Taller 01 - Estimación del Salario de los Ocupados en el Área Metropolitana del Valle de Aburrá

Estudiante

Jesús Andrés Álvarez Alvarado

Docente

Osmar Leandro Loaiza Quintero

Asignatura

Econometría Avanzada

Sede Medellín
10 de abril de 2024

Tabla de Contenidos

Lista de Figuras

Lista de Tablas

1 Variables, Limpieza de Datos y Cruce de Módulos

Lo primero será seleccionar las variables, filtrar por área y en general, alistar los datos para ingresarlos al modelo.

1.1 Variables

Las variables seleccionadas se encuentran en dos módulos, *Características generales, seguridad social en salud y educación* y *Ocupados*. La descripción completa de los módulos y en detalle de cada variable se encuentra en los metadatos (diccionario). Las variables son:

Del módulo de *Características generales, seguridad social en salud y educación*:

- *DIRECTORIO*, *SECUENCIA_P* y *ORDEN*: son los identificadores de la vivienda, hogar y persona, respectivamente. Estas variables se concatenan en una sola para obtener un identificador y este se utiliza a modo de llave para unir los diferentes módulos. La unión de estas variables se nombrará como **id_persona** en los datos limpios

- *AREA*: es la variable que permite filtrar los datos por área. En este caso, como se trabaja con el Área Metropolitana del Valle de Aburrá, se filtrará por el valor 05 de la variable. Conservará el nombre **area** en los datos limpios

- *P3271*: ¿cuál es el sexo al nacer? Para capturar si hay brechas salariales asociadas al sexo. Es una variable cualitativa que toma el valor de 1 si responde *Masculino* y 2 si responde *Femenino*. Se renombrará por **sexo** en los datos limpios y el grupo base será 1 = *Masculino*

- *P6040*: ¿cuántos años cumplidos tiene? Para capturar diferencias salariales entre edades. Es una variable numérica. Se renombrará por **edad** en los datos limpios

- *P6050*: ¿cuál es el parentesco con el jefe del hogar? Para capturar cómo los roles en el hogar pueden influenciar el salario. Es una variable cualitativa que toma el valor de 1 para *Jefe (a) del hogar*; 2 para *Pareja, esposo(a), cónyuge, compañero(a)*; 3 para *Hijo(a), hijastro(a)*; 4 para *Nieto(a)*; 5 para *Otro pariente*; 6 para *Empleado(a) del servicio doméstico y sus parientes*; 7 para *Pensionista*; 8 para *Trabajador*; y 9 para *Otro no pariente*. Se renombrará por **jefe_hogar** en los datos limpios y el grupo base será 9 = *Otro no pariente*

- *P6070*: estado civil. Es una variable cualitativa que toma el valor de 1 para *No está casado(a) y vive en pareja hace menos de dos años*; 2 para *No está casado (a) y vive en pareja hace dos años o más*; 3 para *Está casado (a)*; 4 para *Está separado (a) o divorciado (a)*; 5 para *Está viudo (a)*; y 6 para *Está soltero (a)*. Se renombrará por **estado_civil** en los datos limpios y el grupo base será 6 = *Está soltero (a)*

- *P6080*: ¿se reconoce dentro de alguna etnia? Para capturar brechas salariales entre diferentes etnias. Es una variable cualitativa que toma el valor de 1 para *Indígena*; 2 para *Gitano (a) (Rom)*; 3 para *Raizal del archipiélago de San Andrés, Providencia y Santa Catalina*; 4 para *Palenquero (a) de San Basilio*; 5 para *Negro (a), mulato (a) (afrodescendiente), afrocolombiano(a)*; y 6 para *Ninguno de los anteriores*. Se renombrará por **etnia** en los datos limpios y el grupo base será 6 = *Ninguno de los anteriores*

- *P3042*: ¿cuál es el mayor nivel educativo alcanzado y el último grado o semestre aprobado? Para capturar si se cumple la teoría del capital humano. Es una variable cualitativa que toma el valor de 1 para *Ninguno*; 2 para *Preescolar*; 3 para *Básica primaria (1o - 5o)*; 4 para *Básica secundaria (6o - 9o)*; 5 para *Media académica (Bachillerato clásico)*; 6 para *Media técnica (Bachillerato técnico)*; 7 para *Normalista*; 8 para *Técnica profesional*; 9 para *Tecnológica*; 10 para *Universitaria*; 11 para *Especialización*; 12 para *Maestría*; 13 para *Doctorado*; y 99 para *No sabe, no informa*. Se renombrará por **educacion** en los datos limpios y el grupo base será 1 = *Ninguno*

- *Experiencia potencial*: pretende ser una aproximación a la experiencia laboral en base a la edad y el nivel educativo, dado que no se pregunta directamente cuántos años de experiencia tiene la persona. Es una variable numérica que se calcula así: si la persona tiene un nivel educativo de *Básica secundaria* o menor será la `edad - 15`; si la persona tiene un nivel educativo entre *Bachillerato clásico* y *Normalista* será la `edad - 17`; si la persona tiene un nivel educativo entre *Técnica Profesional* y *Tecnológica* será la `edad - 20`; y, si la persona tiene un nivel educativo de *Universitaria* o superior será la `edad - 22`. Se nombrará **exp_potencial** en los datos limpios

Del módulo de *Ocupados*:

- *DIRECTORIO*, *SECUENCIA_P* y *ORDEN*: son los identificadores de la vivienda, hogar y persona, respectivamente. Estas variables se concatenan en una sola para obtener un identificador y este se utiliza a modo de llave para unir los diferentes módulos. La unión de estas variables se nombrará como **id_persona** en los datos limpios

- *AREA*: es la variable que permite filtrar los datos por área. En este caso, como se trabaja con el Área Metropolitana del Valle de Aburrá, se filtrará por el valor 05 de la variable. Conservará el nombre **area** en los datos limpios

- *INGLABO*: será la variable respuesta, ya que hace referencia a los ingresos laborales mensuales de la persona. Es una variable numérica. Se renombrará por **salario** en los datos limpios

- *P6440*: ¿para realizar este trabajo tiene usted algún tipo de contrato? Pretende ser una aproximación para capturar información de formalidad e informalidad. Es una variable cualitativa que toma el valor de 1 si responde que *Sí* y 2 si responde que *No*. Se renombrará por **contrato** en los datos limpios y el grupo base será 2 = *No*

- *P6920*: ¿está cotizando actualmente a un fondo de pensiones? También intenta aproximar a las personas que son formales e informales. Pues, es poco común que los informales coticen a pensión. No se utiliza afiliación a salud, pues, existe una gran cobertura de salud bajo régimen subsidiado en Colombia, lo que no permite segmentar correctamente formales de informales. Es una variable cualitativa que toma el valor de 1 si responde que *Sí*; 2 si responde que *No*; y 3 si responde que *Ya es pensionado*. Se renombrará por **cotiza_pension** en los datos limpios y el grupo base será 3 = *Ya es pensionado*

- *P6960*: ¿cuántos años lleva afiliado al fondo de pensiones? Se intenta aproximar la experiencia por medio de los años cotizados a pensión para contrastar también con la experiencia potencial. Además, porque la experiencia potencial supone que las personas trabajan de manera continua sin periodos de desempleo. Se debe anotar que esta variable excluirá a la mayoría de población informal, pues estos por lo general no cotizan. Es una variable numérica. Se renombrará por **exp_cotiza** en los datos limpios

1.2 Limpieza de Datos

Una vez elegidas las variables, lo siguiente es alistar los datos para realizar el cruce. Lo primero es leer solo las variables seleccionadas (se tiene la precaución de leer toda la base como caráter, para evitar que se eliminen 0 a la izquierda en columnas de tipo texto). Después, filtrar los datos por área, pues las áreas diferentes al Área Metropolitana del Valle de Aburrá y las demás variables no son de interés para este estudio.

Posteriormente, se crea la variable **id_persona** al concatenar las variables descritas anteriormente. Luego, se eliminan las variables que se utilizaron para crear la llave, ya que no serán de utilidad por separado y se renombran las variables que incluirá el modelo como se indicó anteriormente.

También, se cambia la clase de cada variable a la que le corresponde y se establecen los grupos base en las variables de tipo cualitativas.

Finalmente, una vez las demás variables están listas, se construye la variable *exp_potencial* al restar la edad promedio de los graduados en cada nivel educativo a la edad

Tabla 1: Datos Módulo de Características Generales

id_persona	area	sexo	edad	jefe_hogar	estado_civil	etnia	educacion	exp_potencial
718591311	05	1	59	1	3	6	3	44
718591312	05	2	51	2	3	6	3	36
718591313	05	1	30	3	6	6	9	10
718591314	05	1	24	3	6	6	5	7
718591315	05	1	68	6	5	6	3	53
718591411	05	1	68	1	3	6	11	46

Tabla 2: Datos Módulo de Ocupados

id_persona	area	salario	contrato	cotiza_pension	exp_cotiza
718591311	05	1000000	2	1	15
718591313	05	900000	2	2	NA
718591511	05	1500000	1	2	NA
718591611	05	2000000	1	1	28
718591912	05	1000000	2	1	20
718591915	05	1000000	2	1	7

En las tablas 1 y 2 se puede observar el alistamiento de cada módulo por separado.

1.3 Cruce de módulos

Con los datos ya limpios, se procede a realizar un inner join con la función `merge` del paquete `base`. De esta manera se unirán las variables de los dos módulos solo en los casos en que el **id_persona** se encuentre en ambos módulos.

Tabla 3: Datos Cruce 1

	id_persona	area	sexo	edad	jefe_hogar	estado_civil	etnia
1	718591311	05	1	59	1	3	6
2	718591313	05	1	30	3	6	6
3	718591511	05	1	70	1	3	6
4	718591611	05	2	52	1	4	6
5	718591912	05	1	55	3	6	6
6	718591915	05	2	41	3	6	6

Tabla 4: Datos Cruce 2

	educacion	salario	contrato	cotiza_pension	exp_cotiza	exp_potencial
1	3	1000000	2	1	15	44
2	9	900000	2	2	NA	10
3	10	1500000	1	2	NA	48
4	8	2000000	1	1	28	32
5	5	1000000	2	1	20	38
6	10	1000000	2	1	7	19

Realizada la unión, en las tablas 3 y 4 se puede visualizar la estructura final de los datos de las variables de interés.

2 Análisis Descriptivo de las Variables

2.1 Variables Numéricas

2.1.1 Salario (INGLABO)

Tabla 5: Resumen de la Variable Salario

Mínimo	Cuartil 1	Mediana	Promedio	Cuartil 3	Máximo	NA's	Ceros
0	1.000.000	1.100.000	1.805.544	1.800.000	66.000.000	26	16

En la tabla 5 se observa que la variable salario tiene un claro sesgo positivo, que se nota por la diferencia entre media y mediana. Este se produce por la aparicion de datos de ingreso muy elevados y se deberán tratar de manera adecuada para evitar que se conviertan en datos influyentes en la regresión.

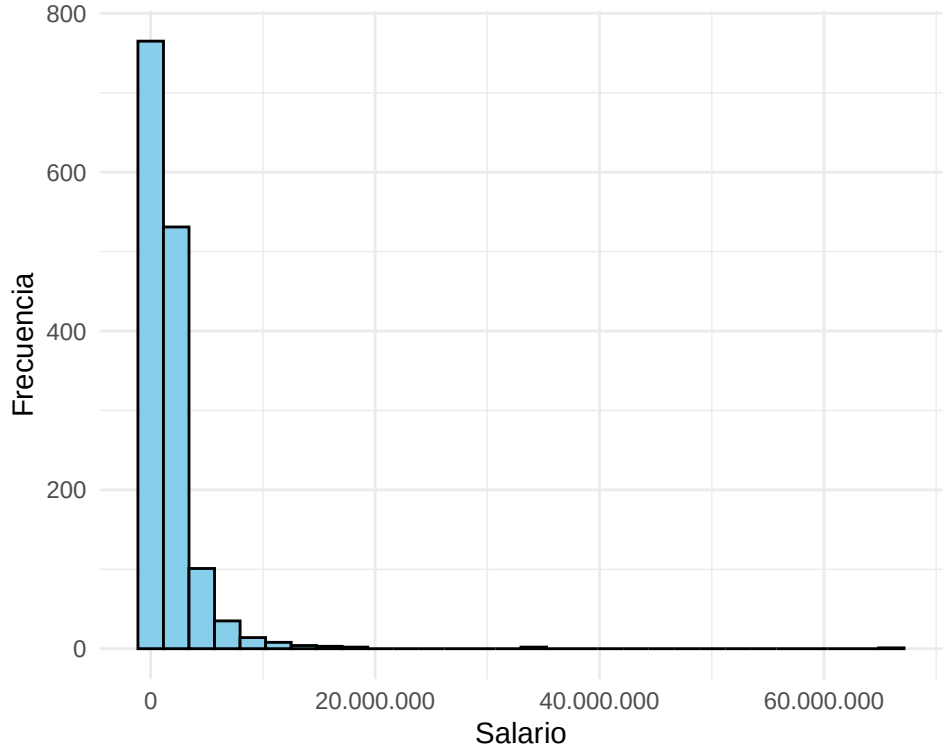

Figura 1: Histograma de Salario

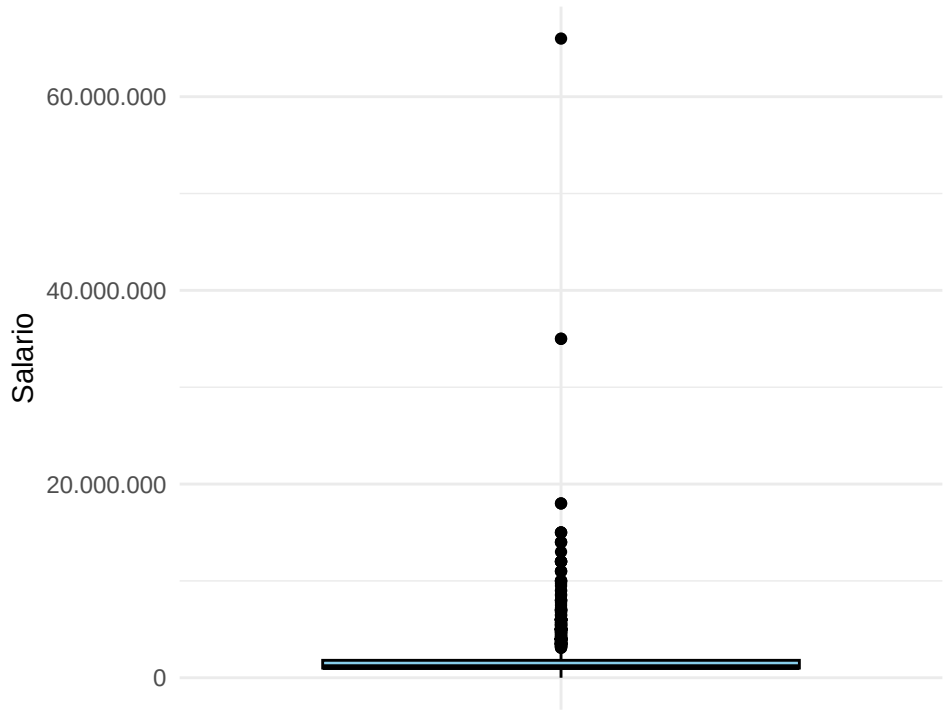

Figura 2: Diagrama de Caja y Bigotes de Salario

En las figuras 1 y 2 se confirma el patrón de sesgo de cola derecha, en donde hay presencia de datos extremos en la distribución. La variable salario se debe incluir después de tomar logaritmo natural, lo cual es común en este tipo de estimación.

Lo primero es eliminar los NA's y valores en 0, dado que no aportarán información al modelo.

Después, mediante el método de Tukey se identifican los valores extremos al calcular un límite inferior y superior mediante la multiplicación de el rango intercuartílico con el cuartil 1 y 3:

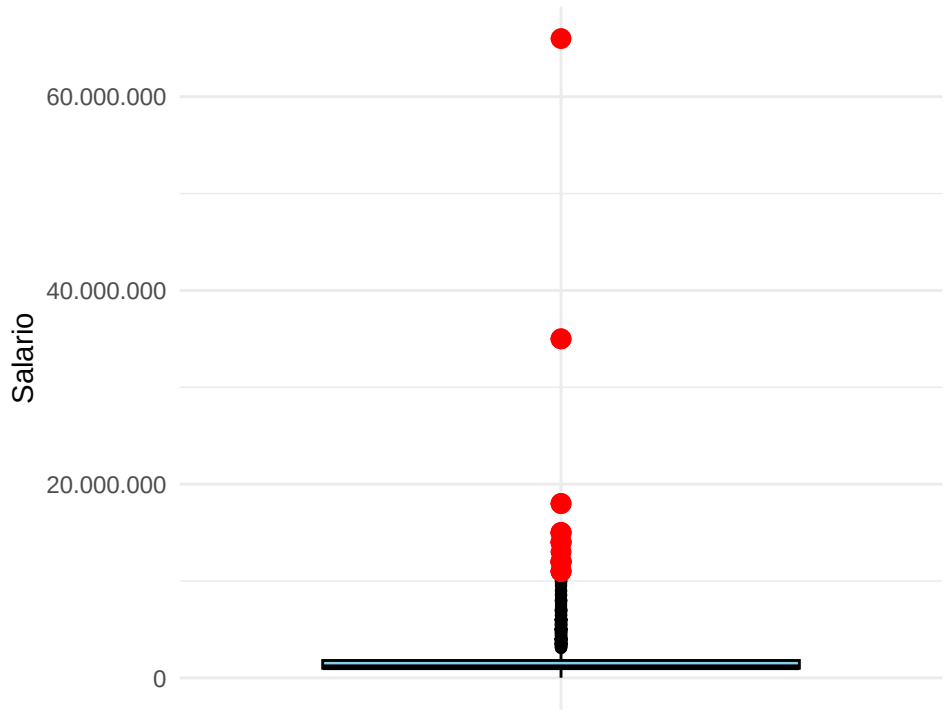

Figura 3: Diagrama de Caja y Bigotes de Salario con Valores Extremos

Luego de identificar los valores extremos potencialmente influyentes, se aplica el método de winsorización, donde los valores extremos se reemplazan por los límites correspondientes.

Finalmente, corregidos los valores extremos se procede a aplicar logaritmo natural y se deja la variable lista para ingresar al modelo. En la figura 4 se observa que el sesgo se eliminó y que es probable que la estimación mejore por la no presencia de datos extremos:

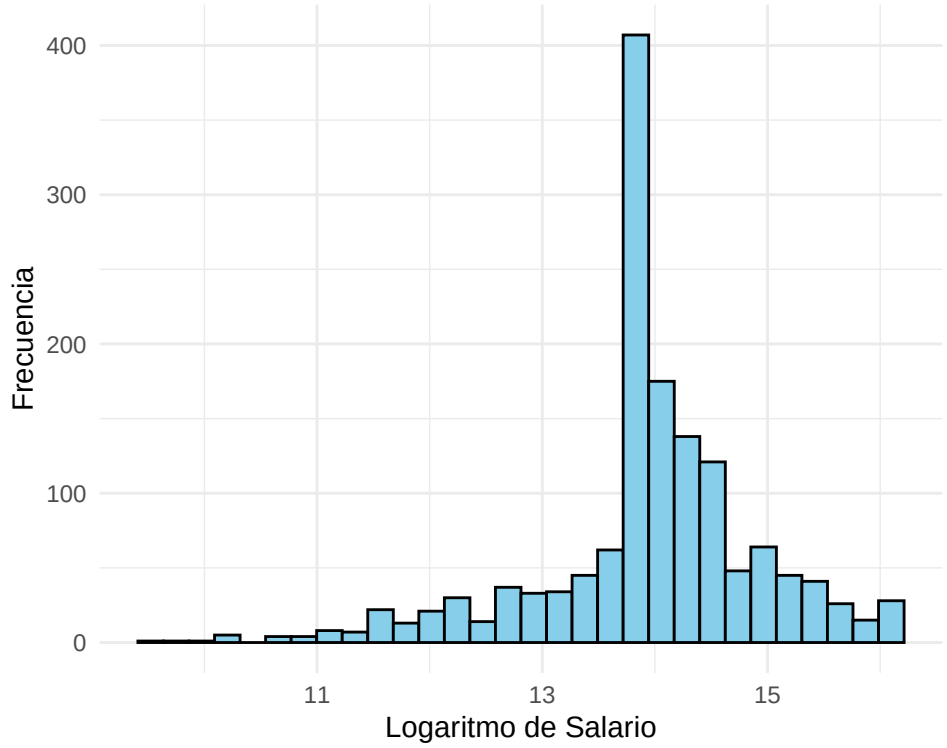

Figura 4: Histograma del Logaritmo de Salario

2.1.2 Edad

Tabla 6: Resumen de la Variable Edad

Mínimo	Cuartil 1	Mediana	Promedio	Cuartil 3	Máximo	NA's
15	29	39	40.33448	51	83	0

En la tabla 6 se observa que la variable edad no tiene sesgo, que se evidencia por la diferencia entre media y mediana. Tampoco tiene datos faltantes, por lo que se intuye que está lista para ingresar al modelo.

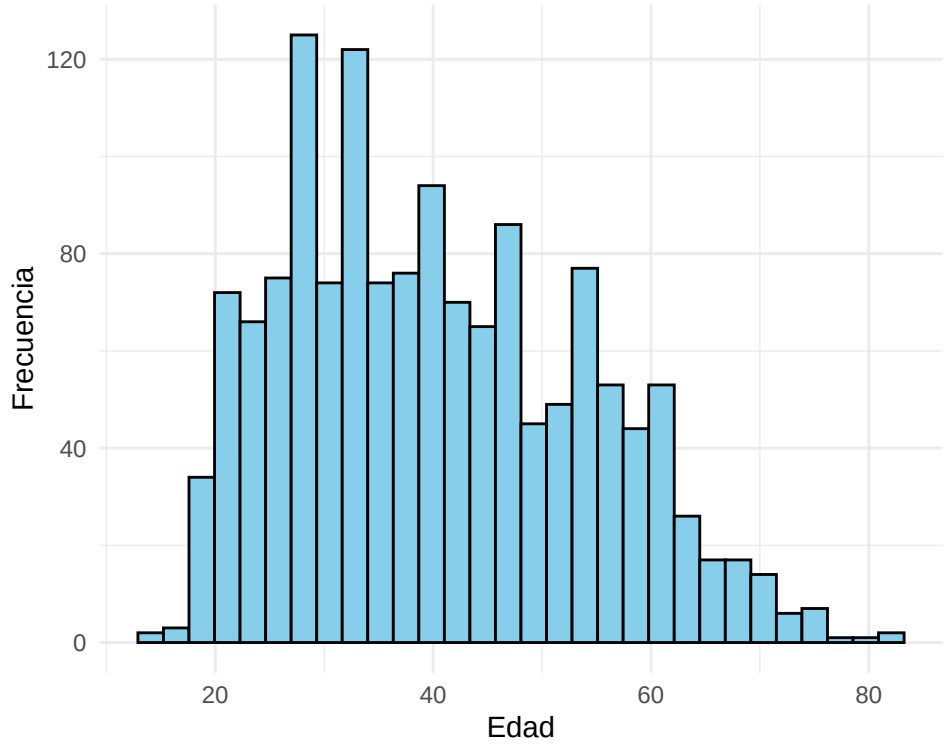

Figura 5: Histograma de Edad

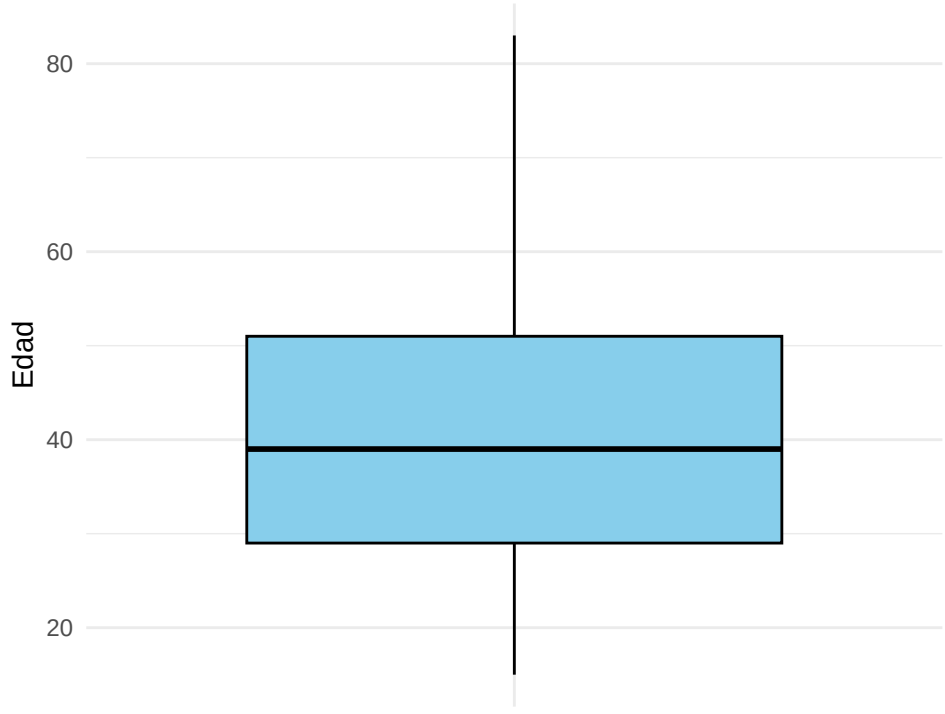

Figura 6: Diagrama de Caja y Bigotes de Edad

En las figuras 5 y 6 se evidencia que no hay señales de sesgo y que la variable no necesita transformación ni tratamiento de datos.

2.1.3 Experiencia Potencial