**Step-by-step**

**Acquisition de datos:**

El proyecto extrae datos de la muestra GEIH 2018.

Ejecute el script que extrae y limpia los datos, que se encuentra en la carpeta de scripts dentro del repositorio. Para realizar el scraping utilizamos Selenium en python.

Si no quiere scrapear los datos puede encontrarlos sin filtrar en la carpeta stores, como “GEIH2018\_sample\_data.csv”.

**Limpieza de datos:**

Siga los scripts proporcionados para limpiar los datos extraídos. Ejecute el script de limpieza de datos, o si prefiere trabajar con la base ya filtrada puede encontrarla en stores como “GEIH2018\_filtered\_data.csv”. En este paso también se realizan ciertas trasformaciones a los datos, se recomienda leer el script para entender los datos con los que se trabaja.

**Estadisticas descriptivas y tablas:**

Con la base de datos filtrada en el script "descriptivas y graficas preliminares.R” se hace una exploracion preliminar de los datos y se realizan graficas descriptivas de las variables relevantes.

**Punto 3**

Para el punto 3 se corre la regresión que se pide, logaritmo del salario en función de la edad y la edad al cuadrado. Para construir el peak- age se deriva respecto a la edad y se iguala a 0 y se reemplaza los betas calculados en la regresión y luego se usa Bootstrap para calcular una distribución de la peak age, es decir de este mismo proceso en 5000 muestras tomadas de la originas. Con la distribución se confirma la media y os intervalos de confianza.

La totalidad de del punto se soliciona con el codigo “punto 3” el cual tiene tanto la estimacion como los graficos

**Punto 4**  
  
**Estimación de la Regresión**  
Se estima una ecuación que relaciona el salario con la edad y el género. Primero, se utiliza un modelo incondicional que considera únicamente la edad y el género. Luego, se estima un modelo condicional que incluye controles adicionales como el nivel educativo, el tipo de empresa y el tamaño de la empresa para ajustar los efectos de estas variables sobre el salario.

**Comparación de Resultados**  
Se comparan los resultados entre el modelo incondicional y el modelo condicional para evaluar cómo la inclusión de controles adicionales afecta la brecha salarial de género. Además, se comparan los resultados obtenidos con el método FWL y el método FWL con bootstrap. Esto permite verificar la consistencia y precisión de las estimaciones del coeficiente de género y los errores estándar. La comparación ayuda a identificar si los cambios en los resultados son robustos frente a diferentes métodos de estimación.

**Análisis de Coeficientes**  
Se examina el cambio en el coeficiente de la variable de género (Female) entre los modelos incondicional y condicional. Se determina si hay una diferencia significativa que podría indicar problemas de selección y discriminación. Los cambios en los coeficientes bajo los diferentes ajustes de los modelos (solo género, incorporación de controles y análisis de modelos condicional e incondicional) permiten identificar si la discriminación y la selección están influyendo en los resultados.

**Estimación de Edades Pico**  
Se estima la edad pico para hombres y mujeres, calculando los intervalos de confianza utilizando métodos de bootstrap. Esto proporciona una medida más robusta y precisa de la edad a la que los salarios alcanzan su punto máximo para cada género. Se discute cómo las diferencias en la edad pico entre géneros pueden reflejar posibles problemas de discriminación en el mercado laboral.

**Discusión de Resultados**  
Se reflexiona sobre los resultados obtenidos de los modelos y la estimación de edades pico. Se considera cómo las diferencias en los coeficientes de género y las edades pico podrían ser indicativas de problemas de selección y discriminación en el mercado laboral.

**Punto 5**

Para este punto luego de dividir la muestra, se usa backward elimination para reducir el número de variables lo más posible después de añadir el cuadrado; luego, se utiliza Lasso para determinar cuáles son las que menos aportan y así reducir aún más el modelo. Se vuelve a hacer backward en las sobrevivientes a Lasso y por último se eliminan las que tiene una correlación alta para terminar con un conjunto de 17 variables sobre las cuales se proponen diversas especificaciones con interacciones y no linealidades para obtener los 2 modelos del último punto. Para todas los modelos se valida con el test y se clasifica según el rmse.

Esta implementación se hizo en python con las librerias sklearn y statsmodels y el punto hasta el c se encuentra en el archivo de python “Punto 5”

**Discusiones importantes (5c)**  
  
En este punto, se realizó un análisis de los resultados y sus respectivas comparaciones, así como una interpretación de los hallazgos. Para la discusión de los errores, se partió de los scripts utilizados en el punto anterior, pero se realizaron modificaciones para generar las distribuciones de los errores estándar y así llevar a cabo un análisis sobre los posibles outliers.

**LOOCV:**

Siguiendo el script “Script LOOCV RMSE test.R” utilizando la base de datos filtrada cargue la librería boot que se usará para la validación cruzada. Se definen dos modelos de regresión lineal generalizada para predecir el logaritmo del salario utilizando diferentes variables. Se aplica (LOOCV) para evaluar la precisión de ambos modelos. Finalmente, se calcula el error cuadrático medio (MSE) y la raíz del error cuadrático medio (RMSE) para cada modelo, proporcionando una medida de su desempeño predictivo. Posteriormente se compara con los modelos anteriores para realizar el analisis.