: Edad del trabajador.
- Gender: Sexo.
- MaritalStatus: Estado civil, (M)arried, (S)ingle, (U)nknown.
- DependentChildren: Número de hijos dependientes.
- DependentsOther: Número de dependientes excluyendo hijos
- WeeklyWages: Salario semanal (en EUR).
- PartTimeFullTime: Jornada laboral, Part time (P) o Full time(F).
- HoursWorkedPerWeek: Número horas por semana.
- DaysWorkedPerWeek: Número de días por semana.
- ClaimDescription: Descripción siniestros.
- InitialIncurredClaimCost: Estimación inicial del coste realizado por la compañía.
- UltimateIncurredClaimCost: Coste total pagado por siniestro.

Estos datos nos ofrecen múltiples posibilidades para consolidar los conocimientos y competencias de manipulación de datos, preprocesado, análisis descriptivo e inferencia estadística.

Nota importante a tener en cuenta para entregar la actividad:

- Es necesario entregar el archivo Rmd y el fichero de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Se debe respetar la misma numeración de los apartados que el enunciado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificulta la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden usar las funciones **head** y **tail** que sólo muestran unas líneas del fichero de datos.
- Se valora la precisión de los términos utilizados (hay que usar de manera precisa la terminología de la estadística).
- Se valora también la concisión en la respuesta. No se trata de hacer explicaciones muy largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de manera clara y concisa.

1 Lectura del fichero y preparación de los datos

Leed el fichero `trainCLEAN.csv` y guardad los datos en un objeto con identificador denominado *claim*. A continuación, verificad que los datos se han cargado correctamente.

1.1 Preparación de los datos

Cambiamos el nombre de las variables a castellano. En concreto, se pide que se denominen de la siguiente forma: `Id`, `Ocurrencia`, `Apertura`, `Edad`, `Sexo`, `Estado`, `Dependientes`, `OtrosDepend`, `Salario`, `Jornada`, `CosteInicio`, `CosteFinal`, `HorasSemana`, `DiasSemana` y `Descripcion`.

- Las variables 'Ocurrencia' y 'Apertura' están clasificadas como factor. Para poder trabajar con ellas hay que convertirlas en fechas.
- Crear una variable denominada 'tiempo' que contabilice en días el tiempo que tarda en abrirse un siniestro por la compañía desde su ocurrencia.

1.2 Clasificación de tiempo

La variable `tiempo` indica la duración de apertura del siniestro de la siguiente forma: "Muy rápido" si se apertura en 15 días o menos, "Rápido" si se apertura entre 16 y 30 días, "Lento" si se apertura entre 31 y 89 días, y "Muy lento" si tarda 90 días o más en aperturarse el siniestro. Cread una variable categórica denominada `Clasificacion`, que clasifique el siniestro según estas categorías.

1.3 Valores ausentes

- Analizad el número de categorías distintas en las variables 'Descripcion', 'Sexo'y 'Estado'. ¿Cuántas descripciones distintas hay de los siniestros?
- Representad los observaciones con la categoría "U" (U=unknown) en las variables 'Sexo'y 'Estado' como missings.
- Comprobad la proporción de observaciones que tienen valores ausentes y sacad conclusiones sobre cómo de serio es el problema de valores ausentes en estos datos.
- Eliminad los valores ausentes del conjunto de datos. Denominamos al conjunto de datos claimNet.

1.4 Salud mental

La compañía está preocupada por las bajas por salud mental. Por este motivo, quiere monitorizar las bajas que incluyan las palabras `Stress`, `Anxiety`, `Harassment` o `Depression`. Se pide:

- Crear la variable dicotómica denominada 'RiesgoSM' si la variable 'Descripcion' incluye alguna de estas palabras.

1.5 Análisis visual

1. Mostrad con diversos diagramas de caja la distribución de la variable 'CosteFinal' en escala logarítmica según la variable 'Sexo', según 'Estado', según 'Clasificacion' y según 'RiesgoSM'.
2. Interpretad los gráficos brevemente.

1.6 Comprobación de normalidad

¿Podemos asumir que la variable `CosteFinal` tiene una distribución normal? Debéis justificar la respuesta a partir de métodos visuales y contrastes.

- Realizad inspección visual de normalidad.

- Realizad contraste de normalidad de Lilliefors (p.ej. con función lillie.test de la libreria nortest).
- Realizad inspección visual y contraste de normalidad a la variable Coste Final en escala logarítmica.

---

2 Estadística inferencial

Utilizamos el conjunto de datos claimNet.

2.1 Intervalo de confianza de la media poblacional de la variable `CosteFinal`

- Calculad manualmente el intervalo de confianza al 95% de la media poblacional de la variable 'CosteFinal' en escala normal (No se pueden utilizar funciones como t.test o z.test para el cálculo).
- A partir del resultado obtenido, explicad cómo se interpreta el intervalo de confianza.

2.2 Contraste de hipótesis para la diferencia de medias

¿Podemos aceptar que la indemnización a las mujeres supera en más de 1000 EUR la de los hombres?

Responded a la pregunta utilizando un nivel de confianza del 95%.

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

Seguid los pasos que se detallan a continuación.

2.2.1 Escribid la hipótesis nula y la alternativa

2.2.2 Justificación del test a aplicar

2.2.3 Cálculos

Realizad los cálculos del estadístico de contraste, valor crítico y valor p con un nivel de confianza del 95%.

2.2.4 Interpretación del test

---

3 Modelo de regresión lineal

Estimad un modelo de regresión lineal múltiple que tenga como variables explicativas: `Edad`, `Sexo`, `Estado`, `Dependientes`, `OtrosDepend`, `Salario`, `Jornada`, `HorasSemana`, `DiasSemana`, `Clasificacion`, `RiesgoSM`, `CosteInicio` y como variable dependiente el `CosteFinal` en escala logarítmica (Nota: se recomienda transformar también a escala logarítmica la variable explicativa `CosteInicio`)

3.1 Interpretación del modelo

Interpretad el modelo lineal ajustado:

– ¿Cuál es la calidad del ajuste?

– Explicad la contribución de las variables explicativas en el modelo.

3.2 Análisis residuos

Por último, para profundizar en la calidad del ajuste se deben analizar los residuos que nos indicarán realmente como se ajusta nuestro modelo a los datos muestrales.

- La salida de 'summary()' presenta los principales estadísticos de la distribución de los residuos. Analizad los valores estimados de los estadísticos.
- Realizad un análisis visual de los residuos

3.3 Predicción

Predecid el coste esperado para las siguientes características: Edad=24, Sexo= "F", Estado="S", Dependientes=1, OtrosDepend=0, Salario=500, Jornada="F",HorasSemana=40,DiasSemana=5, Clasificacion="Lento", RiesgoSM="TRUE" y "CosteInicio"=10000.

(Nota: Debes tener en cuenta que el valor esperado de una variable aleatoria que su logaritmo se distribuye según una normal, i.e. distribución lognormal, es $exp(mu + var/2)$ donde mu y var son la media y la varianza de la transformación logarítmica).

---

4 Regresión logística

4.1 Modelo predictivo

Utilizando las mismas características como variables explicativas, ajustad un modelo predictivo basado en la regresión logística para predecir la probabilidad de que la compañía cuantifique inicialmente el coste del siniestro de forma insuficiente.

Para ello, cread una variable `Deficit` que indique si la valoración inicial del coste del siniestro (`CosteInicio)` es inferior a la indemnización finalmente pagada por la compañía (`CosteFinal`). La variable `Deficit` debe codificarse como una variable dicotómica, que toma el valor 0 cuando la valoración inicial ha sido suficiente y 1 cuando la valoración inicial ha sido insuficiente.

La variable `Deficit` será la variable dependiente del modelo. Analizad la calidad del modelo y las variables que son relevantes.

4.2 Interpretación

Interpretad el modelo ajustado. Concretamente, explicad la contribución de las variables explicativas con coeficiente estadísticamente significativo para predecir si la valoración inicial es insuficiente para cubrir el coste del siniestro.

4.3 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 (valoración inicial del coste insuficiente) si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 en caso contrario. Analizad la matriz de confusión y las medidas de 'sensitivity' y 'specificity'.

Nota: Tomad como categoría de interés que haya déficit en la valoración inicial del coste. Por tanto, déficit igual a 1 será el caso positivo en la matriz de confusión y 0 el caso negativo.

4.4 Predicción

¿Con que probabilidad la valoración inicial del siniestro será insuficiente para un hombre de 20 años de edad, soltero, sin hijos ni otros dependientes, con un salario semanal de 300 EUR, jornada partida, con 30 horas

semanales y cinco días a la semana, una clasificación del tiempo hasta la apertura del siniestro de "Muy lento", una baja que no es por depresión y una valoración inicial de 10000EUR?

---

5 Análisis de la varianza (ANOVA) de un factor

Vamos a realizar un ANOVA para contrastar si existen diferencias en la variable `CosteFinal` en escala logarítmica en función de la clasificación del siniestro en relación al tiempo transcurrido hasta la apertura. Seguid los pasos que se indican.

5.1 Hipótesis nula y alternativa

Escribid la hipótesis nula y la alternativa.

5.2 Modelo

Calculad el análisis de varianza, usando la función `aov` o `lm`. Interpretad el resultado del análisis, teniendo en cuenta los valores: Sum Sq, Mean SQ, F y Pr (> F).

5.3 Efectos de los niveles del factor

Calculad la variabilidad explicada por la variable `Clasificacion` sobre la variable `CosteFinal` mediante la métrica eta squared. Interpretad los resultados.

5.4 Contraste dos-a-dos

Como los factores han resultado significativos hay que hacer los contrastes de las comparaciones múltiples. Se puede utilizar la prueba de Tukey-Kramer que compara dos-a-dos las diferentes categorías de la variable. (Nota: por ejemplo, con la función `HSD.test()` del paquete **agricolae**).

5.5 Adecuación del modelo

Mostrad la adecuación del modelo ANOVA. Se pide lo siguiente:

- Análisis visual de normalidad de los residuos. Podéis usar la función plot sobre el modelo ANOVA calculado.
- Análisis visual de homocedasticidad de los residuos. Podéis usar plot sobre el modelo ANOVA calculado.
- Contraste de normalidad y homocedasticidad.

5.5.1 Normalidad de los residuos

El análisis visual de la normalidad de los residuos se puede hacer a partir del gráfico Normal Q-Q. Mostrad e interpretad este gráfico.

5.5.2 Homocedasticidad de los residuos

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

5.5.3 Contraste de normalidad

Se puede comprobar el supuesto de normalidad de los residuos con las pruebas estadísticas de Shapiro-Wilk o Lilliefors, entre otros. El supuesto de homocedasticidad se puede comprobar a partir de la prueba de Bartleet.

---

6 ANOVA multifactorial

A continuación, se desea evaluar el efecto sobre `CosteFinal` en escala logarítmica según `Sexo` combinado con el factor `RiesgoSM`. Seguid los pasos que se indican a continuación.

6.1 Análisis de los efectos principales y posibles interacciones

Dibujad en un gráfico la variable `CosteFinal` en escala logarítmica en función de `Sexo` y en función de `RiesgoSM`. El gráfico debe permitir evaluar si hay interacción entre los dos factores. Por ello, se recomienda seguir estos pasos:

1. Agrupad el conjunto de datos por `Sexo` y por `RiesgoSM`. Calculad el número de casos disponibles de cada combinación de factores.
2. Calculad la media de coste (en log) para cada grupo.
3. Mostrad en un gráfico el valor medio de la variable `CosteFinal` en escala logarítmica para cada factor.
4. Interpretad el resultado sobre si sólo hay efectos principales o hay interacción entre los factores. Si hay interacción, explicad cómo se observa esta interacción en el gráfico.

6.2 Cálculo del modelo

- Calculad el modelo incluyendo la interacción entre los factores.
- Medid el efecto de los factores sobre la variabilidad explicada del Coste final (en escala logarítmica).
- Análizad dos-a-dos las diferencias de medias entre los distintos factores.
- Adecuación del modelo. Realizar análisis visual de normalidad y homocedasticidad.

6.3 Interpretación de los resultados

---

7 Conclusiones

Resumid las conclusiones principales del análisis. Para ello, podéis resumir las conclusiones de cada uno de los apartados.

Puntuación de la actividad

- Apartados 1 y 2 (15%)
- Apartado 3 (15%)
- Apartado 4 (15%)
- Apartado 5 (15%)
- Apartado 6 (15%)
- Apartado 7 (15%)
- Calidad del informe dinámico (10%)