Comunicación Breve

MULTICOLINEALIDAD ANALIZADA EN EL CURSO DE ANALISIS CUANTITATIVO DE LA MAESTRIA EN CIENCIA DE DATOS

Cristian Bolívar, Kenny Rodríguez, Juan David Borja, Fabian Salazar Figueroa

Palabras Clave: Multicolinealidad, correlación, variables explicativas, factor de inflación de la varianza

RESUMEN

"La multicolinealidad se refiere a la relación lineal entre dos o más variables. Es un problema de datos que puede causar serias dificultades con la fiabilidad de las estimaciones de los parámetros del modelo. En este artículo se considera la multicolinealidad entre las variables explicativas en el modelo de regresión lineal múltiple. Se presentan sus efectos sobre el modelo de regresión lineal y algunos diagnósticos de multicolinealidad para este modelo." (1)

Introducción: La multicolinealidad es un fenómeno que ocurre cuando dos o más variables explicativas en un modelo de regresión están fuertemente correlacionadas. Este fenómeno puede afectar negativamente la precisión y la interpretación de los resultados de la regresión. lo que también significa falta de ortogonalidad entre ellos. Esta relación también es llamada colinealidad o mal condicionamiento

Definición y Tipos: La multicolinealidad se presenta cuando la relación entre dos o más variables del modelo es muy fuerte. Existen dos tipos de multicolinealidad

1. Multicolinealidad exacta: Se produce si ***k*** vectores se encuentran en un subespacio de dimensión menor que ***k***. Esta es la definición de multicolinealidad exacta o dependencia lineal exacta. Cuando una o más variables son combinación lineal de otras variables. En este caso, el coeficiente de correlación entre las variables multicolineales es igual a 1.
2. Multicolinealidad aproximada: Basta con que ***k*** variables sean casi dependientes, lo que ocurre si el ángulo entre una variable y sus proyecciones ortogonales sobre otras es pequeño. Dicho de otra manera: No existe la combinación lineal entre las variables, pero el coeficiente de determinación entre dos o más variables es muy cercano a 1 y, por lo tanto, están fuertemente correlacionadas.

En muchos estudios, la multicolinealidad se ha confundido con la correlación. La correlación es la relación lineal entre solo dos variables, mientras que la multicolinealidad puede existir entre dos variables o entre una variable y la combinación lineal de las otras. Por lo tanto, la correlación es un caso especial de multicolinealidad. Una alta correlación implica multicolinealidad, pero lo contrario no es cierto.

Se puede tener multicolinealidad entre variables explicativas, pero aun así no tener una alta correlación entre pares de estas variables.

La multicolinealidad crea dificultades cuando se construye un modelo de regresión entre la variable de respuesta **Y** contra la variable explicativa **X**, y está definida por la ecuación:

$$\mathbf{y = Xb + E}$$

Donde:

Y, es el vector de dimensión [n x 1]

X, Es la matriz de variables explicativas de dimensión [n x (k + 1)]

b, Es el vector de dimensión [(k + 1) x n]; para los coeficientes de regresión

E, Es el vector de errores de dimensión [n x 1] que se supone que tiene media **0** y varianza matriz de covarianza $\boldsymbol{\sigma 2 I}$.

1. Multicollinearity Aylin Alin∗ Article 2010 John Wiley & Sons, Inc. WIREs Comp Stat 2010 2 370–374

Consecuencias: La multicolinealidad puede tener varias consecuencias negativas en un modelo de regresión:

- Los coeficientes de regresión pueden cambiar cuando se añaden variables que están correlacionadas.
- Se reduce la precisión de la estimación de los parámetros, aumentando el error estándar de los coeficientes de regresión.
- Los p-valores de los coeficientes de regresión se vuelven menos confiables.
- Es probable que se caiga en una situación de sobreajuste, es decir, que el modelo esté demasiado ajustado y, por este motivo, no sirva para hacer predicciones.
- Es probable que se caiga en una situación de sobreajuste, es decir, que el modelo esté demasiado ajustado y, por este motivo, no sirva para hacer predicciones.

Detección y Solución: La multicolinealidad se puede detectar calculando los coeficientes de correlación para todos los pares de variables predictoras. Para solucionar este problema, se pueden utilizar diversas técnicas, como la eliminación de variables redundantes, la recopilación de más datos, o la utilización de métodos de regularización.

Una manera de identificar la multicolinealidad es calcular la matriz de correlación, ya que en ella se recoge el coeficiente de correlación entre todas las variables y, por tanto, se puede observar si algún par de variables están fuertemente correlacionadas. No obstante, con la matriz de correlación solo se puede saber si dos variables están relacionadas entre sí, pero no se puede averiguar si existe una combinación entre un conjunto de variables.

Para ello, se suele calcular el factor de inflación de la varianza.

$$\boldsymbol{VIF_i} = \frac{1}{1-\boldsymbol{R_i^2}}\text{; para } i = 1,2,3\ldots\boldsymbol{k}$$

Donde: **VIF** es el factor de inflación de la varianza de la variable $\boldsymbol{i}$ y $\boldsymbol{R_i^2}$ es el coeficiente de determinación del modelo de regresión que tiene la variable $\boldsymbol{i}$ como variable dependiente y el resto de las variables como variables independientes.

sí pues, según el valor de los factores de inflación de la varianza obtenidos se puede saber si hay multicolinealidad o no:

- **FIV = 1**: cuando el factor de inflación de la varianza es igual a 1, significa que no existe ninguna correlación entre la variable dependiente y las otras variables.
- **1 < FIV < 5**: existe correlación entre las variables, pero es moderada. En principio, no es necesario aplicar ninguna acción para corregir la multicolinealidad.
- **FIV > 5**: si algún factor de inflación de la varianza es mayor que 1, significa que la multicolinealidad del modelo es alta y, por tanto, se debería intentar solucionar.

Si el tamaño de la muestra es pequeño, aumentar el número de datos puede reducir la multicolinealidad aproximada.

Quitar alguna de las variables que producen la multicolinealidad. Si las variables están fuertemente correlacionadas, se perderá poca información en el modelo y la multicolinealidad se verá reducida.

Realizar el modelo de regresión aplicando el criterio de mínimos cuadrados parciales (PLS).

En ocasiones, se puede dejar el modelo de regresión tal y como está, con la multicolinealidad. Por ejemplo, si solo queremos hacer un modelo para hacer predicciones y no necesitamos interpretarlo, podemos utilizar la ecuación del modelo para predecir el valor de la variable dependiente con una nueva observación, suponiendo que el patrón de multicolinealidad se repite en las nuevas observaciones.

Conclusión:

La multicolinealidad es un problema común en la regresión que puede afectar la interpretación y precisión de los resultados. Por lo tanto, es crucial entender, detectar y manejar adecuadamente la multicolinealidad en el análisis de regresión.

1. Multicollinearity Aylin Alin∗ Article 2010 John Wiley & Sons, Inc. WIREs Comp Stat 2010 2 370–374