Trabajo Práctico 1: Regresión lineal

Enfoque Estadístico del Aprendizaje

Fecha y hora límite de primera entrega: 19 de octubre de 2023 a las 23:59 p.m.

Fecha y hora límite de entrega con penalización: 29 de octubre de 2023 a las 23:59 p.m.

Fecha y hora de devolución general: 11 de noviembre de 2023 a las 11 a.m.

INSTRUCCIONES

Deberán realizar el trabajo en un **RNotebook** y entregarlo en formato R Notebook o HTML

El **RNotebook** debe:

- Tener el siguiente nombre: eea2023_tp1_**apellido_nombre**
- Mostrar todo el código que escriban. NO USAR `echo=FALSE`
- Ser ordenado de acuerdo a las consignas propuestas

Una vez terminada la notebook deberán enviarla por mail a eea.uba@gmail.com

CRITERIOS DE EVALUACION

- Explicar los procedimientos y decisiones en el texto
- Comentar el código
- Llegar a los resultados esperados
- Recomendamos fuertemente usar las funciones de `tidyverse`

En caso que los resultados no sean los esperados y no logremos identificar las fuentes de error podemos pedirles que nos compartan el archivo .Rmd y ciertas bases de datos que vayan generando.

DATOS

Los datos con los que se trabajará en este TP provienen de la Encuesta Permanente de Hogares (EPH) provistos por el Instituto de Estadísticas y Censos (INDEC) de la República Argentina. **Link a los datos**

La EPH es una encuesta muestral que permite conocer las características sociodemográficas y socioeconómicas. Recomendamos leer brevemente el **informe** para los datos que se van a utilizar en el trabajo, donde podrán encontrar análisis descriptivos y el diccionario de variables

Los datasets que se comparten corresponden a un recorte del dataset original luego de un preprocesamiento específico para las consignas de este trabajo.

Las variables incluidas son:

- **codusu**: ID de la observación
- **ano4** : Año de relevamiento
- **trimestre** : Trimestre de relevamiento
- **region**: región de residencia
- **aglomerado**: aglomerado urbano de residencia
- **fecha_nacimiento**: Fecha de nacimiento
- **edad**: Años cumplidos

- **asistencia_educacion**: ¿Asiste o asistió a algún establecimiento educativo?
- **nivel_ed**: Nivel educativo en que se encuentra la persona
- **tipo_establecimiento**: ¿El negocio / empresa / institución / actividad en la que trabaja es público o privado?
- **codigo_actividad**: Código de actividad económica (Clasificador de Actividades Económicas para Encuestas Sociodemográficas del Mercosur)
- **sexo**: Sexo (binario)
- **categoria_ocupacion**: Categoría ocupacional
- **cat_cantidad_empleos**: ¿La semana pasada tenía 1 o tenía múltiples empleos?
- **alfabetismo**: ¿Sabe leer y escribir?
- **educacion**: Años de educación estimados
- **experiencia_potencial**: Estimación de la experiencia laboral, calculada a partir de la diferencia entre la edad y una estimación de los años de educación
- **salario_horario**: ingreso por hora trabajada en el mes (de la ocupación principal y ocupaciones secundarias)

El diccionario de todas las variables que forman parte de la base de datos cruda se encuentra en el siguiente **documento**

CONSIGNAS

El objetivo general del trabajo es poder crear una serie de modelos lineales para explicar y predecir el **salario horario** de las personas según la información que proporciona la EPH para el tercer trimestre del año 2022.

Análisis exploratorios

1) **Análisis estructura y correlación**

Leer el archivo "eph_train_2022.csv". ¿Qué puede mencionar sobre su estructura y variables?

¿Cómo es la correlación entre las variables numéricas? Utilice y analice en detalle algún gráfico que sirva para sacar conclusiones sobre la asociación de variables realizando apertura por sexo.

En particular, ¿Cómo es la correlación entre la variable a explicar (salario_horario) y el resto de las variables numéricas?

Modelos

Un modelo clásico del salario es la llamada ecuación de Mincer. Existen varias especificaciones pero la más típica es:

$$E\left[\ln(salario)\right] = \beta_0 + \beta_1 AñosEducacion + \beta_2 ExperienciaLaboral + \beta_3 ExperienciaLaboral^2$$

En los siguientes consignas la idea es ir aproximandose a esta lógica de modelado

2) **Modelos lineales experiencia**

Se va a comenzar con dos modelos lineales que utilicen la información de la experiencia potencial.

Primero, ajustar un modelo de regresión para explicar el salario por hora usando únicamente la experiencia potencial como covariable.

$$E(SalarioHorario) = \beta_0 + \beta_1 ExperienciaPotencial$$

Luego, ajustar otro modelo en donde las únicas covariables sean la experiencia potencial y el cuadrado de la experiencia potencial.

$$E(SalarioHorario) = \beta_0 + \beta_1 ExperienciaPotencial + \beta_2 ExperienciaPotencial^2$$

Responder las siguientes preguntas en base a ambos modelos:

¿Cuál es el impacto de un año adicional de experiencia potencial en el salario horario esperado para cada uno de estos modelos?

¿Cuál es el efecto sobre el salario horario esperado de un año más de experiencia laboral para una persona con 6 años de experiencia laboral? ¿Y para una persona con 35 años de experiencia laboral?

3) **Modelo lineal múltiple**

Se plantea un primer modelo múltiple a partir de la ecuación de Mincer:

$$E(SalarioHorario) = \beta_0 + \beta_1 AñosEducacion + \beta_2 ExperienciaPotencial + \beta_2 ExperienciaPotencial^2 +$$
$$\beta_3 Sexo + \beta_4 Sexo \cdot AñosEducacion$$

Ajustar el modelo planteado y responder las siguientes preguntas: ¿Cuál es la interpretación de las variables incluidas en el modelo? ¿Sus coeficientes son significativos? ¿El modelo resulta significativo para explicar el salario? ¿Qué porcentaje de la variabilidad explica el modelo?

Analizar en profundidad el cumplimiento de los supuestos del modelo lineal para este modelo

4) **Modelo de Mincer "enriquecido"**

Ahora, se procede a modelar según una especicación del modelo de Mincer con ciertas variables adicionales

$$E\left[ln(SalarioHorario)\right] = \beta_0 + \beta_1 AñosEducacion + \beta_2 ExperienciaPotencial + \beta_2 ExperienciaPotencial^2 +$$
$$\beta_3 Sexo + \beta_4 Sexo \cdot AñosEducacion$$

- ¿Cuál es la interpretación del coeficiente asociado a la variable de años de educación? ¿Se observan cambios en la significatividad individual de los coeficientes respecto al modelo anterior?
- ¿Qué porcentaje de la variabilidad del salario horario explica el modelo? ¿Cómo se compara con la variabilidad explicada por el modelo anterior?

  Nota: tenga en cuenta que la variable predicha es el logaritmo del salario horario y se pide el porcentaje de variabilidad explicada del salario horario. Además, como los dos modelos tienen la misma cantidad de covariables es posible compararlos mediante el el R-cuadrado simple
- Analizar en profundidad el cumplimiento de los supuestos del modelo lineal para este modelo y comparar con el análisis del modelo anterior

5) **Modelos propios y evaluación**

Realizar 2 modelos lineales múltiples adicionales y explicar la lógica detrás de los mismos (se valorará la creación y/o inclusión de variables nuevas).

Nota: No se pueden utilizar métodos de selección automática de variables dado que buscamos que analicen otras variables y realicen feature engineering.

Evaluar y comparar la performance del **modelo lineal multiple**, el **modelo de mincer** y los modelos desarrollados en este punto en el dataset de entrenamiento y evaluación (usar dataset "eph_test_2022.csv"). La evaluación de performance consiste en comparar la performance en términos del RMSE y MAE sobre el set de entrenamiento y el set de evaluación.

¿Cuál es el mejor modelo para el objetivo de predecir el salario horario? ¿Por qué?

6) **Modelo lineal robusto**

Leer el archivo "eph_train_outliers_2022.csv". Este último consiste en el dataset original de train con la incorporación de algunas observaciones adicionales que pueden incluir valores atípicos.

Realizar dos gráficos del salario horario, uno para el dataset de entrenamiento sin outliers y otro para el dataset con outliers que permitan observar claramente la diferencia entre ambos sets de datos.

Sobre este nuevo conjunto de datos entrenar el **modelo lineal multiple**, el **modelo de mincer** y un **modelo robusto** (misma especificación que el modelo lineal multiple). Comparar exhaustivamente los coeficientes estimados y su significatividad entre el **modelo lineal multiple** y el **modelo robusto**.

Comparar la performance (RMSE y MAE) de los tres modelos entrenados en este punto en el dataset de entrenamiento (con outliers) y de evaluación ¿Qué puede concluir al respecto?