# Tema 6: Multicolinealidad

Máximo Camacho

# Multicolinealidad

- Bloque I: El modelo lineal clásico
  - Tema 1: Introducción a la econometría
  - Tema 2: El modelo de regresión lineal
  - Tema 3: El método MCO
  - Tema 4: Propiedades de la estimación MCO
  - Tema 5: Inferencia y predicción
- Bloque II: Extensiones al modelo lineal clásico
  - ***Tema 6: Multicolinealidad***
  - Tema 7: Variables ficticias
  - Tema 8: Heteroscedasticidad
  - Tema 9: Endogeneidad

# Descripción de la clase

- Introducción
- Multicolinealidad exacta
  - Consecuencias sobre la estimación
  - ¿Cómo detectarla?
  - ¿Cómo corregirla?
- Multicolinealidad aproximada
  - Consecuencias sobre la estimación
  - ¿Cómo detectarla?
  - ¿Cómo corregirla?
- Conclusiones

# 1. Introducción

## 1.1. Ejemplo de clase

- En California los responsables de educación quieren estudiar notas en 420 colegios. Datos en 1998
  - Notas $Y_i$
  - Ratio estudiantes por profesor $X_{1i}$ (REP)
  - Porcentaje de alumnos que no hablan bien el idioma $X_{2i}$ (PNI)
  - Porcentaje de alumnos que pueden pedir ayuda para comedor $X_{3i}$ (PAC)
  - Porcentaje de alumnos que pueden pedir ayuda por renta baja $X_{4i}$ (PAR)

  **nuevas**
- ¿Cómo estimamos esta relación?
- Modelo lineal clásico

# 1. Introducción

## 1.2. Supuestos del modelo lineal clásico

- Suponemos **relación lineal** entre las variables

  $Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$ $Y = X\beta + \varepsilon$ $Y_i = \chi_i{}'\beta + \varepsilon_i$

- Supuestos
  - Exogeneidad débil $E(\varepsilon_i / \chi_i) = E(\varepsilon_i) = 0$
  - Muestras aleatorias $\Rightarrow$ $E(\varepsilon_i / \chi_j) = E(\varepsilon_i) = 0$ $E(\varepsilon_i \varepsilon_j) = E(\varepsilon_i) E(\varepsilon_j) = 0$
  - Momentos cuartos finitos $0 < E(\varepsilon_i^4) < \infty, 0 < E(X_{1i}^4) < \infty, \ldots, 0 < E(X_{ki}^4) < \infty$
  - No multicolinealidad exacta $X_1, \ldots, X_n$ no son linealmente dependientes
  - Normalidad $\varepsilon / X \sim N$
  - Homoscedasticidad $\operatorname{var}(\varepsilon_i / X) = \sigma^2 \forall i$

# 2. Multicolinealidad exacta

## 2.1. Concepto

- Definición
  - Una o varias variables explicativas son una combinación lineal de otra(s)
- Ejemplos económicos
  - $\text{Renta regional}_i = \beta_0 + \beta_1 \text{interés}_i + \varepsilon_i$
  - $\text{Cotización}_i = \beta_0 + \beta_1 \text{ingresos}_i + \beta_2 \text{gastos}_i + \beta_3 \text{beneficio}_\text{i} + \varepsilon_i$
- La matriz de explicativas $X$ tiene columnas linealmente dependientes

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \overbrace{\begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix}}^{X} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

# 2. Multicolinealidad exacta

## 2.2. Implicación para el modelo

- No podemos encontrar de forma única $\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$

$$rango(X) < K \Rightarrow |X'X| = 0 \Rightarrow \nexists (X'X)^{-1}$$

  - Teoría: Hemos excluido este caso por supuesto
  - Pero podría aparecer en aplicaciones prácticas
- ¿Cómo detectarlo?
  - Los programas “se quejarán” de que no podemos invertir matriz (*X’X*)
  - En Eviews aparece el mensaje “near singular matrix”
- ¿Cómo corregirlo?
  - Se deben a errores del investigador al introducir las explicativas
  - Al aparecer mensaje de error, corregiremos las explicativas

# 2. Multicolinealidad exacta

## 2.2. Implicación para el modelo

- Corrección

  - En el ejemplo de la renta regional, supongamos $interés_i = 4$

$$\text{Renta regional}_i = \beta_0 + \beta_1 \text{interés}_i + \varepsilon_i \Longrightarrow \text{Renta regional}_i = \beta_0 + \beta_1 4 + \varepsilon_i$$

$$\text{Renta regional}_i = \alpha_0 + \varepsilon_i \qquad \alpha_0 = \beta_0 + \beta_1 4$$

  - En el ejemplo de la cotización sabemos $beneficios_i = ingresos_i - gastos_i$

$$\text{Cotización}_i = \beta_0 + \beta_1 \text{ingresos}_i + \beta_2 \text{gastos}_i + \beta_3 \text{beneficio}_i + \varepsilon_i$$

$$\text{Cotización}_i = \alpha_0 + \alpha_1 \text{ingresos}_i + \alpha_2 \text{gastos}_i + \varepsilon_i \quad \begin{cases} \alpha_0 = \beta_0 \\ \alpha_1 = \beta_1 + \beta_3 \\ \alpha_2 = \beta_2 - \beta_3 \end{cases}$$

# 3. Multicolinealidad aproximada

## 3.1. Concepto

- Definición
  - Una o varias variables explicativas son una combinación lineal aproximada de otra(s)
  - Supongamos que $X_1$ es buena explicativa pero comb. lineal aproximada de las demás

$$X_{1i} = \alpha_0 + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki} + u_{1i}$$

- Ejemplos económicos

$\left\{ \begin{array}{l} \text{Porcentaje de alumnos que pueden pedir ayuda por renta baja} \\ \text{Porcentaje de alumnos que pueden pedir ayuda para comedor} \end{array} \right.$

$\left\{ \begin{array}{l} \text{Ayudas que recibe una colegio para libros} \\ \text{Ayudas que recibe el colegio para instalaciones} \end{array} \right.$

$\left\{ \begin{array}{l} \text{Gasto público en carreteras} \\ \text{Gasto público en mejorar otras comunicaciones} \end{array} \right.$

# 3. Multicolinealidad aproximada

## 3.2. Implicación para el modelo

- Podemos encontrar de forma única $\hat{\beta}_{MCO} = (X'X)^{-1} X'Y$

  $rango(X) = K \Rightarrow |X'X| \neq 0 \Rightarrow \exists (X'X)^{-1}$

  - Nota: Hemos excluido la multicolinealidad exacta por el supuesto
  - Estimadores cumplen buenas propiedades y contrastes e intervalos como siempre

- ¿Qué problemas genera en la estimación?
  - Para entenderlo, supongamos que hacemos la regresión

    $$X_{1i} = \alpha_0 + \alpha_2 X_{2i} + \ldots + \alpha_k X_{ki} + u_{1i}$$

  - Y definimos

    $$R_1^2 = 1 - \frac{SCR_1}{STC_1} = 1 - \frac{\sum \hat{u}_{1i}^2}{\sum \left(X_{1i} - \bar{X}_1\right)^2}$$

# 3. Multicolinealidad aproximada

## 3.2. Implicación para el modelo

- ¿Qué problemas genera en la estimación?
  - Podemos demostrar (ejercicios de clase y Wooldridge, pág. 102) que

$$\hat{\beta}_1 = \beta_1 + \frac{\sum \hat{u}_{1i}\varepsilon_i}{\sum \hat{u}_{1i}^2} \Rightarrow \operatorname{var}\left(\hat{\beta}_1 / X\right) = \frac{\sigma^2}{\sum \hat{u}_{1i}^2} = \frac{\sigma^2}{\left(1 - R_1^2\right)\sum\left(X_{1i} - \bar{X}_1\right)^2}$$

  - Cuanta mayor relación lineal entre $X_1$ y el resto mayor varianza de $\hat{\beta}_1$
  - Estimación imprecisa e intervalos de confianza muy grandes $\left(\hat{\beta}_1 \pm t_{n-K,\alpha/2}\sqrt{\widehat{\operatorname{var}}\left(\hat{\beta}_1\right)}\right)$
  - Ejemplos:
    - Si $X_1$ no se relacionara ( $R^2_1 = 0$ ) $\operatorname{var}\left(\hat{\beta}_1 / X\right) = \sigma^2 / \sum\left(X_{1i} - \bar{X}_1\right)^2$
    - Si $R^2_1 = 0.5$ esa varianza se duplica
    - Si $R^2_1 = 0.9$ esa varianza se multiplica por 10

# 3. Multicolinealidad aproximada

## 3.2. Implicación para el modelo

- Ejemplo simulado (Novales, pág 346)
  - Se generan 50 tripletas de vectores (112 x 1) $\left(\varepsilon^j, X_1^j, X_2^j\right)$ (j=1, 2, …50)
  - Bajo 3 supuestos de $R^2_1$ (0, 0.9 y 0.99). Siempre se cumplen los supuestos clásicos
  - Se generan $Y^j$ conociendo la recta poblacional $Y_i^j = 8 + 5X_{1i}^j - 3X_{2i}^j + \varepsilon_i^j$
  - Se estima MCO 50 veces y los resultados medios

| Medias | $R^2_1 = 0.00$ | $R^2_1 = 0.90$ | $R^2_1 = 0.99$ |
|---|---|---|---|
| $\hat{\beta}$ | ( 8.0 , 5.1, -3.0 ) | ( 7.9, 5.1,-3.2 ) | ( 7.9, 5.2, -3.3 ) |
| $\text{var}\left(\hat{\beta}_1\right)$ | 0.23 | 1.29 | 11.22 |
| $\text{var}\left(\hat{\beta}_2\right)$ | 0.23 | 1.06 | 11.03 |

casi esperanzas (insesgados)

aumentan mucho

# 3. Multicolinealidad aproximada

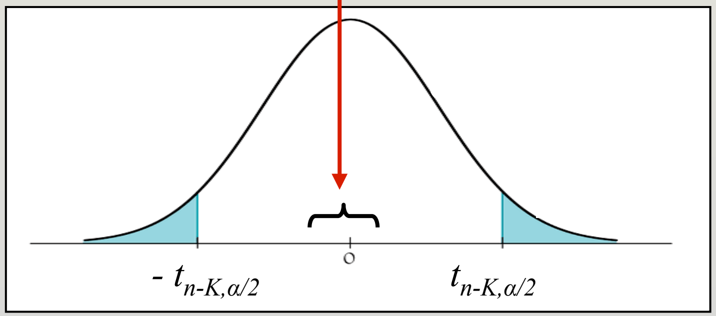
## 3.2. Implicación para el modelo

- Puede indicar artificialmente variables no son significativas individualmente
  - Supongamos que hacemos el contraste

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i \qquad H_0 : \beta_i = 0 \qquad H_a : \beta_i \neq 0$$

  - El estadístico tiene varianza muy grande y tiende a caer en zona de no $RH_0$
    - Con independencia de que $X_1$ se relacione con $Y$
    - Aunque el $R^2_c$ sea alto y no caiga al quitar $X_1$

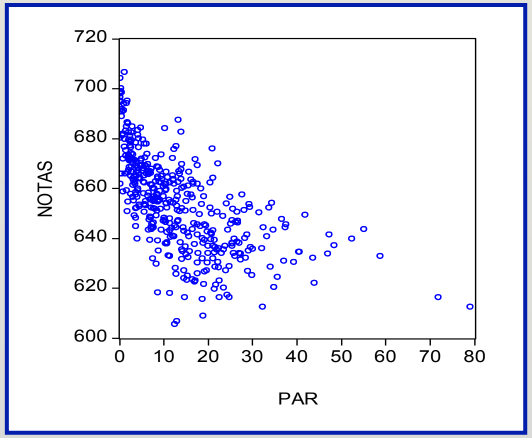
$$t^* = \frac{\hat{\beta}_i}{\uparrow\sqrt{\hat{\mathrm{var}}\left(\hat{\beta}_i\right)}} \sim t_{n-K}$$

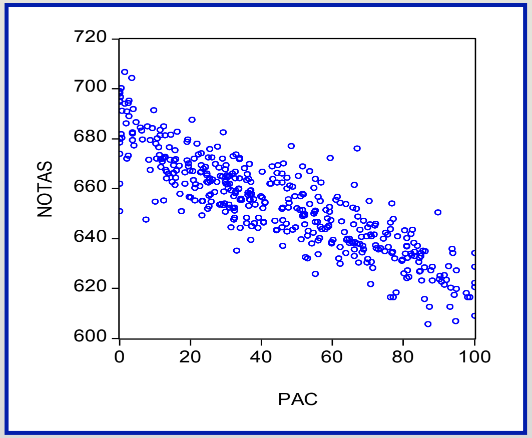


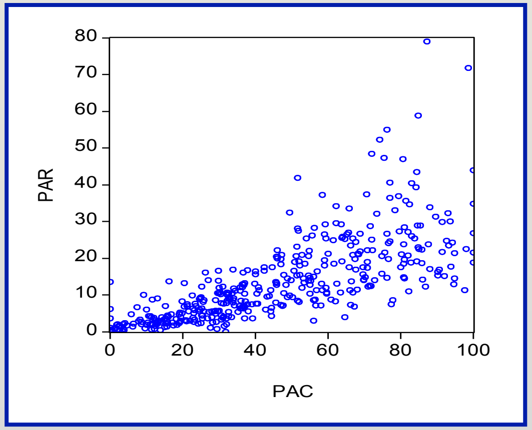
# 3. Multicolinealidad aproximada

## 3.3. Métodos de detección de multicolinealidad

- Método 1: gráficos de las explicativas
  - En nuestro ejemplo $Notas_i = \beta_0 + \beta_1 REP_i + \beta_2 PNI_i + \beta_3 PAC_i + \beta_4 PAR_i + \varepsilon_i$
  - Esperamos que PAR y PAC se relacionen (negativamente) con la dependiente
  - Esperamos que PAR y PAC se relacionen linealmente (y positivamente)
  - **Crítica**: los gráficos siempre nos pueden engañar



# 3. Multicolinealidad aproximada

## 3.3. Métodos de detección de multicolinealidad

- Método 2: regresar explicativas entre ellas y ver $R^2_j$
  - Sabemos que $$\operatorname{var}\left(\hat{\beta}_j / X\right) = \frac{\sigma^2}{\left(1 - R_j^2\right)\sum\left(X_{ji} - \bar{X}_j\right)^2}$$
  - El problema viene dado por $R^2_j$ altos que indican alta relación lineal
  - Veamos como son los $R^2_j$

$$REP_i = \alpha_0 + \alpha_1 PNI_i + \alpha_2 PAC_i + \alpha_3 PAR_{3i} + \varepsilon_i \Rightarrow R_1^2 = 0.04$$

$$PNI_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PAC_i + \alpha_3 PAR_{3i} + \varepsilon_i \Rightarrow R_2^2 = 0.49$$

$$PAR_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PNI_i + \alpha_3 PAC_{3i} + \varepsilon_i \Rightarrow R_3^2 = 0.60$$

$$PAC_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PNI_i + \alpha_3 PAR_{3i} + \varepsilon_i \Rightarrow R_4^2 = 0.74$$

  - Parece que el problema puede estar con PAR o con PAC

# 3. Multicolinealidad aproximada

## 3.3. Métodos de detección de multicolinealidad

- **Crítica** ¿cómo de grande debe ser $R^2_j$ para preocuparnos por multicolinealidad?
  - En la literatura ha habido algunos intentos de acotarlo
  - Ej: Klien (1962). Sólo nos preocupa si $R^2_j > R^2$

$$Notas_i = \beta_0 + \beta_1 REP_i + \beta_2 PNI_i + \beta_3 PAC_i + \beta_4 PAR_i + \varepsilon_i \Rightarrow R^2 = 0.77$$

$$PAR_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PNI_i + \alpha_3 PAC_{3i} + \varepsilon_i \Rightarrow R_3^2 = 0.60$$

$$PAC_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PNI_i + \alpha_3 PAR_{3i} + \varepsilon_i \Rightarrow R_4^2 = 0.74$$

  - ¿Debemos preocuparnos por multicolinealidad?

- No hay ninguna razón objetiva para usar esta cota

# 3. Multicolinealidad aproximada

## 3.3. Métodos de detección de multicolinealidad

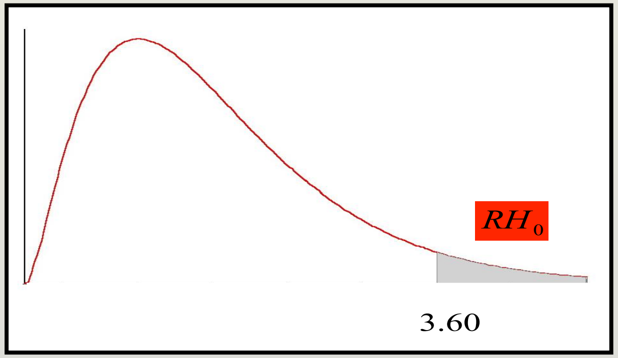
- Método 3: regresar explicativas entre ellas y contrastes significatividad conjunta
  - Sabemos que $REP_i = \alpha_0 + \alpha_1 PNI_i + \alpha_2 PAC_i + \alpha_3 PAR_i + \varepsilon_i$
  - CSC: $H_0 : \alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0$ $\quad F_j^* = \dfrac{R_j^2 / 3}{\left(1 - R_j^2\right)/(n-4)}$

$$REP_i = \alpha_0 + \alpha_1 PNI_i + \alpha_2 PAC_i + \alpha_3 PAR_{3i} + \varepsilon_i \Rightarrow F_1^* = 5.77$$

$$PNI_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PAC_i + \alpha_3 PAR_{3i} + \varepsilon_i \Rightarrow F_2^* = 133$$

$$PAR_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PNI_i + \alpha_3 PAC_{3i} + \varepsilon_i \Rightarrow F_3^* = 208$$

$$PAC_i = \alpha_0 + \alpha_1 REP_i + \alpha_2 PNI_i + \alpha_3 PAR_{3i} + \varepsilon_i \Rightarrow F_4^* = 394$$




- **Crítica**. Con $n$ grande $RH_0$ con demasiada frecuencia: recuerda que $R^2_1 = 0.04$ !!!

# 3. Multicolinealidad aproximada

## 3.3. Métodos de detección de multicolinealidad

- Método 4: contradicción contraste significatividad global e individuales
  - Supongamos que en el modelo

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

  - Las explicativas $X_{2i}$ , $X_{3i}$ y $X_{4i}$ se relacionan casi linealmente
  - Si las variables se relacionan con la dependiente
    - El contraste de significatividad global, menos afectado por la multicolinealidad, puede indicar significatividad (incluso al 1%)
    - Y el $R^2$ puede ser alto indicando un buen ajuste
  - Sus contrastes de significatividad individuales
    - Pueden indicar no significatividad
    - Sólo una variable ($X_{1i}$ ) significativa

# 3. Multicolinealidad aproximada

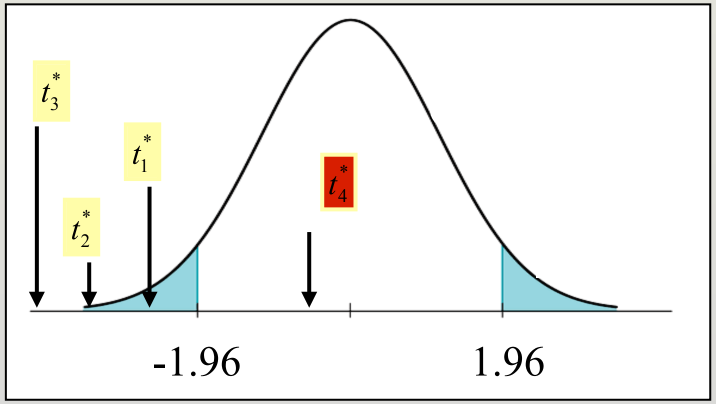
## 3.3. Métodos de detección de multicolinealidad

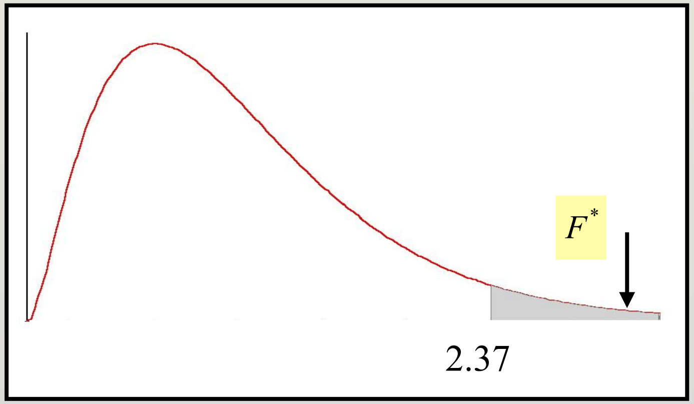
- **Crítica**: casos tan claros no los tendremos en la realidad

$$\hat{Notas}_i = \underset{(4.69)}{700.39} - \underset{(0.23)}{1.01} REP_i - \underset{(0.03)}{0.13} PNI_i - \underset{(0.03)}{0.52} PAC_i - \underset{(0.06)}{0.04} PAR_i \Rightarrow R^2 = 0.77$$

$$t_i^* = \frac{\hat{\beta}_i}{\sqrt{\hat{\mathrm{var}}\left(\hat{\beta}_i\right)}} \begin{cases} t_1^* = -4.23 \\ t_2^* = -3.81 \\ t_3^* = -16.42 \\ t_4^* = -0.78 \end{cases}$$

$$F^* = \frac{R^2/4}{\left(1-R^2\right)/(n-5)} = 357$$



# 3. Multicolinealidad aproximada

## 3.3. Métodos de detección de multicolinealidad

- Globalmente significativas
- PAR es la única no significativa
- ¿Se debe a multicolinealidad o a que no es una buena explicativa?
  - Si se debe a multicolinealidad: vamos a ver soluciones
  - Si no se relaciona con la dependiente: deberíamos quitarla del modelo
  - ¿Pero cómo lo sabremos?
  - Lo mejor es acudir al sentido común
- ¿Cuáles son las soluciones a la multicolinealidad?

# 3. Multicolinealidad aproximada

## 3.4. Soluciones a la multicolinealidad

- Solución 1: añadir información extra-muestral
  - Ampliar la muestra: usando otros colegios de California
    - **Crítica**: debemos usar toda la información disponible en la primera estimación
  - Usar la estimación de otra muestra
    - Supongamos que usando datos de Texas hemos estimado

$$\hat{Notas}_i = -0.20 PAR_i$$

    - Para los datos de California estimamos

$$Notas_i = \beta_0 + \beta_1 REP_i + \beta_2 PNI_i + \beta_3 PAC_i - 0.20 PAR_i + \varepsilon_i$$

    - **Crítica**: ¿por qué $\hat{\beta}_4^{Texas} = \hat{\beta}_4^{California}$ ? $\Rightarrow$ Restricciones falsas sesgan la estimación

# 3. Multicolinealidad aproximada

## 3.4. Soluciones a la multicolinealidad

- Solución 2: Usar estimadores alternativos a MCO
  - Estimador de Cresta
    - Buscamos un *c* y un estimador de menor varianza

$$\hat{\beta}_C = (X'X + cI)^{-1} X'Y \qquad \operatorname{var}(\hat{\beta}_c / X) = \sigma^2 (X'X + cI)^{-1} X'X (X'X + cI)^{-1}$$

    - **críticas**: el estimador propuesto es sesgado y cómo elegir *c*
  - Estimador de componentes principales
    - Buscamos combinación lineal de las columnas de $X$: $Z = XB$
    - $B$ se busaca de forma que las $Z_i$ sean ortogonales (elimina multicolinealidad)
    - Regresamos usando Z $\quad Y = Z\alpha + \varepsilon$
    - **críticas**: $\hat{\alpha} \begin{cases} \text{difíciles de interpretar} \\ \text{sesgados} \end{cases} \Bigg\}$ da igual si objetivo es predecir

# 3. Multicolinealidad aproximada

## 3.4. Soluciones a la multicolinealidad

- Solución 3: Eliminar la variable no significativa ($PAR_i$)
  - (Ver tema 5: consecuencias de imponer restricciones ciertas y falsas)
  - Supongamos que eliminamos la variable: imponemos $\beta_4 = 0$
    - Si no se relaciona con las notas ⇒ Imponemos una restricción cierta

      (No sesgo y reducimos la varianza)
    - Si se relaciona con las notas ⇒ Imponemos una restricción falsa

      (Introducimos sesgo y reducimos la varianza)
  - ¿Cuándo merece la pena asumir el riesgo de eliminarla?
    - Analizaremos el error cuadrático medio

# 3. Multicolinealidad aproximada

## 3.4. Soluciones a la multicolinealidad

- Supongamos un modelo más sencillo

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \Longrightarrow \hat{\beta}_1$$

- Pensamos que $X_2$ presenta problemas de multicolinealidad con $X_3$ : si $\beta_2 = 0$

$$Y_i = \beta_1 X_{1i} + \varepsilon_i \Longrightarrow \tilde{\beta}_1$$

- ¿ECM? (Ejercicios de clase y Novales, pág 361)

$$ECM\left(\tilde{\beta}_1\right) = \mathrm{var}\left(\hat{\beta}_1\right) + sesgo^2$$

$$\frac{ECM\left(\tilde{\beta}_1\right)}{ECM\left(\hat{\beta}_1\right)} = 1 + R_2^2\left(t^2 - 1\right) \qquad t = \left(\frac{\beta_i}{\sqrt{\mathrm{var}(\beta_i)}}\right)$$

- Sólo merece la pena si | $t$ |<1 ⇒ algunos autores piden | $t^*$ |<<1
- **críticas**:
  - ¿Cuánto podemos considerar $t^*$<<1?
  - ¿Queremos estimadores sesgados?

# 3. Multicolinealidad aproximada

## 3.4. Soluciones a la multicolinealidad

- En nuestro ejemplo

$$\hat{Notas}_i = \underset{(4.69)}{700.39} - \underset{(0.23)}{1.01}\, REP_i - \underset{(0.03)}{0.13}\, PNI_i - \underset{(0.03)}{0.52}\, PAC_i - \underset{(0.06)}{0.04}\, PAR_i$$

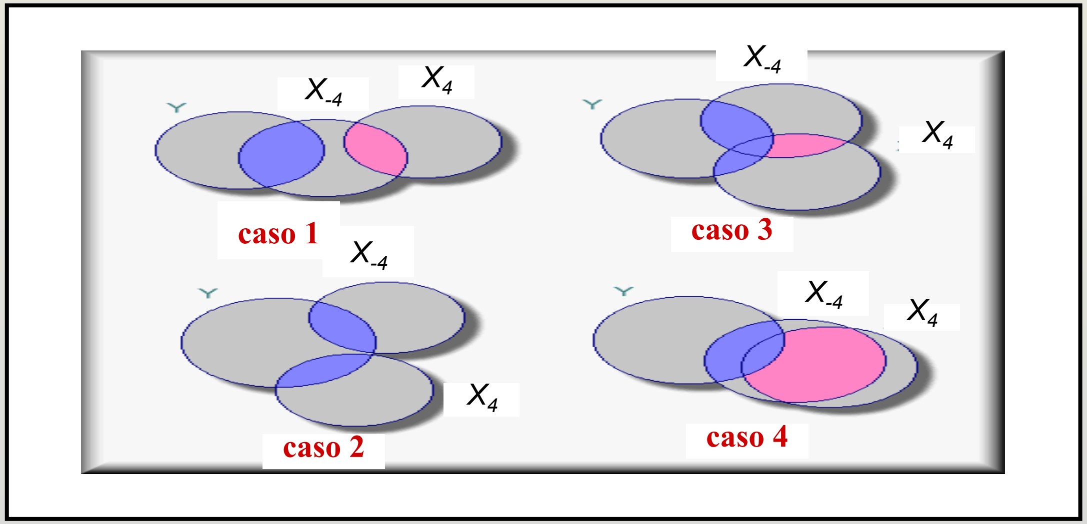
$$\left|t_4^*\right| = \left|\frac{\hat{\beta}_4}{\sqrt{\hat{\mathrm{var}}\left(\hat{\beta}_4\right)}}\right| = 0.78$$

- ¿Qué hacemos?
  - Mi consejo es acudir al sentido común (teoría económica)
  - Parece que tanto PAC como PAR contienen la misma información sobre los colegios: nos dicen cuáles tienen más alumnos pobres
  - Podemos eliminar PAR porque PAC ya contiene la información necesaria para estimar el modelo de notas

$$\hat{Notas}_i = \underset{(4.69)}{700.15} - \underset{(0.24)}{0.99}\, REP_i - \underset{(0.03)}{0.12}\, PNI_i - \underset{(0.02)}{0.55}\, PAC_i$$

# 4. Conclusión

- $X_4$ puede tener distintas relaciones con las demás ( $X_{-4}$ ) y la dependiente
  - $X_4$ no explicativa y multico, : $X_4$ la eliminamos con contrastes
  - $X_4$ explicativa y no multico, $X_4$ explicativa: $X_4$ será explicativa final
  - $X_4$ explicativa y multico no severa: $X_4$ será explicativa tras los contrastes
  - $X_4$ explicativa y multico severa: acudir al sentido común más contrastes



# 5. ¿Qué hemos aprendido?

- Multicolinealidad
  - Relación lineal entre las variables explicativas del modelo
- Exacta
  - Teoría: nunca hay porque suponemos
  - Práctica: fácil de detectar y resolver
- Aproximada
  - Da lugar a estimaciones imprecisos, varianzas e intervalos de confianza grandes
  - Difícil de detectar y de corregir
- ¿Qué hacer?
  - Asegurarnos de que tenemos multicolinealidad: todas pruebas y teoría económica
  - No significativa y la información sobre $Y$ ya está en las que quedan: eliminarla
  - No significativa pero la información sobre $Y$ ya está en las que quedan: la dejamos