Examen Final. Análisis de Datos. ECO-A.

2022-12-19

Ejercicio 1 (1.5 pts)

La base de datos `cars` **de R** relaciona la velocidad de un vehículo con la distancia necesaria para el frenado. Contiene 50 valores de velocidad y su correspondiente distancia de frenado. En concreto, las variables toman los siguientes nombres

- `speed`: velocidad en millas por hora.
- `dist`: distancia de frenado en pies.

Así, por ejemplo, la observación con `speed` = 24 y `dist` = 93, indica que un vehículo que viaje a 24 millas por hora, necesita una distancia de 93 pies para frenar por completo.

Ajusta un modelo lineal en el que la **variable respuesta** sea la **distancia de frenado** y la **variable predictora** sea la velocidad. La fórmula del modelo sería:

$dist = \beta_0 + \beta_1 \cdot speed$

¿Cuál es el valor del coeficiente β_1? ¿Cómo se interpreta este valor?

Nota: Los datos están en R bajo el nombre `cars` y por tanto no necesitarás cargalos. No obstante, en Canvas encontrarás un fichero con los datos llamado `cars.csv` que solo utilizarás si no pudieses acceder a los datos `cars`.

Interpretación: si incrementamos la velocidad en 1 milla por hora, el modelo predice que la distancia de frenado aumentará en 3.93 pies.

Ejercicio 2 (1.25 pts)

Utilizando el modelo anterior, construye una gráfica donde el eje x corresponda a la velocidad y el eje y a la distancia de frenado. En esta gráfica han de aparecer:

- Los valores reales de distancia de frenado y velocidad.
- Una línea roja que represente las predicciones de distancia de frenado para cada velocidad según el modelo del ejercicio anterior.

Ejercicio 3 (1.25 pts)

Calcula y visualiza los residuos del modelo del ejercicio 1 mediante una gráfica donde en el eje x aparezca la variable `speed` y en el eje y los residuso correspondientes.

¿Qué conclusión extraes?

Conclusión: al no observarse patrones claros en la distribución de los residuos, concluímos que el modelo ha captado satisfactoriamente los patrones presentes en los datos.

Ejercicio 4 (1.5 pts)

El fichero `salarios.csv` contiene 534 observaciones de las siguientes variables (recogidas en EEUU en el año 1985) para diferentes individuos:

- `wage` salario (en dólares por hora)
- `educ` número de años de educación recibidos
- `race` raza, con dos niveles "NW" (no blanco) o "W" (blanco)
- `sex` sexo, con niveles "F" (femenino) "M" (masculino)
- `hispanic` se refiere a si la persona es hispánica o no. Contiene dos niveles: "Hisp" (Hispánico) "NH" (No hispánico)
- `south` se refiere a si la persona es o no del sur de EEUU. Contiene dos niveles: "NS" (no es del sur) y "S" (sí es del sur).
- `married` indica si la persona está o no casada. Contiene dos niveles: "Married" (Casado), "Single" (Soltero)
- `exper` número de años de experiencia laboral
- `union` indica si la persona pertenece o no a un sindicato. Cintiene dos niveles: "Not" (no pertenece a sindicato) "Union" (sí pertenece a un sindicato)
- `age` edad en años
- `sector` indica el sector al que pertenece la persona. Posibles niveles: "clerical", "const", "manag", "manuf", "other", "prof", "sales" y "service"

Importa los datos a R y construye una tabla que contenga el **salario medio** de la **gente de raza blanca** y el de la **gente de raza no blanca**. Comenta lo que observas.

Conclusión: La gente de raza blanca cobra más, en media, que la gente de otras razas.

Ejercicio 5 (1.25 pts)

Considerando únicamente la gente menor de 40 años, crea una tabla donde aparezca el salario mediano de la **gente de raza blanca** y el de la **gente de raza no blanca** según **pertenezcan o no a un sindicato**. Comenta lo que observas.

Conclusión: El salario mediano de la gente afiliada a un sindicato es mayor que el de los no afiliados. Además, independientemente de la pertenencia a un sindicato, la gente blanca sigue cobrando más, en mediana, que la gente de otras razas.

Ejercicio 6 (1.25 pts)

Usando los datos de los salarios del ejercicio anterior, crea una nueva base de datos que cumpla lo siguiente:

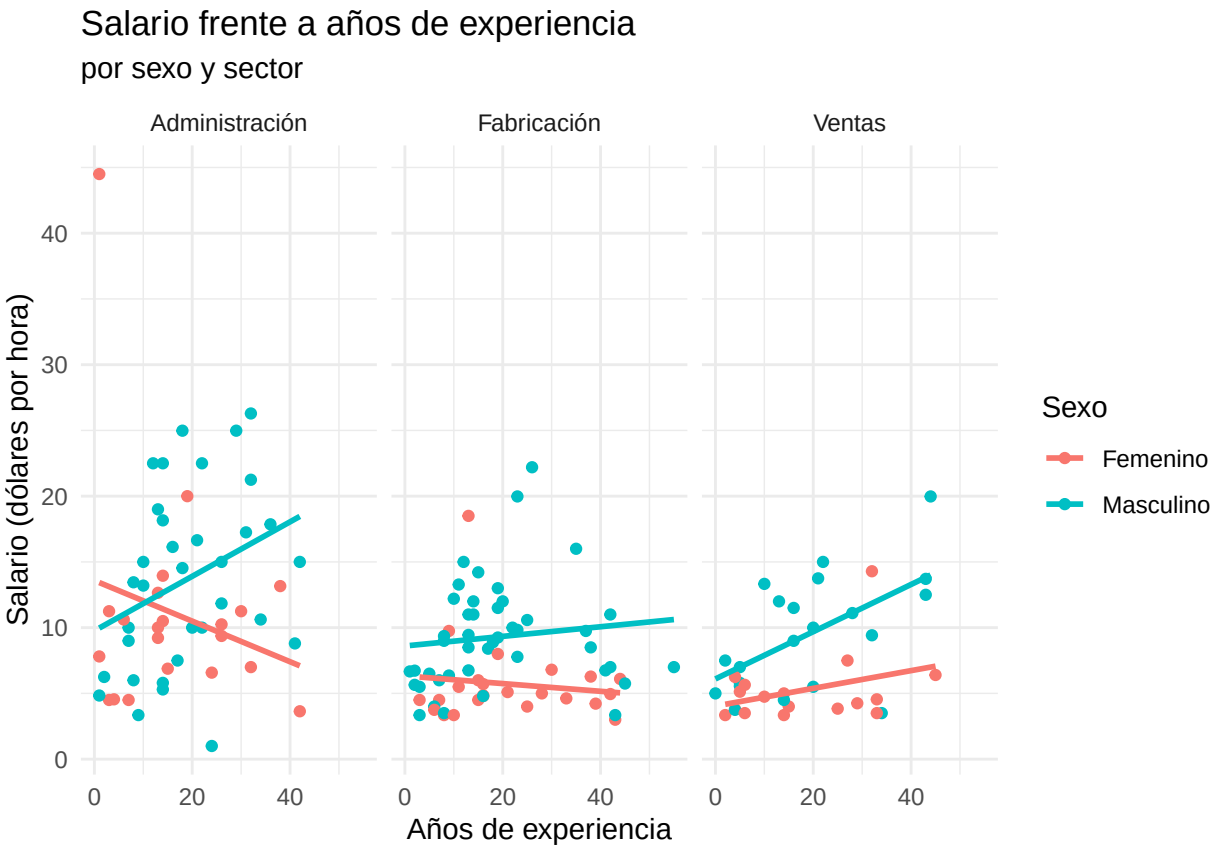
- Contenga una variable de tipo factor llamada `Sexo` creada a partir de la variable `sex`, con dos niveles "Masculino" si sex = "M" y "Femenino" si sex = "F".
- Contenga únicamente observaciones para las cuales el sector es o bien "manag", o bien "manuf" o bien "sales".
- Contenga una variable de tipo factor llamada `Sector` creada a partir de la variable `sector`, con tres niveles "Administración" si sector = "manag", "Fabricación" si sector = "manuf" y "Ventas" si sector = "sales".

Pista: para convertir una variable llamada X que no es factor a una llamada Y que sí lo es, puede usar `mutate(Y = factor(X))`.

Ejercicio 7 (3 pts)

Utilizando los datos creados en el ejercicio anterior, reproduce siguiente gráfica.

Extrae dos conclusiones válidas.



Conclusión:

- En la mayoría de sectores, y prácticamente de forma independiente a los años de experiencia, las mujeres cobran menos que los hombres.
- En el sector de la adminsitración, el salario de los hombres crece con los años de experiencia mientras que el de las mujeres, decrece.