A4 - Análisis de varianza y repaso del curso

Enunciado

Semestre 2019.1

Índex

Introducción

El conjunto de datos Mid-Atlantic Wage Data contiene 11 variables supuestamente relacionadas con el salario bruto de un grupo de 3000 trabajadores hombres de la región del Atlántico Medio de Estados Unidos, correspondiente al año 2011. Las variables son:

- year (año en que se obtuvo el dato sobre el salario)
- age (edad del trabajador)
- maritl (estado civil: 1. Never Married, 2. Married, 3. Widowed, 4. Divorced, y 5. Separated)
- race (variable categórica que indica raza: 1. White, 2. Black, 3. Asian, y 4. Other)

- education (variable categórica sobre el nivel de educación: 1. < HS Grad, 2. HS Grad, 3. Some College, 4. College Grad, 5. Advanced Degree)
- region (región, solo mid-atlantic)
- jobclass (variable categórica tipo de trabajo: 1. Industrial, 2. Information)
- health (variable categórica indicando el nivel de salud: 1. <=Good, 2. >=Very Good)
- health_ins (variable catgórica que indica si la persona tiene seguro de salud: 1. Yes, 2. No)
- logwage (Logaritmo del salario del trabajador)
- wage (Salario, $1000s)

Nota importante a tener en cuenta para entregar la actividad: Es necesario entregar el fichero Rmd y el fichero de salida (PDF o html). El fichero de salida debe incluir: el código y el resultado de la ejecución del mismo (paso a paso). Se debe respetar la misma numeración de los apartados que el enunciado.

1 Lectura del fichero y tipo de variables

Leed el fichero `Wage.csv` el cual contiene los datos del estudio **Mid-Atlantic Wage Data**.
A continuación, verificad el tipo de cada variable. ¿Qué variables son de tipo numérico? ¿Qué variables son de tipo categórico? Realizad conversiones de tipo si es necesario.

2 Estadística descriptiva y visualización

2.1 Análisis descriptivo

Realizad un análisis descriptivo numérico de los datos (resumid los valores de las variables numéricas y categóricas).

2.2 Visualización

Mostrad con diversos diagramas de caja la distribución de la variable `wage` según: `race`, `jobclass`, `health` y `health_ins`. Interpretar los gráficos brevemente.

2.3 Comprobación de normalidad

¿Podemos asumir que la variable `wage` tiene una distribución normal? Justificar la respuesta a partir de métodos visuales.

3 Estadística inferencial

3.1 Intervalo de confianza de la variable `age`

a) Calcular el intervalo de confianza al 95% de la variable `age` de los trabajadores. A partir del valor obtenido, explicad como se interpreta el resultado del intervalo de confianza.

b) Calcular los intervalos de confianza al 95% de la variable `age`, segregando los trabajadores por la variable `jobclass`. ¿Qué conclusión se puede extraer de la comparación de los dos intervalos, en relación a si existe solapamiento o no en los intervalos de confianza? Justificad la respuesta.

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

3.2 Contraste de hipótesis para la diferencia de medias

¿Podemos aceptar que los trabajadores que tienen contratado un seguro médico variable (`health_ins`) tienen un salario (`wage`) que supera en más de 20$ (en miles de dólares) el salario de los que no tienen seguro médico? Calcularlo para un nivel de confianza del 95%.

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

Se asumirá que la variable `wage` tiene distribución normal. Seguid los pasos que se detallan a continuación.

3.2.1 Escribid la hipótesis nula y la alternativa

3.2.2 Justificar qué método aplicaríais

3.2.3 Cálculos

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

3.2.4 Interpretación

Interpretad los resultados.

3.3 Contraste no paramétrico

3.3.1 Aplicación de contraste no paramétrico

Aplicad un contraste no paramétrico para responder la misma pregunta anterior. Podéis usar funciones R.

3.3.2 Interpretar el resultado

3.3.3 Paramétrico vs no paramétrico

Justificad qué tipo de contraste (paramétrico/no paramétrico) se debería aplicar en este caso.

4 Regresión logística

4.1 Modelo predictivo

Ajustad un modelo predictivo basado en regresión logística para predecir la probabilidad de tener un salario superior a la media en función de las variables: `health_ins`, `jobclass` y `age`. Tomad como salario medio

el valor de la media muestral de la variable `wage`. Podéis codificar como 0 cuando el salario es inferior a la media y 1 cuando el salario es superior o igual.

4.2 Interpretación

Interpretad el modelo ajustado. Concretamente, explicad la contribución de las variables explicativas para predecir si el salario es superior o inferior a la media.

4.3 Importancia del nivel de estudios

Añadid al modelo anterior la variable `education`. Interpretad los niveles de la variable `education` a partir del **odds ratio**. ¿En qué porcentaje se ve incrementada la probabilidad de tener un salario superior al salario medio según el nivel educativo? Proporcionad intervalos de confianza del 95% de los odds ratio.

4.4 Predicción

¿Superaría el salario medio un trabajador con seguro médico, que trabaja en el ámbito de la información y con 42 años de edad y formación de graduado? ¿Y si se trata de un trabajador del ámbito industrial?

5 Análisis de la varianza de un factor (ANOVA)

5.1 Nivel de educación y salario

Seleccionar las observaciones que tengan un salario inferior a 150000$. Para este grupo de trabajadores realizad un Anova para contrastar si existen diferencias en el salario según el nivel de educación.

5.1.1 Hipótesis nula y alternativa

5.1.2 Modelo

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

5.1.3 Cálculos

Para profundizar en la comprensión del modelo ANOVA, calcular manualmente la suma de cuadrados intra y la suma de cuadrados entre grupos. Los resultados deben coincidir con el resultado del modelo ANOVA. Como referencia, podéis obtener las fórmulas de López-Roldán i Fachelli (2015), páginas 29-33.

5.1.4 Interpretación de los resultados

5.2 Adecuación del modelo

Mostrad visualmente la adecuación del modelo ANOVA. Podéis usar **plot** sobre el modelo ANOVA calculado. En los apartados siguientes, realizad la interpretación de estos gráficos.

5.2.1 Normalidad de los residuos

Interpretar la normalidad de los residuos a partir del gráfico Normal Q-Q que habéis mostrado en el apartado anterior.

5.2.2 Homocedasticidad de los residuos

Los gráficos "Residuals vs Fitted", "Scale-Location" y "Residuals vs Factor levels" proporcionan información sobre la homcedasticidad de los residuos. Interpretad estos gráficos.

5.3 ANOVA no paramétrico

Si las asunciones de normalidad y homocedasticidad no se cumplen, se puede aplicar un contraste no paramétrico como el test de Kruskal-Wallis.

5.3.1 Test Kruskal-Wallis

Aplicad el test de Kruskal-Wallis para contrastar si hay diferencias en el salario según la raza (`race`). Como en el apartado anterior, seleccionad las observaciones con salario inferior a 150 y además, descartad los casos en que la variable `race` tenga el valor "4. Other". Podéis usar funciones R que calculen el test Kruskal-Wallis.

5.3.2 Interpretación de los resultados

Interpretad los resultados del test Kruskal-Wallis.

6 ANOVA multifactorial

A continuación, se desea evaluar el efecto de la raza combinado con otro factor. Primero se realizará el análisis con el factor tipo de trabajo (`jobclass`) y posteriormente, con el factor nivel de educación (`education`). En este apartado seleccionad las observaciones con un salario inferior a 150 y además, descartad las observaciones en que la variable `race` tengan el valor "4. Other", como se ha hecho anteriormente.

6.1 Factores: raza y tipo de trabajo

6.1.1 Análisis visual de los efectos principales y posibles interacciones

Dibujar en un gráfico la variable `wage` en función de la raza (`race`) y en función del tipo de trabajo (`jobclass`). El gráfico ha de permitir evaluar si hay interacción entre los dos factores. Por este motivo, se deben seguir los pasos siguientes:

1. Agrupar el conjunto de datos por raza y por tipo de trabajo. Calcular la media de salario para cada grupo. Para realizar este proceso, se pueden usar las funciones `group_by` y `summarise` de la librería `dplyr`.
2. Mostrar el conjunto de datos en forma de tabla, donde se muestre la media de cada grupo según raza y tipo de trabajo.
3. Mostrar en un gráfico el valor medio de la variable `wage` para cada raza y tipo de trabajo. Podés inspiraros en los gráficos de López-Roldán y Fachelli (2015), p.38. Podéis realizar este tipo de gráficos usando la función `ggplot` de la librería `ggplot2`.

4. Interpretar el resultado sobre si hay efectos principales o existe interacción entre los factores. Si hay interacción, explicad como se observa esta interacción en el gráfico.

6.1.2 Modelo ANOVA

Aplicar un modelo anova con estos factores y su posible interacción. A continuación, analizar si la interacción es significativa.

6.1.3 Adecuación del modelo

Interpretar la adecuación del modelo ANOVA obtenido usando los gráficos de los residuos.

6.2 Factores: raza y nivel de educación

Seguid los mismos pasos que el apartado anterior para aplicar un modelo ANOVA con los factores raza y nivel de educación.

6.2.1 Análisis visual de los efectos principales y posibles interacciones

6.2.2 Modelo ANOVA

6.2.3 Adecuación del modelo

7 Comparaciones múltiples

Tomando como referencia el modelo ANOVA multifactorial, con los factores raza y tipo de trabajo, aplicar el test de comparación múltiple Scheffé. Interpretar el resultado del test e indicar qué grupos son significativamente diferentes entre si.

8 Conclusiones

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

9 Comentarios importantes sobre la actividad

1. No se puede inspeccionar ni corregir de manera manual el fichero de datos. Por ejemplo, no se pueden realizar instrucciones de este tipo:

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

   Este tipo de transformaciones se deben hacer con funcionalidades de búsqueda (buscar los registros que tienen errores o inconsistencias) y luego hacer las correcciones oportunas con funcionalidades de R. Así el procedimiento es útil, independientemente del fichero de datos y de la posición y valores concretos del archivo.

2. **No se pueden hacer listados completos de los datos del fichero a pantalla**, porque generan archivos de salida excesivamente grandes. Si se desean validar los resultados de una instrucción sobre los datos, se puede usar la función **head** que muestra las primeras filas de la tabla de datos o **tail** que muestra las últimas.

Puntuación de la actividad

- Apartados 1 y 2 (10%)
- Apartado 3 (10%)
- Apartado 4 (10%)
- Apartado 5 (10%)
- Apartado 6 (20%)
- Apartado 7 (20%)
- Apartado 8 (10%)
- Calidad del informe dinámico (10%)