UNIVERSIDAD DE GRANADA

ECONOMETRÍA

DOBLE GRADO EN INGENIERÍA INFORMÁTICA Y GADE

MODELO PARA LA PREDICCIÓN DE SOBREPESO

Autores

JOSÉ ÁNGEL CARRETERO MONTES, ISMAEL SALLAMI MORENO, FERNANDO JOSÉ GRACIA CHOIN

FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

Granada, a 2 de diciembre de 2024

ÍNDICE GENERAL

1

INTRODUCCIÓN

1.1 INTRODUCCIÓN

La obesidad es un problema de salud pública que ha adquirido proporciones alarmantes en las últimas décadas, afectando a personas de todas las edades y contextos socioeconómicos. Este fenómeno no solo está vinculado a un mayor riesgo de desarrollar enfermedades crónicas, como la diabetes tipo 2 y enfermedades cardiovasculares, sino que también representa una carga significativa para los sistemas de salud a nivel global. Para abordar este problema, es esencial entender los factores que influyen en su desarrollo. En este trabajo, se propone la construcción de un modelo econométrico cuyo objetivo es estimar el grado de obesidad en función de diversas variables relacionadas con características individuales y hábitos de vida. Entre las variables incluidas en el análisis se encuentran la edad, altura, peso, antecedentes familiares de sobrepeso, hábitos alimenticios, nivel de actividad física, consumo de agua, uso de dispositivos tecnológicos, y otros factores clave que pueden contribuir a la obesidad. A través de este enfoque, se busca identificar los determinantes más relevantes que influyen en la obesidad, lo que podría facilitar la implementación de políticas y estrategias de intervención más eficaces para combatir este problema creciente.

1.2 VARIABLES

1.2.1 *Dependientes*

El objetivo del modelo será estimar el **grado de obesidad** en función de varias variables. Por tanto, la variable dependiente será cualquiera que represente de forma fiel el grado de obesidad que sufra un determinado individuo. En nuestro caso, será el peso.

1.2.2 *Independientes*

Para estimar el grado de obesidad de una persona, hemos identificado las siguientes variables independientes como relevantes:

- **Edad (Age)**: Variable continua que representa la edad de la persona.
- **Altura (Height)**: Variable continua que representa la altura de la persona.
- **Antecedentes familiares de sobrepeso (family_history_with_overweight)**: Variable binaria que indica si algún familiar ha sufrido o sufre de sobrepeso.
- **Consumo frecuente de alimentos calóricos (FAVC)**: Variable binaria que indica si la persona consume frecuentemente alimentos altos en calorías.
- **Consumo de vegetales (FCVC)**: Variable entera que indica la frecuencia con la que la persona consume vegetales en sus comidas.
- **Número de comidas principales (NCP)**: Variable continua que indica cuántas comidas principales tiene la persona al día.
- **Consumo de alimentos entre comidas (CAEC)**: Variable categórica que indica si la persona consume alimentos entre comidas.
- **Consumo de agua (CH_2O)**: Variable continua que indica la cantidad de agua que la persona bebe diariamente.
- **Actividad física (FAF)**: Variable continua que indica la frecuencia con la que la persona realiza actividad física.
- **Uso de dispositivos tecnológicos (TUE)**: Variable entera que indica la cantidad de tiempo que la persona utiliza dispositivos tecnológicos como teléfonos móviles, videojuegos, televisión, entre otros.
- **Consumo de alcohol (CALC)**: Variable categórica que indica con qué frecuencia la persona consume alcohol.
- **Medio de transporte (MTRANS)**: Variable categórica que indica el medio de transporte que la persona utiliza usualmente.

1.3 DETALLES DE LA BASE DE DATOS

1.3.1 *Delimitar el tema sobre el que se desea trabajar*

El tema central de este estudio es analizar y predecir el **grado de obesidad** (que mediremos con el peso) en la población a partir de diversos factores relacionados con el estilo de vida, los hábitos alimenticios y las características físicas. Este trabajo se centrará en identificar las variables más relevantes que influyen en el desarrollo de la

obesidad, así como en construir un modelo que permita predecir el nivel de obesidad en base a dichas variables.

El estudio se enfoca en el análisis de datos obtenidos de individuos de diversas edades y antecedentes, utilizando un conjunto de variables independientes que incluyen factores como la actividad física, el consumo de alimentos y agua, entre otros.

1.3.2 *Búsqueda de bibliografía básica sobre el tema*

Para este estudio, se ha utilizado un **dataset** obtenido de la plataforma **Kaggle**, la cual es conocida por proporcionar conjuntos de datos de alta calidad en una amplia variedad de campos. El **dataset** en cuestión contiene información detallada sobre variables relacionadas con la obesidad y los hábitos de vida de individuos.

Puede acceder a la base de datos aquí.

1.3.3 *Especificación inicial del modelo.*

El modelo econométrico propuesto sigue la forma de un **modelo lineal múltiple**, que se puede expresar de la siguiente manera:

$$Y_t = \beta_0 + \beta_1 \cdot \text{Edad}t + \beta_2 \cdot \text{Altura}_t + \beta_3 \cdot \text{AntFam}_t + \beta_4 \cdot \text{FAVC}_t + \cdots + \beta_k \cdot Xkt + u_t \quad (1)$$

Donde:

- Y_t representa el **peso** de la persona t, que es nuestra variable dependiente.
- β_0 es el **término independiente**, que refleja el valor esperado de Y_t cuando todas las variables explicativas son iguales a cero.
- $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ son los **coeficientes de regresión** que indican el cambio promedio en Y_t dado un cambio en cada variable explicativa.
- X_{kt} son las **variables independientes** del modelo (Edad, Altura, Peso, Antecedentes familiares, etc.).
- u_t es el **término de error**, que capta los factores no observados que influyen en el grado de obesidad y no están recogidos por las variables explicativas.

Cada uno de los coeficientes β_k mide el impacto parcial de la correspondiente variable explicativa en el grado de obesidad, manteniendo las demás constantes. Por ejemplo, β_3 mide el efecto marginal del peso sobre la obesidad, controlando por edad, altura y las demás variables.

Este modelo lineal múltiple permitirá estimar cómo estos factores influyen en el grado de obesidad y evaluar cuáles son los más significativos.

Para estimar los parámetros, usaremos el método de mínimos cuadradors ordinarios: El modelo de **Mínimos Cuadrados Ordinarios (MCO)** es una técnica econométrica ampliamente utilizada para estimar los parámetros de un modelo lineal. Este método se basa en la minimización de la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo. En el contexto de nuestro estudio sobre el grado de obesidad, utilizaremos MCO para estimar cómo las variables independientes, como la edad, altura, peso y hábitos de alimentación, influyen en la variable dependiente, que es el grado de obesidad.

La ventaja del MCO radica en su simplicidad y en la capacidad de proporcionar estimaciones eficientes y consistentes de los parámetros del modelo, siempre que se cumplan ciertos supuestos, como la linealidad, la homocedasticidad y la independencia de los errores.

En la *figura 1* se presenta un gráfico que ilustra la relación entre una variable independiente (por ejemplo, el peso) y el grado de obesidad, donde se puede observar la nube de puntos y la recta de regresión que representa la relación estimada por el modelo.

1.3.4 *Disponibilidad de datos del fenómeno.*

Para el estudio vamos a usar una base de datos de naturaleza de sección cruzada o también conocida como **transversal**, la cual consiste en datos de mútiples individuos observados en un mismo instante de tiempo.

1.3.5 *Ámbito de aplicación.*

Hemos escogido una base de datos la cual se corresponde a un **ámbito local**, por ello esto nos permite estudiar la obesidad, y que factores influyen en ella, en las zonas geograficas de Perú, Colombia y México.

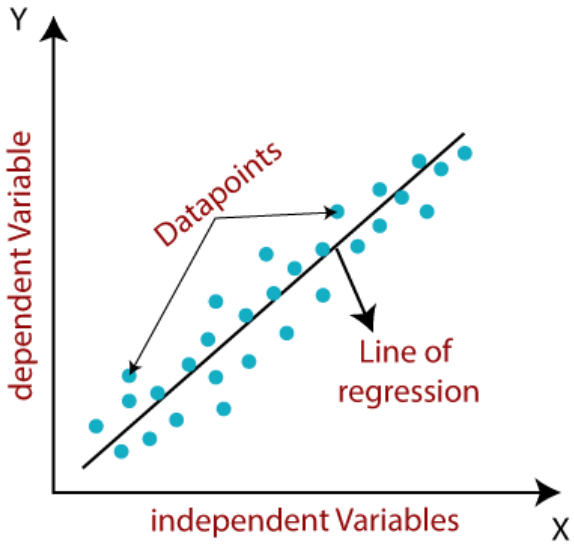


Figura 1: Nube de puntos y recta de regresión que representa la relación entre el peso y el grado de obesidad.

1.4 ELABORACIÓN DE LA BASE DE DATOS

1.4.1 *Datos Extraños*

En este caso hay ciertos datos que no concuerdan con la lógica real, como puede ser el caso de pesos fuera de lo común y otras variables con valores extraños. Por ende, a la hora de realizar el modelo tendremos cuidado con estos casos.

2

MODELO PRELIMINAR

2.1 ESTIMACIÓN DEL MODELO PRELIMINAR

El modelo de regresión lineal múltiple estimado se representa de la siguiente forma:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$$

donde:

- Y_i es la variable dependiente (Peso en este caso),
- $X_{1i}, X_{2i}, \ldots, X_{ki}$ son las variables independientes (Altura, Edad, Género, etc.),
- β_0 es el término constante,
- u_i es el término de error aleatorio.

Los resultados preliminares de la regresión OLS para este modelo se presentan en la siguiente tabla:

OLS Regression Results

Dep. Variable:	Weight	R-squared:	0.576
Model:	OLS	Adj. R-squared:	0.572
Method:	Least Squares	F-statistic:	177.5
Date:	Tue, 22 Oct 2024	Prob (F-statistic):	0.00
Time:	07:09:09	Log-Likelihood:	-8983.6
No. Observations:	2111	AIC:	1.800e+04
Df Residuals:	2094	BIC:	1.810e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-207.9919	10.281	-20.230	0.000	-228.155	-187.829
Gender	-4.7575	1.026	-4.639	0.000	-6.769	-2.746
Age	0.7403	0.082	9.069	0.000	0.580	0.900
Height	125.2342	5.894	21.247	0.000	113.675	136.793
family_history_with_overweight	17.0101	1.115	15.259	0.000	14.824	19.196
FAVC	7.9764	1.252	6.373	0.000	5.522	10.431
FCVC	9.0480	0.750	12.070	0.000	7.578	10.518
NCP	1.0882	0.506	2.151	0.032	0.096	2.080
CAEC	8.2303	0.869	9.470	0.000	6.526	9.935
SMOKE	-0.2063	2.664	-0.077	0.938	-5.430	5.017
CH2O	1.3156	0.645	2.041	0.041	0.052	2.580
SCC	-6.9609	1.878	-3.706	0.000	-10.645	-3.277
FAF	-2.7221	0.484	-5.619	0.000	-3.672	-1.772
TUE	-1.6760	0.652	-2.571	0.010	-2.954	-0.398
CALC	-4.1794	0.764	-5.472	0.000	-5.677	-2.682
MTRANS	3.6899	0.382	9.658	0.000	2.941	4.439
NObeyesdad	2.3519	0.220	10.680	0.000	1.920	2.784

Omnibus:	37.671	Durbin-Watson:	0.741
Prob(Omnibus):	0.000	Jarque-Bera (JB):	21.704
Skew:	-0.047	Prob(JB):	1.94e-05
Kurtosis:	2.512	Cond. No.	812.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

2.2 INTERPRETACIÓN DE LOS ESTIMADORES

A continuación, se ofrece una interpretación de los coeficientes más relevantes del modelo:

- **Constante**: El coeficiente de la constante es -207.99, lo que indica el valor estimado del peso cuando todas las variables independientes son 0. Aunque en la práctica, esta interpretación rara vez es útil, ya que variables como la altura no pueden ser cero.

- **Género**: El coeficiente de Género es -4.757. Esto indica que, manteniendo todas las demás variables constantes, ser mujer (asumiendo que el género está codificado como 1 para mujeres y 0 para hombres) está asociado con una reducción de 4.76 unidades en el peso.
- **Edad**: El coeficiente de 0.7403 indica que por cada año adicional de edad, el peso aumenta en promedio 0.74 unidades, manteniendo las demás variables constantes.
- **Altura**: El coeficiente de altura es 125.23, lo que sugiere que, por cada unidad adicional de altura, el peso aumenta en 125.23 unidades, manteniendo todas las demás variables constantes.
- **Historial Familiar de Sobrepeso**: El coeficiente es 17.01, indicando que tener antecedentes familiares de sobrepeso está asociado con un aumento promedio de 17.01 unidades en el peso.
- **FAVC (Comida grasa frecuente)**: El coeficiente positivo de 7.976 indica que consumir frecuentemente comida grasa aumenta el peso en 7.976 unidades, manteniendo todo lo demás constante.
- **FCVC (Consumo de verduras)**: El coeficiente de 9.048 indica que un mayor consumo de verduras está asociado con un aumento de 9.05 unidades en el peso, lo cual es inesperado y podría requerir mayor análisis.
- **NCP (Número de comidas diarias)**: El coeficiente de 1.088 sugiere que, por cada comida adicional consumida al día, el peso aumenta en promedio 1.088 unidades.
- **CAEC (Comer entre comidas)**: El coeficiente de 8.230 sugiere que comer entre comidas aumenta el peso en 8.23 unidades, manteniendo el resto de factores constantes.
- **Fumar (SMOKE)**: El coeficiente es -0.206, indicando que fumar no tiene un impacto estadísticamente significativo en el peso, dado su alto valor p ($P > 0{,}05$).
- **Consumo de agua (CH2O)**: El coeficiente de 1.315 indica que por cada unidad adicional de agua consumida, el peso aumenta en 1.315 unidades, lo cual puede ser contraintuitivo y merece un análisis más profundo.
- **SCC (Conteo de calorías)**: El coeficiente de -6.961 sugiere que el control de calorías está asociado con una reducción del peso en 6.96 unidades, lo cual es consistente con lo esperado.
- **FAF (Actividad física)**: El coeficiente de -2.722 indica que un aumento en la actividad física reduce el peso en promedio 2.72 unidades, manteniendo constantes las demás variables.

- **TUE (Uso de tecnología)**: El coeficiente de -1.676 sugiere que más tiempo usando tecnología está asociado con una reducción del peso, lo cual puede requerir más análisis.
- **CALC (Consumo de alcohol)**: El coeficiente de -4.179 indica que consumir alcohol está asociado con una reducción del peso, lo cual es un resultado inesperado.
- **MTRANS (Medios de transporte)**: El coeficiente de 3.689 indica que el uso de ciertos medios de transporte se asocia con un aumento del peso en 3.689 unidades.
- **NObeyesdad (Nivel de obesidad)**: El coeficiente de 2.351 indica que un mayor nivel de obesidad previo está asociado con un aumento de 2.351 unidades en el peso.

2.3 MEDIDAS DE BONDAD DE AJUSTE

Para evaluar la calidad del modelo estimado, utilizamos diversas medidas de bondad de ajuste que nos permiten juzgar cómo de bien el modelo ajusta a los datos observados. Las medidas que se consideran son el coeficiente de determinación R^2, el coeficiente de determinación ajustado R^2_{ajustado}, y los criterios de información de Akaike (AIC) y Schwarz (BIC).

2.3.1 *Coeficiente de Determinación R^2*

El coeficiente de determinación R^2 mide la proporción de la varianza total de la variable dependiente que es explicada por el modelo. En otras palabras, indica qué tan bien las variables independientes explican las variaciones en la variable dependiente.

$$R^2 = 1 - \frac{\text{Suma de Cuadrados de los Residuos (SCR)}}{\text{Suma Total de Cuadrados (SCT)}}$$

Donde:

- La **Suma Total de Cuadrados (SCT)** mide la variabilidad total de la variable dependiente en torno a su media:

$$\text{SCT} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- La **Suma de Cuadrados de los Residuos (SCR)** mide la variabilidad que no ha sido explicada por el modelo:

$$\text{SCR} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Para este modelo, se ha obtenido un R^2 de **0.576**, lo que significa que el **57.6 %** de la variabilidad del peso puede explicarse por las variables independientes incluidas en el modelo. Aunque este valor indica que el modelo captura más de la mitad de la variación en el peso, hay un **42.4 %** de la variabilidad que no está siendo explicada, lo que sugiere que existen otros factores que no se han tenido en cuenta o que pueden requerir un tratamiento más sofisticado.

2.3.2 *Coeficiente de Determinación Ajustado $R^2_{ajustado}$*

El coeficiente de determinación ajustado R^2_{ajustado} ajusta el R^2 para el número de predictores en el modelo. Esto es importante porque el R^2 tiende a aumentar al agregar más variables, incluso si esas variables no tienen una relación significativa con la variable dependiente. El R^2_{ajustado} penaliza el uso de variables adicionales, lo que permite obtener una mejor medida del poder explicativo real del modelo.

$$R^2_{\text{ajustado}} = 1 - \frac{\frac{\text{SCR}}{n-k}}{\frac{\text{SCT}}{n-1}}$$

Donde:

- n es el número de observaciones,
- k es el número de variables explicativas incluidas en el modelo.

El valor obtenido para el R^2_{ajustado} es de **0.572**, lo cual es muy cercano al valor de R^2. Esto indica que el modelo no está sobreajustado (es decir, no ha añadido variables innecesarias), ya que el R^2_{ajustado} no disminuye considerablemente respecto al R^2 original.

2.3.3 *Criterio de Información de Akaike (AIC)*

El criterio de información de Akaike (AIC) es una medida que penaliza la complejidad del modelo. Es útil cuando se comparan varios modelos, ya que tiene en cuenta

tanto la bondad del ajuste (medida por la suma de cuadrados de los residuos) como el número de parámetros estimados. Cuanto menor sea el valor del AIC, mejor es el modelo.

$$AIC = \ln\left(\frac{\text{SCR}}{n}\right) + \frac{2k}{n}$$

Donde:

- n es el número de observaciones,
- k es el número de parámetros estimados (incluyendo la constante).

En este modelo, el valor de **AIC** es $\mathbf{1.800 \times 10^4}$. Para comparar diferentes modelos, se elegiría aquel que minimice este valor. Un menor AIC indica que el modelo no solo ajusta mejor los datos, sino que también evita la inclusión innecesaria de parámetros.

2.3.4 *Criterio de Información de Schwarz (BIC)*

El criterio de información bayesiano de Schwarz (BIC) es otra medida que, como el AIC, penaliza la complejidad del modelo, pero de manera más estricta. El BIC introduce una penalización más fuerte por la inclusión de más variables en el modelo, lo que lo hace útil para evitar el sobreajuste.

$$BIC = \ln\left(\frac{\text{SCR}}{n}\right) + \frac{k}{n} \cdot \ln(n)$$

En este modelo, el valor del **BIC** es $\mathbf{1.810 \times 10^4}$. Al igual que con el AIC, un menor valor de BIC es preferible, y su objetivo es identificar el modelo que ofrece el mejor ajuste sin sobrecargarlo con demasiadas variables.

2.3.5 *Comparación y Selección del Mejor Modelo*

Las medidas de bondad de ajuste nos indican diferentes aspectos de la calidad del modelo. En resumen:

- Un R^2 de 0.576 indica que el modelo es razonablemente bueno para explicar la variabilidad del peso.
- El R^2_{ajustado} de 0.572 sugiere que no hay un sobreajuste significativo.

- Los criterios de información AIC y BIC nos permiten comparar este modelo con otros posibles modelos. Si generamos otros modelos, escogeremos aquel con los valores más bajos de AIC y BIC.

En función de estas medidas, este modelo tiene un ajuste aceptable, pero se puede mejorar probando otros modelos que optimicen las medidas de AIC y BIC.

2.4 CONCLUSIONES Y SELECCIÓN DEL MODELO

El modelo preliminar parece explicar bien la variabilidad del peso, con un R^2 razonable y estimadores significativos. Para mejorar el modelo, podemos considerar variables adicionales o la eliminación de aquellas que no resulten significativas en otros ensayos.

Además, el uso de AIC y BIC será clave para la selección del mejor modelo en función de los diferentes criterios de ajuste que hemos estudiado.

3

MULTICOLIENALIDAD

3.1 ANÁLISIS DE LA MULTICOLINEALIDAD

3.1.1 *Obtención de las principales medidas de diagnóstico de la multicolinealidad*

El análisis de multicolinealidad se llevó a cabo utilizando dos indicadores principales:

- **Número de Condición (Cond. No.)**: Calculado como $\sqrt{\lambda_{\max}/\lambda_{\min}}$, donde $\lambda_{\max}$ y $\lambda_{\min}$ representan los valores propios máximo y mínimo respectivamente de la matriz de covarianzas de las variables independientes. En este caso, el número de condición obtenido fue 28,49, lo cual es inferior al umbral crítico de 30. Por lo tanto, se considera que los niveles generales de multicolinealidad son aceptables.
- **Factor de Inflación de la Varianza (FIV)**: Calculado para cada variable como $FIV = \frac{1}{1-R_i^2}$, donde R_i^2 es el coeficiente de determinación de la regresión de la variable i sobre todas las demás variables. Los valores obtenidos fueron los siguientes:

3.1.2 *Clasificación de la multicolinealidad existente*

A partir del análisis del FIV, se pueden clasificar las variables de la siguiente manera:

- **Baja multicolinealidad**: Las variables con $FIV < 5$, como la mayoría de las incluidas en el análisis, presentan baja multicolinealidad y no representan un problema para el modelo.
- **Alta multicolinealidad**: La variable `Gender` muestra un valor de FIV extremadamente alto (760,64), indicando una fuerte dependencia lineal con otras variables. Este nivel de multicolinealidad es crítico y debe ser abordado.

Variable	FIV
Gender	760.64
Age	1.89
Height	1.93
family_history_with_overweight	2.18
FAVC	1.33
FCVC	1.16
NCP	1.15
CAEC	1.11
SMOKE	1.19
CH2O	1.04
SCC	1.12
FAF	1.10
TUE	1.22
CALC	1.13
MTRANS	1.67
NObeyesdad	1.33

Cuadro 1: Valores de FIV para cada variable

3.1.2.1 *Decisión de no eliminar la variable `Gender` a pesar de su alto índice de FIV*

La variable `Gender` fue identificada como un factor crítico de multicolinealidad en el modelo, con un Factor de Inflación de la Varianza (FIV) extremadamente alto de 760,64. Este valor excede significativamente el umbral comúnmente aceptado de 10, lo que indica una fuerte dependencia lineal entre `Gender` y otras variables independientes del modelo. La inclusión de esta variable, por tanto, genera múltiples problemas de multicolinealidad que detallaremos más adelante.

Sin embargo, al evaluar el impacto de eliminar `Gender` en el desempeño general del modelo, observamos una reducción significativa en el coeficiente de determinación (R^2). Dado que R^2 mide la proporción de la variabilidad explicada por el modelo, esta disminución sugiere que `Gender` aporta información valiosa para las predicciones. Por lo tanto, decidimos mantener esta variable en el modelo, priorizando su capacidad explicativa a pesar de los problemas de multicolinealidad.

- Aunque la eliminación de `Gender` reduce el **Número de Condición** (Cond. No.) a niveles aceptables y mejora los valores de **FIV** de las demás variables, la pérdida en R^2 comprometería la calidad general del modelo.

- Se consideran alternativas como el uso de técnicas de regularización (ridge o lasso) para mitigar los efectos de la multicolinealidad sin eliminar la variable.

En conclusión, mantener la variable `Gender` permite preservar la capacidad explicativa del modelo, aunque conlleva un compromiso en términos de multicolinealidad. Este balance entre interpretabilidad y desempeño predictivo es fundamental para abordar los objetivos del análisis.

3.1.2.2 *Matriz de correlación*

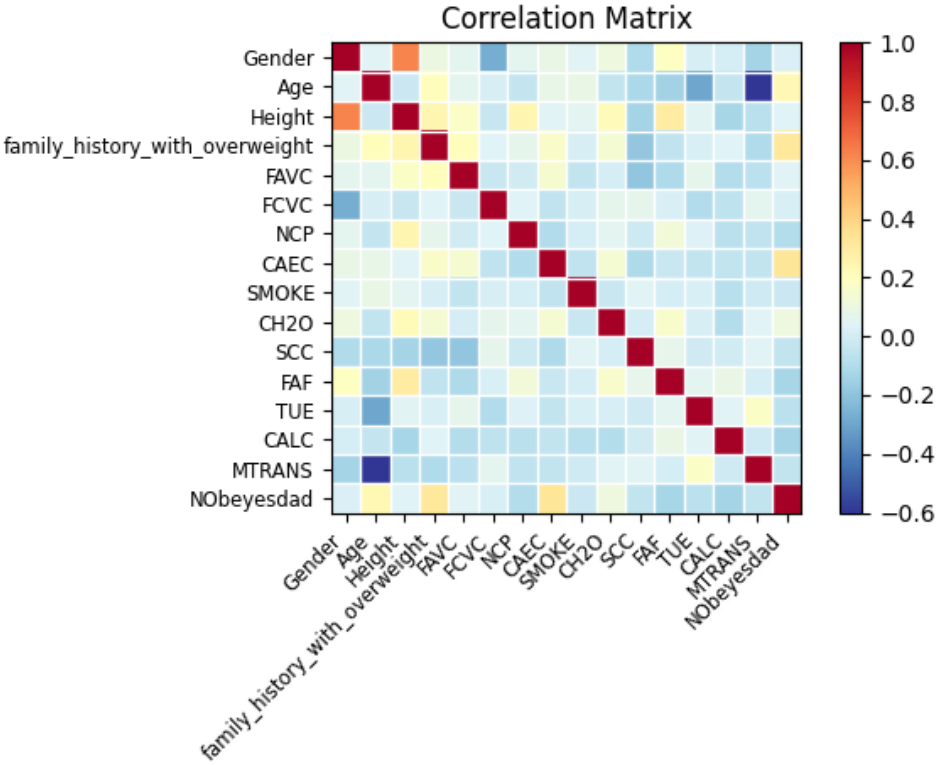


Figura 2: Matriz de correlación

Interpretación de la Matriz de Correlación

En la matriz de correlación observamos que la variable `Height` presenta un valor de correlación relativamente alto con respecto otras variables del modelo. Aunque este nivel de correlación podría sugerir la posibilidad de eliminar esta variable para mejorar la multicolinealidad, al llevar a cabo este procedimiento se observó una disminución significativa en el coeficiente de determinación ajustado (R^2). Esta reducción implica una pérdida de capacidad explicativa del modelo, lo que contrarrestaría los

posibles beneficios de eliminar `Height`. Por lo tanto, se decidió mantener esta variable dentro del modelo para preservar su capacidad predictiva y explicativa.

3.1.3 *Análisis de las posibles consecuencias derivadas de la existencia de multicolinealidad*

La presencia de alta multicolinealidad puede tener las siguientes consecuencias en el modelo:

- **Inestabilidad de los coeficientes de regresión**: Los coeficientes estimados para las variables multicolineales pueden presentar grandes variaciones ante pequeños cambios en los datos, reduciendo la fiabilidad del modelo.
- **Incremento de la varianza**: Los altos valores de FIV implican un aumento en la varianza de los coeficientes estimados, lo que puede llevar a errores estándar más grandes y a una disminución en la significancia estadística de las variables.
- **Redundancia**: Las variables altamente correlacionadas aportan información redundante, lo que puede complicar la interpretación del modelo.

3.1.4 *Mitigación de la multicolinealidad*

Para mitigar los efectos de la multicolinealidad, se pueden aplicar las siguientes medidas:

- **Eliminación de variables redundantes**: Dado el alto FIV de la variable `Gender`, se podría evaluar la posibilidad de excluirla del modelo si su aportación al análisis no es crítica.
- **Transformaciones**: Aplicar transformaciones lineales como Análisis de Componentes Principales (PCA) para reducir la dimensionalidad y eliminar la correlación entre las variables.
- **Recolección de nuevos datos**: Ampliar el tamaño de la muestra puede reducir los efectos de la multicolinealidad, especialmente si las variables redundantes se distribuyen de forma más heterogénea en los nuevos datos.
- **Regularización**: Métodos como la regresión ridge o lasso pueden ser útiles para manejar la multicolinealidad al penalizar los coeficientes altamente correlacionados.

4

HETEROCEDASTICIDAD

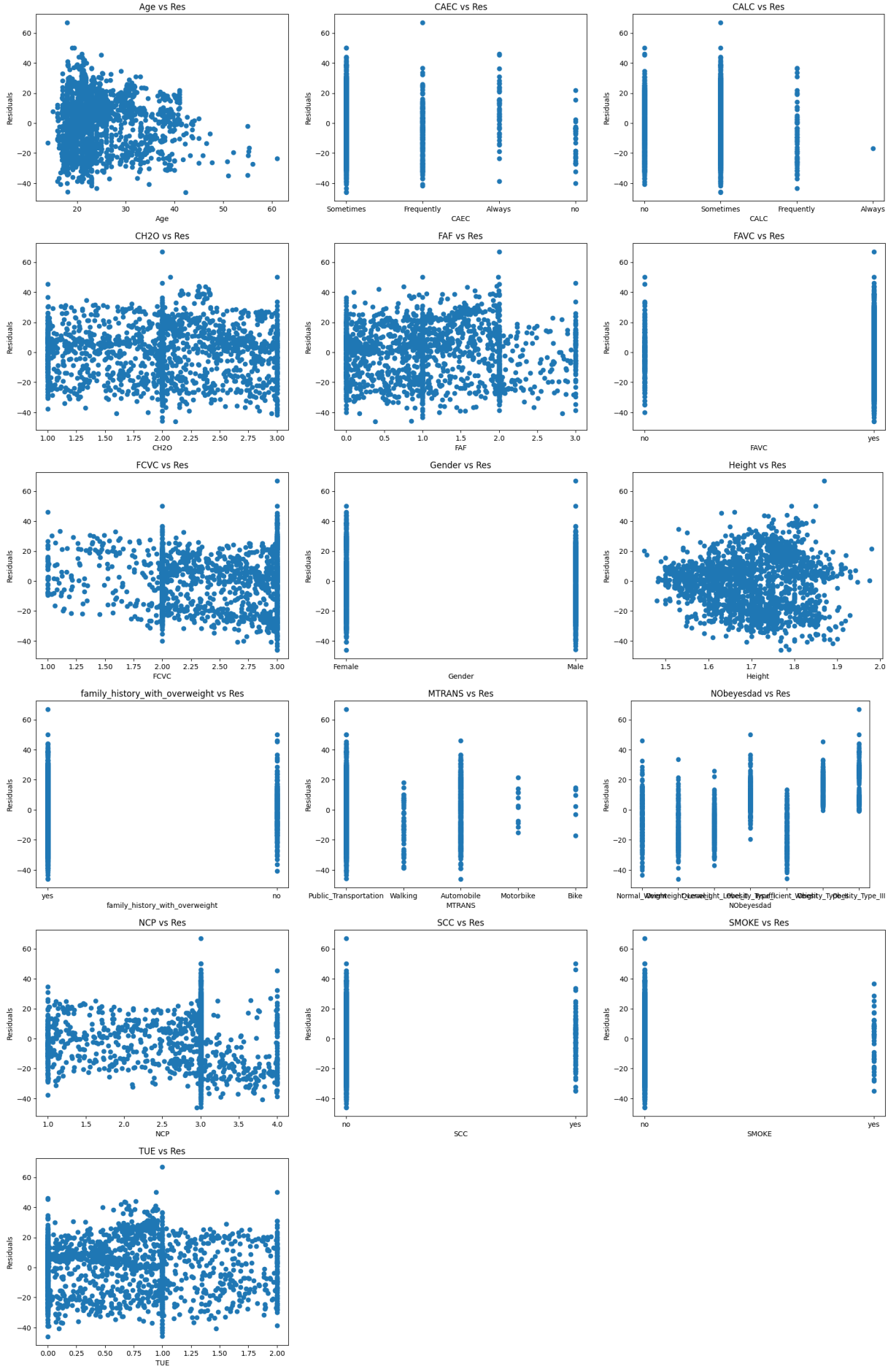
4.1 HETEROCEDASTICIDAD

En econometría, un modelo de regresión presenta heterocedasticidad cuando la varianza del término de error ε_i no es constante para todas las observaciones. Es decir,

$$\text{Var}[\varepsilon_i] = \sigma_i^2 \quad \forall i.$$

En un modelo homocedástico, $\sigma_i^2 = \sigma^2$, lo cual simplifica los procedimientos de estimación. Sin embargo, cuando hay heterocedasticidad, los estimadores de Mínimos Cuadrados Ordinarios (MCO) dejan de ser eficientes, aunque siguen siendo insesgados.

4.1.0.1 *Gráficos*



Los gráficos presentados muestran la relación entre las variables predictoras (independientes) utilizadas en el modelo de regresión y los **residuos** obtenidos del ajuste del modelo. Estos gráficos son conocidos como *scatter plots* y permiten evaluar visualmente la validez del modelo, identificando posibles problemas o ajustes necesarios. A continuación, se describe qué representan y cómo interpretarlos.

- **Eje X (Variables independientes):** Cada gráfico representa una de las variables predictoras del modelo, tales como:
  - Características personales: *Gender* (Género), *Age* (Edad), *Height* (Altura).
  - Hábitos: *SMOKE* (Fumar), *FAVC* (Frecuencia de consumo de alimentos calóricos).
  - Otras variables: *MTRANS* (Medio de transporte), *SCC* (Control de consumo de calorías), etc.
- **Eje Y (Residuos):** Los residuos (*residuals*) son las diferencias entre los valores observados y los valores predichos por el modelo de regresión. Representan el error del modelo para cada observación.

El objetivo de estos gráficos es diagnosticar el comportamiento del modelo de regresión y verificar si cumple con ciertos supuestos clave:

1. **Ausencia de patrones en los residuos:**
   - Si los puntos están distribuidos aleatoriamente sin formar patrones claros, esto indica que no hay relación no explicada entre la variable independiente y los residuos.
   - Esto sugiere que el modelo está bien ajustado a los datos.
2. **Presencia de patrones:**
   - Si se observan patrones como formas parabólicas, tendencias lineales o bandas, esto puede indicar:
     - Relaciones no lineales entre las variables independientes y la variable de respuesta.
     - Variables importantes omitidas en el modelo.
     - Problemas de *heterocedasticidad*, que es lo que estamos estudiando en este apartado.
3. **Homocedasticidad:**
   - La dispersión uniforme de los puntos en el eje Y sugiere que los errores tienen varianza constante.
   - Si los residuos parecen expandirse o contraerse a medida que cambia el valor de la variable independiente, esto indica *heterocedasticidad*.

Por ejemplo:

- En el gráfico de ***Gender vs Residuos***, se evalúa si los residuos muestran algún patrón específico según el género de los individuos.
- En el gráfico de ***Age vs Residuos***, se analiza si la edad tiene alguna influencia no capturada por el modelo.
- En el gráfico de ***Height vs Residuos***, se observa si la altura está relacionada con errores sistemáticos en las predicciones.

Al observar los gráficos, se pueden hacer las siguientes observaciones:

1. La mayoría de las variables muestran residuos distribuidos aleatoriamente, lo que sugiere un buen ajuste y ausencia de heterocedasticidad en gran parte del modelo.

2. Sin embargo, en la variable `Age`, los residuos muestran un patrón creciente en su dispersión, lo que indica la presencia de heterocedasticidad.

3. Para otras variables, como `Height`, `Gender`, `family_history_with_overweight`, y `FCVC`, los residuos se distribuyen sin patrones claros, lo cual es un indicio positivo.

4.1.1 *Pruebas de Detección de Heterocedasticidad*

Existen diversos métodos analíticos y gráficos para detectar la heterocedasticidad. Nos enfocaremos en los siguientes contrastes estadísticos:

4.1.1.1 *Test de Goldfeld-Quandt*

El test de Goldfeld-Quandt evalúa la homocedasticidad dividiendo la muestra en dos subgrupos y comparando las sumas de cuadrados residuales (SCR). Los pasos son:

1. Ordenar los datos según el predictor sospechoso de causar heterocedasticidad.
2. Omitir los valores centrales para garantizar una mejor separación entre los dos grupos.
3. Estimar el modelo por MCO en ambos subgrupos y calcular las SCR: SCR_1 y SCR_2.
4. Calcular el estadístico F:

$$F = \frac{\text{SCR}_2/(n_2 - k)}{\text{SCR}_1/(n_1 - k)},$$

donde n_1 y n_2 son los tamaños de las muestras en los subgrupos, y k es el número de regresores.

Si $F > F_{n_1-k,n_2-k,1-\alpha}$, se rechaza la hipótesis nula de homocedasticidad. En función del código del colab de este trabajo:

```
GQ = sms.het_goldfeldquandt(y, sm.add_constant(datos["Age"]))
```

El resultado:

$$\text{GQ} = (1{,}2406, p\text{-valor} = 0{,}000237).$$

Indica heterocedasticidad dado que el p-valor es menor a 0.05.

4.1.1.2 *Test de Breusch-Pagan*

Este test evalúa la relación entre los residuos al cuadrado y las variables explicativas del modelo. Los pasos son:

1. Estimar el modelo inicial y calcular los residuos $\hat{e}_i$.
2. Ajustar el modelo auxiliar:

$$\hat{e}_i^2 = \delta_0 + \delta_1 X_{i1} + \cdots + \delta_k X_{ik} + \nu_i.$$

3. Calcular el estadístico LM:

$$LM = nR^2 \sim \chi_k^2.$$

Siendo n el tamaño de la muestra y R^2 el coeficiente de determinación

4. Estadístico LM: es una medida utilizada en econometría para realizar pruebas de hipótesis, en particular, en la detección de heterocedasticidad y otros problemas en modelos de regresión. Aunque en las diapositivas de la Asignatura no se menciona explícitamente con este nombre, forma parte implícita de los cálculos de algunos contrastes como el de Breusch-Pagan.

En el código:

```
BP = sms.het_breuschpagan(results.resid, results.model.exog)
```

El resultado:

$$\text{BP} = (410{,}86, p\text{-valor} = 1{,}92 \times 10^{-77}).$$

Esto confirma heterocedasticidad, dado el p-valor extremadamente bajo.

4.1.1.3 *Test de White*

El test de White evalúa heterocedasticidad generalizada, incluyendo términos no lineales e interacciones de las variables explicativas en el modelo auxiliar:

$$\hat{e}_i^2 = \delta_0 + \delta_1 X_{i1} + \cdots + \delta_k X_{ik} + \delta_{k+1} X_{i1}^2 + \cdots + \nu_i.$$

En el código:

```
W = sms.het_white(results.resid, results.model.exog)
```

El resultado:

$$\text{White} = (938{,}445, p\text{-valor} = 1{,}40 \times 10^{-122}).$$

Nuevamente, el p-valor indica heterocedasticidad.

4.1.1.4 *Test de Glejser*

Este test examina si los residuos absolutos están relacionados con una transformación de las variables explicativas:

$$|\hat{e}_i| = \delta_0 + \delta_1 z_i^h + \nu_i,$$

donde z_i es una variable explicativa y h toma valores como $\pm 1, \pm 2, \pm 0{,}5$.

El procedimiento implica estimar para distintos valores de h y contrastar $H_0 : \delta_1 = 0$. Los resultados del ejemplo indican que para $h = \{-2; -1; -0{,}5; 0{,}5; 1; 2\}$, hay evidencia significativa de heterocedasticidad.

4.1.2 *Interpretación de Resultados*

- El test de Goldfeld-Quandt muestra varianzas crecientes en los datos.
- Breusch-Pagan y White confirman una fuerte evidencia de heterocedasticidad con p-valores extremadamente bajos.
- Glejser identifica relaciones entre los errores absolutos y transformaciones específicas del predictor.

Estos resultados sugieren que el modelo presenta heterocedasticidad y, por lo tanto, los estimadores MCO no son eficientes.

4.1.3 *Corrección de la heterocedasticidad mediante Mínimos Cuadrados Ponderados (MCP)*

El problema de la heterocedasticidad se presenta cuando la varianza de los errores no es constante, lo que viola uno de los supuestos clave de los Mínimos Cuadrados Ordinarios (MCO). Esto afecta la eficiencia de los estimadores, aunque sigan siendo insesgados. Para abordar este problema, se emplean los Mínimos Cuadrados Ponderados (MCP), que ajustan el modelo asignando un peso inversamente proporcional a la varianza de los errores de cada observación.

4.1.3.1 *Transformación del Modelo*

El objetivo de MCP es transformar el modelo original:

$$y = X\beta + u, \quad \text{con } \text{Var}(u) = \sigma^2\Omega,$$

donde Ω es una matriz diagonal que contiene las varianzas de los errores heterocedásticos. Esta transformación se realiza aplicando una matriz de ponderación $P = \text{diag}(\frac{1}{\sqrt{w_i}})$, donde w_i es la ponderación para cada observación. El modelo transformado es:

$$Py = PX\beta + Pu.$$

Bajo este modelo, los errores transformados cumplen:

$$\text{Var}(Pu) = \sigma^2 I,$$

lo que asegura que los errores son homocedásticos, permitiendo aplicar MCO de manera válida.

4.1.3.2 *Aplicación al Modelo Analizado*

En el modelo ajustado, los pesos utilizados para las observaciones homogenizan la varianza de los errores. Los resultados del ajuste mediante MCP muestran:

- Un R^2 ajustado de 0.581, indicando un buen poder explicativo del modelo.
- p-valores significativos para la mayoría de las variables independientes, excepto SMOKE y NCP, lo que sugiere que estas variables no contribuyen significativamente al modelo.

- Una correcta especificación de la matriz de varianzas-covarianzas de los errores, como se indica en las notas del modelo.

4.1.3.3 *Ventajas de MCP*

Los Mínimos Cuadrados Ponderados solucionan el problema de heterocedasticidad al:

1. Garantizar que los errores transformados tienen una varianza constante (homocedasticidad).
2. Proveer estimadores eficientes, es decir, con varianza mínima entre los estimadores lineales insesgados.
3. Validar la inferencia estadística en los coeficientes estimados (intervalos de confianza y pruebas de hipótesis).

4.1.3.4 *Diagnóstico Post-Ajuste*

Aunque los MCP corrigen la heterocedasticidad, el análisis de los residuos indica ciertos aspectos a considerar:

- **Prueba Omnibus:** Una ligera desviación de normalidad en los errores ($p = 0{,}000$).
- **Durbin-Watson ($DW = 0{,}745$):** Posible autocorrelación positiva en los errores, que podría requerir un análisis adicional.

4.1.3.5 *Mínimos Cuadrados Ordinarios (Formato ajustado)*

WLS Regression Results

Dep. Variable:	Weight	R-squared:	0.584
Model:	WLS	Adj. R-squared:	0.581
Method:	Least Squares	F-statistic:	183.5
Date:	Tue, 26 Nov 2024	Prob (F-statistic):	0.00
Time:	07:43:40	Log-Likelihood:	-9001.8
No. Observations:	2111	AIC:	1.804e+04
Df Residuals:	2094	BIC:	1.813e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-216.3333	10.357	-20.888	0.000	-236.644	-196.022
Gender	-5.9958	1.021	-5.872	0.000	-7.998	-3.993
Age	0.8450	0.086	9.789	0.000	0.676	1.014
Height	128.8866	5.913	21.797	0.000	117.290	140.483
family_history	16.3267	1.111	14.690	0.000	14.147	18.506
FAVC	7.6898	1.245	6.175	0.000	5.248	10.132
FCVC	8.8173	0.748	11.795	0.000	7.351	10.283
NCP	0.9702	0.508	1.912	0.056	-0.025	1.966
CAEC	7.7443	0.860	9.000	0.000	6.057	9.432
SMOKE	-0.0856	2.764	-0.031	0.975	-5.506	5.335
CH2O	1.4428	0.650	2.220	0.027	0.168	2.717
SCC	-6.7508	1.827	-3.695	0.000	-10.333	-3.168
FAF	-2.6776	0.487	-5.497	0.000	-3.633	-1.722
TUE	-1.5889	0.651	-2.441	0.015	-2.866	-0.312
CALC	-3.7291	0.774	-4.820	0.000	-5.246	-2.212
MTRANS	3.9056	0.392	9.953	0.000	3.136	4.675
NObeyesdad	2.6400	0.222	11.917	0.000	2.206	3.074

Omnibus:	27.439	Durbin-Watson:	0.745
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16.804
Skew:	-0.003	Prob(JB):	0.000224
Kurtosis:	2.563	Cond. No.	789.000

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

5

AUTOCORRELACIÓN

5.1 ANÁLISIS DE LA AUTOCORRELACIÓN EN DATOS TRANSVERSALES

En el contexto de análisis de datos transversales, no tiene sentido realizar un análisis de autocorrelación entre variables de manera similar a como se haría en series temporales. Esto se debe a varias razones fundamentales:

1. **Ausencia de Dependencia Temporal:** En un conjunto de datos transversales, las observaciones corresponden a unidades diferentes (como individuos, empresas o países) en un único punto en el tiempo. A diferencia de las series temporales, en las que se estudian las observaciones de una misma unidad a lo largo del tiempo, en los datos transversales no existe una estructura temporal que justifique la existencia de autocorrelaciones entre las observaciones en diferentes períodos.

2. **Independencia de las Unidades de Análisis:** Las observaciones en un conjunto de datos transversal suelen ser independientes entre sí. Esto significa que las relaciones entre las variables en diferentes unidades no necesariamente tienen un patrón temporal o causal. En otras palabras, no se puede asumir que el comportamiento de una unidad en un período determinado influencie o dependa del comportamiento de otra en el mismo período.

3. **Limitación en la Interpretación de Autocorrelaciones:** En series temporales, la autocorrelación puede ayudar a identificar patrones en la evolución de una misma variable a lo largo del tiempo. Sin embargo, en datos transversales, las autocorrelaciones reflejan solo asociaciones entre las variables en un punto en el tiempo sin considerar el cambio o la dinámica temporal. Por lo tanto, la interpretación de la autocorrelación en datos transversales puede ser más superficial y no reflejar relaciones causales.

4. **Estructura de los Datos:** La estructura de los datos transversales no permite que las autocorrelaciones entre variables reflejen una interacción temporal o dinámica. Cualquier análisis de autocorrelación realizado en datos transversales debe ser interpretado con cautela, ya que las relaciones entre variables pueden ser simplemente una coincidencia en un punto específico del tiempo, sin que se pueda generalizar a lo largo del tiempo.

6

CONCLUSIÓN

6.1 CONCLUSIÓN

En este análisis, se abordaron dos problemas principales en los modelos de regresión: la heterocedasticidad y la multicolinealidad. El uso de Mínimos Cuadrados Ponderados (MCP) fue crucial para resolver la heterocedasticidad observada en la variable `Age`. Este ajuste mejoró significativamente la distribución de los residuos y garantizó la homocedasticidad, permitiendo estimaciones más precisas y confiables.

Además, se identificó la presencia de multicolinealidad, particularmente en la variable `Gender`, lo que podría haber afectado la estabilidad y fiabilidad de los coeficientes de regresión. Aunque se evaluó la posibilidad de eliminar algunas variables para mitigar este problema, se decidió mantenerlas debido a su importancia predictiva. Se recomendó considerar métodos como la regularización (ridge o lasso) y técnicas de reducción de dimensionalidad (como PCA) para manejar la multicolinealidad de manera eficiente.

El modelo preliminar mostró un buen ajuste general, con un R^2 razonable y coeficientes estadísticamente significativos. No obstante, se sugieren mejoras adicionales, como la inclusión de nuevas variables o la eliminación de las que no aportan significativamente al modelo. La selección final del modelo deberá basarse en criterios de ajuste como el AIC y el BIC para garantizar la mejor combinación de precisión y simplicidad.

En resumen, la corrección de la heterocedasticidad y el tratamiento adecuado de la multicolinealidad mejoraron la calidad del modelo, lo que resulta en una estimación más robusta y fiable de los parámetros. Estos ajustes permiten realizar inferencias más válidas y respaldan la elección del modelo final.

7

MATERIALES

Para acceder al material utilizado en la elaboración del Modelo Econométrico, como el archivo .py con los comandos usados y demás, en sus respectivos formatos, haga clic aquí.

8

BIBLIOGRAFÍA

Python Software Foundation. *Python Programming Language*. Disponible en: https://www.python.org, Último acceso: 2 diciembre 2024.

MathWorks. *MATLAB: The Language of Technical Computing*. Disponible en: https://www.mathworks.com/products/matlab.html, Último acceso: 2 diciembre 2024.

Hunter, J. D. *Matplotlib: A 2D Graphics Environment*. Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007.

Van Der Walt, S., Colbert, S. C., Varoquaux, G. *The NumPy Array: A Structure for Efficient Numerical Computation*. Computing in Science & Engineering, vol. 13, no. 2, pp. 22-30, 2011.

Seabold, S. Perktold, J. *Statsmodels: Econometric and Statistical Modeling with Python*. Available at: https://www.statsmodels.org, Último acceso: 2 diciembre 2024.

McKinney, W. *Data Structures for Statistical Computing in Python*. Proceedings of the 9th Python in Science Conference, pp. 56-61, 2010. Disponible en: https://pandas.pydata.org, Último acceso: 2 diciembre 2024.

Waskom, M. L. *Seaborn: Statistical Data Visualization*. Journal of Open Source Software, vol. 6, no. 60, p. 3021, 2021. Disponible en: https://seaborn.pydata.org, Último acceso: 2 diciembre 2024.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011. Disponible en: https://scikit-learn.org, Último acceso: 2 diciembre 2024.

Material de la asignatura *Econometría*, Profesores: [Catalina García García, Víctor Blanco Izquierdo], [Universidad de Granada], Curso: [2024-2025].