Análisis de Multicolinealidad

José Ángel Carretero Montes
Ismael Sallami Moreno
Fernando José Gracia Choin

Noviembre 2024

1 Análisis de la Multicolinealidad

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:

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

Table 1: Valores de FIV para cada variable

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.

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.

1.2.2 Matriz de correlación

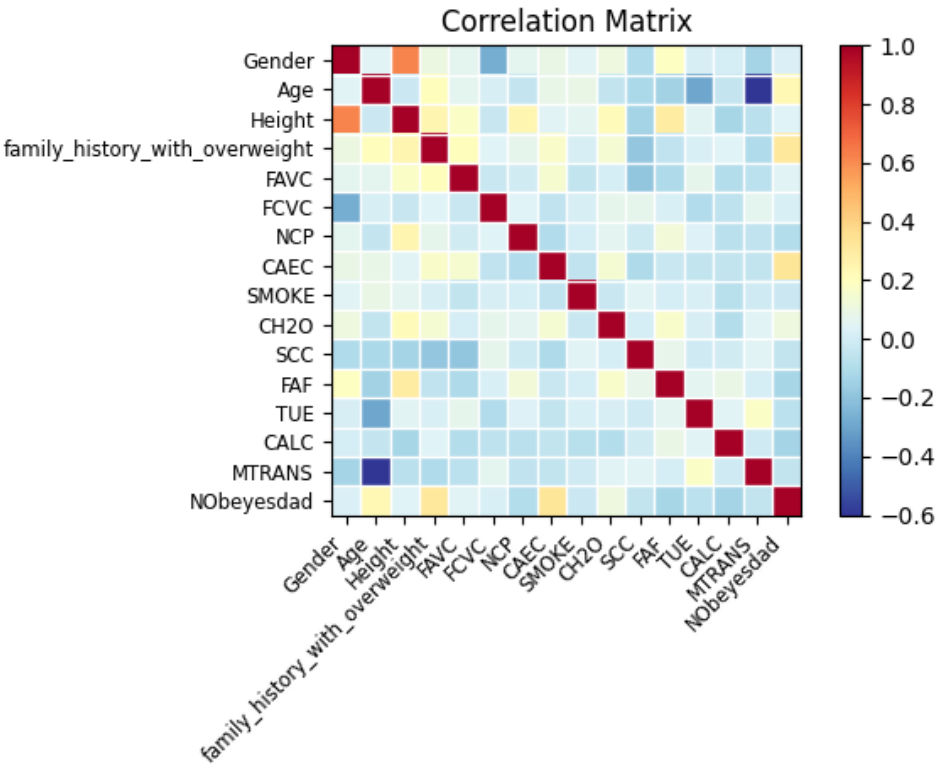

Figure 1: Matriz de correlación

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.

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.

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.