**Modelo de salarios individuales por hora**

* w = salario por hora
* X = matriz de posibles variables explicativas/predictores
* Centrados en

1. **Introducción.**

Exposición breve del problema y si hay antecedentes. Describe brevemente los datos y su idoneidad para abordar la cuestión del conjunto de problemas. Contiene una vista previa de los resultados y las principales conclusiones.

1. **Datos.**

**Restricción:** personas empleadas mayores de 18 años

¿Datos faltantes o salarios en $0?

* Restringir los datos a estos individuos y realizar un análisis descriptivo de las variables utilizadas en el conjunto de problemas.
* Describir brevemente la GEIH, incluyendo su propósito y cualquier otra información relevante.
* Describir el proceso de adquisición de los datos y si hay alguna restricción para acceder a estos datos.
* Describir el proceso de limpieza de datos
* Un análisis descriptivo de los datos. Incluir una tabla de estadísticas descriptivas con su interpretación, un análisis profundo que ayude al lector a comprender los datos, su variación y la justificación de las elecciones de datos.

1. **Perfil Edad-Salario**

* Tabla de regresión
* Interpretación de los coeficientes y su significado.
* Discusión del modelo en el ajuste de muestra.
* Un gráfico del perfil estimado de edad-ingresos implícito en la ecuación anterior. Incluir discusión de la "edad pico" con sus respectivos intervalos de confianza. (Bootstrap para intervalos de confianza).

1. **Brecha salarial por género**

* *Female* = 1, cuando el individuo es mujer.
* *¿Igual salario por igual labor?* Estimar una brecha de ingresos condicional incorporando variables de control como características similares del trabajador y del trabajo. Estimar la brecha salarial condicional:
  + Primero, usando FWL
  + Segundo, FWL con boostrap. Comparar las estimaciones y los errores estándar.
* Trazar el perfil de edad-salario calculado y estimar las "edades pico" implícitas con los respectivos intervalos de confianza por género.
* Incluir:
* Tabla de regresión, con las estimaciones lado a lado de las diferencias salariales condicionales y no condicionales, destacando el coeficiente de interés. Los controles de "ruido" no deben incluirse en la tabla, sino anotarse.
* Interpretación de los coeficientes "femeninos", comparación entre los modelos y el ajuste en la muestra.
* Discusión sobre las edades pico implícitas y su similitud / diferencia estadística.
* Discusión reflexiva sobre la brecha salarial incondicional y condicional, buscando responder si los cambios en el coeficiente son evidencia de un problema de selección, un "problema de discriminación", una mezcla o ninguno de estos problemas.

1. **Predicción de ganancias**

* Dividir la muestra en dos: una muestra de entrenamiento (70%) y una muestra de prueba (30%). (establecer una semilla para lograr la reproducibilidad).
* Informar y comparar el rendimiento predictivo de todas las especificaciones anteriores con al menos cinco (5) especificaciones adicionales que exploren las no linealidades y la complejidad.
* Discusión de resultados:
  + Acerca de la métrica de rendimiento elegida y justificación para elegirla.
  + Acerca de la especificación con el error de predicción más bajo.
  + Para la especificación, explorar aquellas observaciones que parecen "perder la marca"; calcular los errores de predicción en la muestra de prueba y examinar su distribución. ¿Están las observaciones en las colas de la distribución de errores de predicción? ¿Son estos valores atípicos personas potenciales que la DIAN debería investigar, o son solo el producto de un modelo defectuoso?
* LOOCV. Para los dos modelos con el error predictivo más bajo en la sección anterior, calcular el error predictivo utilizando Leave-one-out-cross-validation (LOOCV). Comparar los resultados del error de prueba con los obtenidos con el enfoque de conjunto de validación y explorar los posibles vínculos con la estadística de influencia.