# Multicolinealidad

*(Guía de clase)*

## 1. Significado y naturaleza de la multicolinealidad

✎ El supuesto de ausencia de multicolinealidad es el segundo relativo a la parte sistemática del modelo de regresión. Aunque suele ser definido generalmente como un "problema de datos" (Véase Gujarati 2004: 349; Chaterjee et al 2000: 226), aquí se extenderá la idea para incluir situaciones en las que claramente el problema surge además por: a) la naturaleza teórica del problema; b) como resultado de "la propiedad de intercambiabilidad de los índices"; c) incorrecta especificación del modelo de regresión.

✎ En un sentido amplio, el supuesto de la regresión lineal establece que ninguno de los regresores incorporados al modelo puede ser el resultado de la combinación lineal de los restantes regresores.

FICHA Nº 5

- Obsérvese con detalle que el supuesto refiere a la ausencia de una combinación lineal de los regresores independientes y no meramente a una relación entre dos regresores (colinealidad).
- La referencia a una "combinación lineal de varios regresores" lleva directamente a las medidas de correlación múltiple ("R") y de coeficiente de determinación ("$R^{2}$").
- La diferencia conceptual entre "multicolinealidad" y "colinalidad" es central para comprender a cabalidad este supuesto.
- Es decir, el supuesto no se puede representar adecuadamente con la técnica de las correlaciones bivariadas o de Pearson.
- Esto repercutirá a la hora de elegir técnicas para identificar violaciones al supuesto, así como también al momento de resolverlos.

# Multicolinealidad

*(Guía de clase)*

---

✎ La multicolinealidad puede afectar a dos regresores (en el caso más simple de todos), a un subconjunto de los regresores incluidos, o incluso a todos los regresores del modelo.

✎ Si bien el supuesto aplica a todos los análisis, la severidad con que puede afectar las estimaciones e inferencias que se realicen dependerá del tamaño de la base de datos y del tipo de problemas analizados.

- Es más preocupante el problema cuando se tienen pequeñas bases, por ejemplo con menos de 30 observaciones que cuando se tienen grandes bases de más de mil observaciones. El punto crucial aquí son los grados de libertad disponibles para realizar las pruebas de hipótesis en los coeficientes. Este aspecto se relaciona con el supuesto referido al rango de la matriz y a la noción de "micronumerosidad" introducida por Goldberger .

- En las series temporales, el problema de multicolinealidad puede aparecer conjuntamente con un problema de auto-correlación de los errores.

## 2. Consecuencias de la multicolinealidad

✎ En términos sintéticos, en presencia de algún grado severo de multicolinealidad en un el modelo de regresión, los estimadores MCO de los parámetros seguirán siendo consistentes y no sesgados pero ya no serán los de menor varianza (v.g. serán ineficientes).

- Se desprende de la anterior afirmación que será necesario examinar la varianza o el error estándar de los coeficientes de regresión.

- Para desarrollar esta idea, examinaré primero un caso extremo de

---

multicolinelidad perfecta y para luego tratar el caso más general y frecuente.

- Debe recordarse en todo momento que las consecuencias de la multicolinealidad están directamente relacionadas con la magnitud con que se presenta. De aquí la continua apelación al calificativo de "severidad".

✎ Sea el siguiente modelo de regresión lineal múltiple que se ha ajustado con cuatro regresores:

$$Y=b_0+b_1X_1+b_2X_2+b_3X_3+b_4X_4+e$$

✎ Supóngase formalmente que el regresor 4 incluido en el modelo puede expresarse como una combinación lineal perfecta de otros tres regresores también incluidos:

$$X_4=a_0+a_1X_1+a_2X_2+a_3X_3$$

$$R_{X_4}{}^2=1.0$$

☞ Dado que señaló que es una combinación lineal perfecta, el coeficiente de determinación para esta regresión auxiliar es igual a 1.

☞ Una combinación de este tipo ocurre cuando por ejemplo, una variable nominal con tres categorías ha sido convertida en tres variables dicotómicas y las tres se han incluido en el modelo de regresión.

✎ Supongamos que el paquete estadístico ha estimado el modelo de regresión (una cuestión que difícilmente ocurre y que puede prácticamente verificarse). Recordando la fórmula del error estándar del coeficiente de regresión, en el caso del regresor 4 será:

# Multicolinealidad

*(Guía de clase)*

$$ee_{b_4} = \sqrt{\frac{\hat{\sigma}^2}{S_{X_4}{}^2 * (1 - R_{X_4}{}^2)}}$$

$$ee_{b_4} = \sqrt{\frac{\hat{\sigma}^2}{S_{X_4}{}^2 * (1 - 1)}}$$

$$ee_{b_4} = \sqrt{\frac{\hat{\sigma}^2}{S_{X_4}{}^2 * 0}}$$

- Al re-emplazar en el denominador de la fórmula el valor que tiene el coeficiente de determinación para la ecuación auxiliar ajustada para $X_4$ se llega una división entre 0, de cociente indeterminado.

- En consecuencia, no es posible hallar el error estándar para este coeficiente. Se cae la estimación.

- Pero, tampoco es posible computar los valores del estadístico t para realizar la prueba de hipótesis porque como se recordará se requiere del error estándar del coeficiente como denominador.

$$t_{b_4} = \frac{b_4 - 0}{ee_{b_4}}$$

$$t_{b_4} = \frac{b_4 - 0}{?}$$

✎ Fuera de este caso extremo de multicolinealidad perfecta en que el denominador se vuelve cero, la estimación de los errores estándares de los

coeficientes estará afectado en diversos grados de severidad.

- Claramente, dicha severidad vendrá determinada por el coeficiente de determinación para la regresión auxiliar del "p-ésimo regresor".

- Este coeficiente puede ser interpretado como la proporción de varianza en $X_p$ que no está correlacionado con los restantes del modelo. Al ser restado de 1 tal como exige la fórmula, puede interpretarse ese resultado como una proporción de varianza libre.

- Al reducirse progresivamente esta varianza libre que es el multiplicador de la varianza de $X_p$ tenderá:

  - i) hacer que el denominador del error estándar sea cada vez más grande;

  - ii) a su vez el cociente será grande, dando lugar a una magnitud de la varianza del coeficiente que será cada vez mayor conforme sea más grave la multicolinealidad presente en el modelo;

  - iii) la raíz, esto es el error estándar será también mayor;

  - iv) dando lugar (por fórmula) a un valor reducido de *t*.

  - v) aumentando las posibilidades de que no pueda rechazarse la hipótesis nula.

## 3. Detección de la multicolinealidad

✎ Para detectar una violación a este supuesto se disponen de distintos instrumentos, desde aquellos más intuitivos o informales hasta un conjunto de pruebas formales o estadísticos.

- Se recomienda fuertemente que las decisiones finales respecto de este

problema se tomen únicamente con el uso de pruebas formales.

✎ Desde la etapa de revisión de la bibliografía se pueden anticipar problemas de colinealidad entre las variables por la naturaleza del problema que se está trabajando.

- En la sociología económica, las variables que representan atributos del ciclo económico como ser: inflación, tasa de crecimiento del producto bruto interno, tasa de desempleo, incidencia de la pobreza. El efecto del ciclo económico se verá reflejado en cada una de estas variables.

- Algunas variables pueden ser tratadas como "variables resumen" de un conjunto de otras propiedades por efecto de la historia. Si se realiza un análisis a nivel de municipios en México y se utilizan como independientes el porcentaje de población indígena, el porcentaje de personas que no completaron primaria, el grado de marginación socioeconómica, la tasa de abandono escolar en primaria y la tasa de mortalidad infantil, se encontrará que estas variables está correlacionadas por razones históricas de marginación y exclusión de la población indígena.

✎ En la etapa preliminar del análisis, donde se construyen gráficos y se calculan estadísticos descriptivos para las variables:

- Los gráficos entre variables independientes permitirán ver la dispersión de los puntos y proporcionar así un primer indicio de que dos variables están correlacionadas.

- Conviene presentar y analizar una matriz de correlaciones entre los predictores. Esta matriz permitirá detectar el caso más simple de violación del supuesto: correlaciones bivariadas fuertes. Sin embargo, este examen con correlaciones no es suficiente, porque no permitirá detectar combinaciones lineales entre varios regresores.

✎ El ajuste del modelo de regresión puede proporcionar información valiosa para

detectar multicolinealidad en tres casos típicos, comenzando del más extremo de todos.

- La prueba de significación F para todos los coeficientes informa que "existe modelo", pero ninguna de las pruebas t aplicadas a los coeficientes resulta estadísticamente significativa. Tal falta de consistencia es un indicio de que existe multicolinealidad severa entre los predictores.

- El modelo es significativo, algunos o varios coeficientes son significativos pero el signo de uno de los coeficientes resulta contrario al hipotetizado, en una forma inesperada y paradójica.

  - Por ejemplo, los estudios sobre ingreso informan consistentemente que la relación entre edad e ingresos es parabólica, con un signo positivo para el término de primer grado y negativo para el término de segundo grado. Sin embargo, se ha hecho en el ajuste hecho los signos están invertidos.

- El modelo es inestable.

  - La incorporación o eliminación de una variable genera fuertes y grandes cambios en la magnitud de los coeficientes.
  - La incorporación de nuevas observaciones al análisis, o la eliminación de algunas observaciones, genera fuertes cambios en las magnitudes de los coeficientes.

✎ La primera medida estadística formal para detectar la multicolinealidad disponible en el SPSS ® es el **"Tolerance"** o **Tolerancia**, representada por "T".

- Supone que se ajusta una ecuación auxiliar para cada regresor del modelo de regresión ajustado y se calcula el coeficiente de determinación respectivo.

- La TOLERANCIA, "*T*" , mide la proporción de lo que se denominó más arriba como "varianza libre" para cada uno de los regresores del modelo.

- Habrá por tanto, tantas tolerancias como regresores se han incluido en un

modelo de regresión. Por tanto, cada regresor tendrá distinta tolerancia y habrá que prestar atención a aquella medida más baja.

- Formalmente: $$T_p = 1 - R_p^2$$

- De la fórmula anterior se desprende que la tolerancia tiene un valor máximo de 1 cuando el regresor en cuestión no tiene ningún grado de multicolinealidad con los restantes, hasta un valor mínimo de 0 cuando el regresor p es una combinación lineal perfecta de los otros regresores.

- Es deseable que tolerancia sea lo mayor posible, idealmente igual a uno, y en general que sea superior a 0,40.

- En el paquete SPSS ® se solicita esta medida: "analyze/regression/linear/" y en la opción de "statistics" "colliearity diagnostics". Aparecerá como una columna adicional en la tabla de los coeficientes de regresión, luego de la significación.

✎ La segunda medida para identificar multicolinealidad es el estadístico de **"variance inflation factor" o VIF**.

- La idea del VIF es sencilla de entender. A medida en que es mayor la multicolinealidad presente en uno de los regresores del modelo, la varianza de su coeficiente comienza a crecer porque el denominador de la fórmula se hace más chico. Es decir, la multicolinealidad "infla" la varianza del coeficiente.

- Formalmente: $$VIF_p = \frac{1}{1 - R_{X_p}^2}$$

- Se observa claramente que el VIF se define como el inverso de la tolerancia.

- El VIF tomará valores entre un mínimo de 1 cuando no hay ningún grado de multicolineadlidad y no tendrá límite superior por definición en el caso de

multicolinealidad perfecta.

✎ Una crítica que se ha dirigido a las dos medidas de multicolinealidad anteriores, la tolerancia y el VIF es que pueden no resultar concluyentes. El argumento se fundamenta en que el error estándar de un coeficiente también depende también de la varianza del p- avo regresor analizado y no solamente en el grado en que éste puede expresarse como una combinación lineal de los otros regresores.

☞ Se recuerda que el denominador del error estándar es:

$$S^2{}_{X_p} * (1 - R^2{}_{X_p})$$

☞ Este argumento recuerda en forma importante que "no hay recetas" para diagnosticar cuando un grado determinado de multicolinealidad puede ser severo.

✎ De esta crítica surge una tercera medida estadística para detectar multicolinealidad, originada en la aplicación del **análisis factorial** como técnica auxiliar al modelo de regresión (Véase una aplicación en Cortés 1987). En forma sencilla se explica a continuación alguna de las nociones elementales.

➷ La idea central detrás de un análisis factorial es que si un conjunto de variables están correlacionadas entre sí es que se debe a que están midiendo *al menos* un concepto o factor subyacente. En este sentido, el análisis factorial resume información mediante la construcción de índices.

☞ El concepto de modernización no observable directamente puede sin embargo suponerse que se representa a través de distintos indicadores, tales como urbanización, industrialización, escolarización, participación de la mujer en la fuerza de trabajo, nivel de participación electoral, crecimiento económico, etc. Puede suponerse que algunas, sino todas, estas variables pueden presentar fuertes correlaciones en un análisis

transversal de países o de regiones en un país.

☞ Se enfatiza la idea de que *al menos* un concepto pudiera subyacer, porque el análisis factorial está diseñado para "extraer" todos los factores-resumen que pueden existir en la matriz de correlaciones entre un conjunto de variables.

- El objetivo de aplicar esta técnica a los regresores es identificar cuál es la estructura factorial subyacente en el modelo original y detectar la existencia de "factores" que puedan expresarse como una fuerte combinación de dos o más regresores. Este tipo de estrategia puede ser útil luego como medida de corrección, en la medida que pueden subsituirse los regresores por el o los índices factoriales construidos.

- En el análisis factorial, cada factor subyacente que se identifica tiene asociada una varianza propia o Eigen-Value. Dado que las variables se ingresan al análisis en forma estandarizada, cuando la varianza de un factor es mayor a 1 es un indicio de que dos o más variables se combinan para generar un factor.

- De los "valores propios" o Eigen values se proponen como medida el índice de condición o "condition index". Se define como la raíz cuadrada de la razón entre la Eigen value máximo obtenido en el análisis y el mínimo valor de Eigen value obtenido. Formalmente (Gujarati 2004: 348):

$$IC=\sqrt{\frac{Eigenval_{MAX}}{Eigenval_{MIN}}}$$

- En el paquete SPSS ® proporciona como parte de los diagnósticos de multicolinealidad un "análisis factorial ad-hoc" para los regresores incluidos en el modelo, cuyos resultados difieren del realizado "por fuera" de la regresión, pero que cumplen con la misma finalidad.

- Las reglas propuestas por Gujarati (2004: 348) para interpretar los valores del

índice de condición son dos. Si el IC está entre 10 y 30, existe multicolinealidad entre moderada a fuerte. Si el IC excede de 30, existe multicolinealidad severa.

## 4. Tratamiento de la multicolinealidad

✎ Tal como se ha insistido largamente aquí el problema de multicolinealidad es de grados. Por lo que las medidas que aquí se proponen deben tomarse siempre y cuando la severidad del problema sea tan importante como para que una o varias variables del modelo de regresión se presenten como estadísticamente no significativas.

☞ En el caso de que exista multicolinealidad en un grado leve, perfectamente puede continuarse el trabajo de análisis sin adoptar medidas remediales.

✎ Si se trata de un problema de multicolinealidad casi perfecta o perfecta, es razonable pensar que ésta se debe a la incorporación de dos regresores que miden el mismo concepto pero de forma alternativa.

☞ En tal caso, se sugiere revisar el proceso de operacionalización de los conceptos y suprimir la redundancia hallada.

✎ En presencia de multicolinealidad, una medida remedial bastante lógica puede ser quitar del modelo aquella variable con más alto VIF ( o más baja tolerancia).

✎ Dado que una de las razones por las que aparece multicolinealidad se debe a los términos cuadráticos o cúbicos incluidos innecesariamente en el modelo, una medida remedial razonable consiste en quitar tales términos del modelo.

✎ La medida remedial más apropiada es substituir la o las variables

correlacionadas por un índice que resume la información provista por cada una de ellas.

- En ocasiones, las variables incluidas muy correlacionadas están indicando un concepto más abstracto que el que se supone hace independientemente cada una de las variables.

  - Por ejemplo, en un análisis hecho con países o regiones, las variables incluidas como independientes tales como alfabetización, urbanización, empleo en el sector servicios, razón de teléfonos por habitante, mortalidad infantil pueden presentar fuerte multicolinealidad porque representan conjuntamente el concepto más abstracto de "modernización".

- En tal caso, se procede a computar un índice mediante alguna técnica, desde la más simple, como ser un índice sumatorio simple hasta uno más sofisticado como ser un índice calculado sobre la base de un análisis factorial.

  - Lo más apropiado es utilizar el análisis factorial para extraer los factores subyacentes en el caso de que la multicolinealidad sea severa.

| SUPUESTO | PRUEBA | VIOLACIÓN | CONSECUENCIAS | TRATAMIENTOS |
|---|---|---|---|---|
| Ausencia de multi-colinealidad<br><br>Cov ($X_j$ , $X_{j+1}$ ) =0 | 1) el SPSS detiene los cálculos; no hay ajuste posible del modelo | 1) colinealidad perfecta entre dos predictores ($R^2_{X,X}$=1). | Múltiples hiperplanos de regresión se pueden ajustar a las observaciones. La estimación se para. | 1) revisar la especificación del modelo (en general esto resulta de un error en las variables)<br><br>Ajustar un modelo dividiendo todas las variables entre uno de los predictores (que no sea correlacionado). Cortés & Rubalcaba (1983). |
| | 2) matriz de correlaciones<br><br>3) Indicador de Tolerancia:<br><br>$T = 1 - R^2$<br><br>4) VIF<br><br>5) Análisis factorial | 2) uno de los predictores es una función perfecta de los restantes predictores<br><br>2) uno (o varios) de los predictores es una función de los restantes predictores con un $R^2$ alto (70% o más por ejemplo)<br><br>3) existen moderadas correlaciones (bivariadas) entre alguno de los regresores | No se introducen sesgos en los coeficientes, pero se afectan sus varianzas (son coeficientes ineficientes)<br><br>No se introducen sesgos en los coeficientes.<br>Se afectan levemente sus varianzas (son coeficientes ineficientes) | 1) No haga nada (Green 1999).<br><br>2) elimine uno de los regresores más correlacionados<br><br>3) Sustituya los regresores correlacionados por un índice<br><br>4) Haga un análisis factorial (Cortés 1987)<br><br>5) Considere una re-especificación del modelo con interacciones |