Tema 5. Diagnosis y validación del modelo de regresión lineal múltiple

1. Multicolinealidad
    - 1.1 Identificación y consecuencias
    - 1.2 Tratamiento
2. El análisis de los residuos
    - 2.1 Propiedades de los residuos
    - 2.2 Análisis gráfico de los residuos
3. Errores de especificación
4. Observaciones influyentes y atípicas
    - 4.1 Robustez a priori. Los efectos palanca de las observaciones
    - 4.2 La robustez a posteriori del modelo
    - 4.3 Datos atípicos

En este tema vamos a ver como:

- **identificar los problemas** que surgen al construir el modelo de regresión,
- sus **efectos** sobre las propiedades del modelo,
- cómo **reformular el modelo** para adecuarlo a la realidad estudiada.

1. Multicolinealidad

El primer problema que surge es la **dependencia de las variables explicativas (o regresores) entre sí**, es decir, la existencia de una o más combinaciones lineales entre las columnas de la matriz $\mathbf{X}$

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ & & \dots & \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix}, \quad \text{rang}(\mathbf{X}) = k+1$$

Problema! cuando $\text{rang}(\mathbf{X}) < k+1$.

Cuando esto ocurre es difícil separar los efectos de cada variable explicativa y medir la contribución individual, con lo que los estimadores individuales serán **inestables** y con **gran varianza**. A este problema se le denomina **multicolinealidad** y consiste en **querer extraer de la muestra más información de la que contiene**.

Existen dos tipos de multicolinealidad:

1. **Multicolinealidad perfecta**

   Una de las variables explicativas es **combinación lineal exacta** de las demás:

   $\text{rang}(\mathbf{X}) < k+1 \Rightarrow \det(\mathbf{X}'\mathbf{X}) = 0 \Rightarrow$ no se puede calcular $(\mathbf{X}'\mathbf{X})^{-1}$

   El sistema de ecuaciones que determina el vector $\hat{\boldsymbol{\beta}}$ no tiene solución única.

2. **Alta multicolinealidad**

   Cuando alguna o todas las variables explicativas están **altamente correlacionadas** entre sí (pero el coeficiente de correlación no llega a ser 1 ni -1). En este caso las columnas de la matriz $\mathbf{X}$ tienen un alto grado de dependencia entre sí, pero sí puede calcularse el vector $\hat{\boldsymbol{\beta}}$, aunque:

   a) Los estimadores $\hat{\boldsymbol{\beta}}$ tendrán **varianzas muy altas,** lo que provocará **mucha imprecisión** en la estimación de los $\hat{\beta}_j$ y, por tanto, los **I.C. serán muy anchos**.

   b) Los estimadores $\hat{\beta}_j$ serán **muy dependientes entre sí**, puesto que tendrán **altas covarianzas** y habrá poca información sobre lo que ocurre al variar una variable si las demás permanecen constantes.

1.1 Identificación y consecuencias de la multicolinealidad

Consecuencias:

- Los estimadores $\hat{\beta}_j$ serán **muy sensibles** a pequeñas variaciones en el tamaño muestral o a la supresión de una variable aparentemente no significativa. A pesar de esto, la predicción no tiene por qué verse afectada ante la multicolinealidad, ni ésta afecta al vector de residuos, que está siempre bien definido.
- Los coeficientes de regresión pueden ser **no significativos individualmente** (puesto que las varianzas de los $\hat{\beta}_j$ van a ser grandes), aunque el contraste global del modelo sea significativo.
- La multicolinealidad puede afectar mucho a algunos parámetros y nada a otros. Los parámetros que estén **asociados a variables explicativas poco correlacionadas con el resto no se verán afectados y podrán estimarse con precisión**.

Identificación de la multicolinealidad

La identificación de variables correlacionadas se realiza examinando

1) La matriz de correlaciones entre las variables explicativas, $\mathbf{R}$, y su inversa, $\mathbf{R}^{-1}$,
2) Los autovalores de $\mathbf{X'X}$ o de $\mathbf{R}$.

1) La presencia de **correlaciones altas** entre variables explicativas es un **indicio** de multicolinealidad. Pero, es posible que exista una relación perfecta entre una variable y el resto y, sin embargo, sus coeficientes de correlación sean bajos (por ejemplo, cuando sea el caso de una relación no lineal).

Definimos la **matriz de correlaciones** como

$$\mathbf{R} = \begin{pmatrix} 1 & r_{12} & r_{13} & \dots & r_{1k} \\ r_{12} & 1 & r_{23} & \dots & r_{2k} \\ & & & \dots & \\ r_{1k} & r_{2k} & r_{3k} & \dots & 1 \end{pmatrix}, \quad \text{donde } r_{ij} = \frac{s_{x_i,x_j}}{s_{x_i} s_{x_j}}, \quad -1 \le r_{ij} \le 1,$$

que es una matriz de orden k, simétrica, con unos en la diagonal.

La inversa de la matriz de correlaciones:

$$\mathbf{R}^{-1} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1k} \\ \gamma_{12} & \gamma_{22} & \gamma_{23} & \cdots & \gamma_{2k} \\ & \cdots & & \cdots & \\ \gamma_{1k} & \gamma_{2k} & \gamma_{3k} & \cdots & \gamma_{kk} \end{pmatrix}$$

tiene en cuenta la **dependencia conjunta.** Los elementos de su diagonal se denominan **factores de incremento** o **de inflación de la varianza** y verifican:

$$\gamma_{ii} = FIV(i) = \frac{1}{1 - R^2_{i.resto}}, \quad i = 1,...,k,$$

donde $R^2_{i.resto}$ es el coeficiente de determinación de la regresión de la variable x_i en función del resto de variables explicativas, es decir, $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k$.

Por tanto, si para algún i se tiene que:

$$\gamma_{ii} > 10 \Leftrightarrow \frac{1}{1 - R^2_{i.resto}} > 10 \Leftrightarrow 1 - R^2_{i.resto} < 0.1 \Leftrightarrow R^2_{i.resto} > 0.9$$

es decir, la variable x_i se explica en un 90% por el resto de variables explicativas, por tanto, estamos en una situación de **alta multicolinealidad**.

Inconveniente: $\mathbf{R}^{-1}$ se calculará con poca precisión cuando $\mathbf{R}$ sea casi singular ($\det(\mathbf{R}) \approx 0$).

2) Las mejores medidas de singularidad de $\mathbf{X'X}$ o de $\mathbf{R}$ utilizan los autovalores de estas matrices. Un índice de singularidad, que se utiliza en cálculo numérico, es el **índice de condicionamiento** (*condition number*).

Si $\mathbf{M}$ es una matriz de orden k, simétrica y definida positiva, y $\lambda_1 < \lambda_2 < \ldots < \lambda_k$ son sus autovalores, se define el **índice de condicionamiento** de $\mathbf{M}$ como:

$$\text{cond}(\mathbf{M}) = \sqrt{\frac{\lambda_k}{\lambda_1}} \geq 1$$

Es más conveniente calcular este índice para $\mathbf{R}$ que para $\mathbf{X'X}$, con el fin de evitar la influencia de las escalas de medida de los regresores.

Para saber si existe o no multicolinealidad, calcularemos $\text{cond}(\mathbf{R})$ y si

- $\text{cond}(\mathbf{R}) > 30$ → **alta multicolinealidad**
- $10 < \text{cond}(\mathbf{R}) < 30$ → **multicolinealidad moderada**
- $\text{cond}(\mathbf{R}) < 10$ → **ausencia de multicolinealidad** (la matriz $\mathbf{R}$ está bien definida).

1.2 Tratamiento de la multicolinealidad

Cuando la recogida de datos se diseñe a priori, la multicolinealidad puede evitarse tomando las observaciones de manera que la matriz $\mathbf{X'X}$ sea diagonal, lo que aumentará la precisión en la estimación (los estimadores tendrán varianza pequeña).

La multicolinealidad **es un problema de la muestra** y, por tanto, **no tiene solución simple,** ya que estamos pidiendo a los datos más información de la que contienen.

Las dos únicas soluciones son:

1) Eliminar regresores, reduciendo el número de parámetros a estimar,
2) Incluir información externa a los datos.

1) La primera solución conduce a eliminar, bien variables muy correlacionadas con las que se incluyen, o bien ciertas combinaciones lineales de ellas mediante componentes principales.

a) **Eliminación de variables de la ecuación**. Se puede mejorar, en promedio, el error cuadrático medio de la estimación de los parámetros, si se eliminan aquellas variables cuyo estadístico

$$\left|t_{\exp}\right| = \left|\frac{\hat{\beta}_j}{s_R\sqrt{q_{jj}}}\right| < 1.$$

b) En lugar de eliminar directamente variables, se pueden considerar **las componentes principales** (Técnica de Análisis Multivariante) y considerar como regresores las compionentes más importantes, es decir, las que tienen mayor autovalor asociado, que son las que explican mayor porcentaje de la variabilidad de los datos

2) La segunda solución es introducir información externa mediante el **enfoque bayesiano** que conduce a los **estimadores contraídos** que se presentan en el apéndice 13B de Peña (1995, vol. II). Estos estimadores pueden justificarse como método de reducción del error cuadrático medio, y su utilización es polémica.

- **Ver ejemplo multicolinealidad**

2. El análisis de los residuos

Modelo de regresión lineal múltiple: $\mathbf{Y}=\mathbf{X}\ \beta\ +\mathbf{U}$, donde $\mathbf{X}$ es la matriz de regresores $nx(k+1)$, β es el vector de parámetros $(k+1)x1$ y $\mathbf{U}$ es un vector $nx1$ con ley normal multivariante $NM_n(\mathbf{0},\sigma^2\ \mathbf{I})$.

Una vez estimado el modelo de regresión lineal múltiple tendremos que comprobar las hipótesis de linealidad, normalidad, homocedasticidad e independencia, realizando un estudio de los residuos.

2.1 Propiedades de los residuos

Matricialmente se definen como:

$$\left.\begin{aligned}\mathbf{e}=\mathbf{Y}-\hat{\mathbf{Y}}=\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}=\mathbf{Y}-\underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{Y}=\mathbf{Y}-\mathbf{H}\mathbf{Y}=(\mathbf{I}-\mathbf{H})\mathbf{Y}\\ =(\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}+\mathbf{U})=\mathbf{X}\boldsymbol{\beta}-\mathbf{H}\mathbf{X}\boldsymbol{\beta}+\mathbf{U}-\mathbf{H}\mathbf{U}=\mathbf{X}\boldsymbol{\beta}-\underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}_{\mathbf{X}\boldsymbol{\beta}}+\mathbf{U}-\mathbf{H}\mathbf{U}\\ =\mathbf{U}-\mathbf{H}\mathbf{U}=(\mathbf{I}-\mathbf{H})\mathbf{U}\end{aligned}\right\}\Rightarrow \mathbf{e}=\underbrace{(\mathbf{I}-\mathbf{H})}_{\text{constante}}\mathbf{U}$$

Proposición: Puesto que $\mathbf{U}$ tiene ley normal multivariante $NM_n(\mathbf{0}, \sigma^2\,\mathbf{I})$, el vector $\mathbf{e}$ de residuos también tiene ley normal multivariante con vector de esperanzas y matriz de covarianzas:

$$\mathrm{E}(\mathbf{e}) = \mathbf{0},\ \ \mathrm{var}(\mathbf{e}) = \sigma^2(\mathbf{I}-\mathbf{H})\,.$$

Demostración:

$$\mathrm{E}(\mathbf{e}) = \mathrm{E}((\mathbf{I}-\mathbf{H})\mathbf{U}) = (\mathbf{I}-\mathbf{H})\,\mathrm{E}(\mathbf{U}) = \mathbf{0}$$

$$\mathrm{var}(\mathbf{e}) = \mathrm{E}(\mathbf{ee'}) = \mathrm{E}((\mathbf{I}-\mathbf{H})\mathbf{UU'}(\mathbf{I}-\mathbf{H})') = (\mathbf{I}-\mathbf{H})\underbrace{\mathrm{E}(\mathbf{UU'})}_{\sigma^2\mathbf{I}}\underbrace{(\mathbf{I}-\mathbf{H})'}_{\mathbf{I}-\mathbf{H}} = \sigma^2(\mathbf{I}-\mathbf{H})^2 = \sigma^2(\mathbf{I}-\mathbf{H}),$$

puesto que $\mathbf{I-H}$ es una matriz idempotente, al serlo también el proyector ortogonal $\mathbf{H}$:

$$(\mathbf{I}-\mathbf{H})^2 = (\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = (\mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{HH}) = (\mathbf{I}-\mathbf{H})$$

$$\mathbf{H}^2 = \mathbf{HH} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'X}(\mathbf{X'X})^{-1}\mathbf{X'} = \mathbf{H}$$

En particular, para $i=1,2,\ldots,n$, se tiene que $\mathrm{var}(e_i)=\sigma^2(1-v_{ii})$, donde v_{ii} es el elemento i-ésimo de la diagonal del proyector ortogonal $\mathbf{H}$.

Sustituyendo σ^2 por la varianza residual s_R^2 se obtendrá una estimación de la varianza del i-ésimo residuo.

Para comparar los residuos entre sí suele ser más ilustrativo estandarizarlos. Se definen los **residuos estandarizados** como:

$$r_i = \frac{e_i}{s_R\sqrt{1-v_{ii}}}$$

Problema! En la expresión anterior, numerador y denominador son **dependientes**, puesto que el residuo e_i se utiliza en el cálculo de s_R^2.

Esto puede solucionarse eliminando la observación i-ésima de la matriz de datos y estimando de nuevo el modelo con las n-1 observaciones restantes. Sean $\hat{\boldsymbol{\beta}}_{(i)}$ y $s_{R(i)}^2$ los estimadores así obtenidos (es decir, sin la observación i-ésima). Se demuestra que $s_{R(i)}^2$ tiene la siguiente expresión:

$$s_{R(i)}^2 = \frac{(n-k-1)s_R^2 - e_i^2 \Big/ \sqrt{1-v_{ii}}}{n-k-2}$$

Lo que significa que para obtener los $s_{R(i)}^2$ para i=1,2,…, n **no es necesario** re-estimar el modelo n veces, sino que se obtienen a partir de los valores de s_R^2, e_i^2 y v_{ii} del modelo completo.

Puesto que la observación i-ésima no interviene en el cálculo de $s^2_{R(i)}$, el residuo i-ésimo e_i es independiente de $s^2_{R(i)}$. Se define el **residuo estudentizado** como:

$$t_i = \frac{e_i}{s_{R(i)}\sqrt{1-v_{ii}}} \sim t_{n-k-2}$$

Estos tres tipos de residuos: e_i , r_i , t_i aportan información valiosa sobre los datos.

Si n es grande y los datos no contienen valores extremos, los tres tipos de residuos se comportan por igual. Pero en caso contrario, r_i y t_i suelen ser más informativos para detectar deficiencias en el modelo.

2.2 Análisis gráfico de los residuos

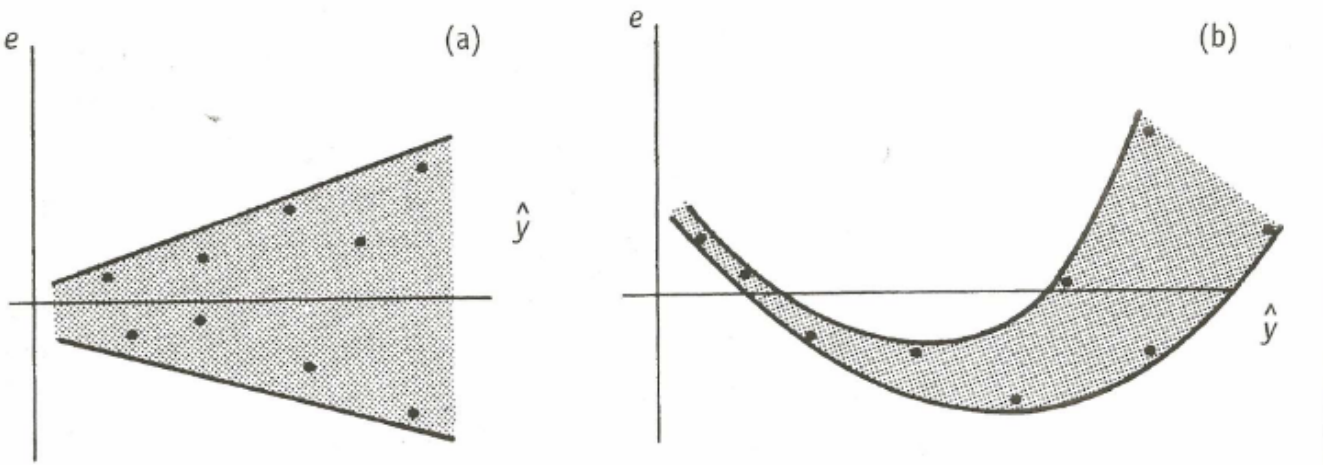
- **Histograma y gráfico probabilístico normal**

Útiles para analizar la **normalidad** de los residuos e identificar **valores atípicos**.
Cuando el número de datos sea al menos cuatro veces mayor que el número de parámetros estimados (es decir, $n > 4(k+1)$) podemos despreciar la dependencia entre los residuos.

- **Gráficos de residuos frente a valores predichos**

Útiles para identificar la **falta de linealidad**, **heterocedasticidad** y **valores atípicos**.
Su uso es análogo al que vimos en el Tema 2 de regresión lineal simple.
El gráfico puede realizarse con cualquiera de los tres tipos de residuos, aunque suelen utilizarse preferentemente e_i o r_i .

Figura 6.3 Heterocedasticidad, con falta de linealidad en (b)



- **Gráficos de residuos frente a variables explicativas**

Ayudan a identificar **si la falta de linealidad o la heterocedasticidad es debida a alguna variable explicativa**.

Es conveniente complementarlos con los gráficos parciales de residuos.

- **Gráficos parciales de residuos**

Útiles para estudiar la **relación entre la variable respuesta**, y, **y una variable regresora** x_j, eliminando el efecto de las demás variables explicativas $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$.

Para ello deberíamos:

1) Eliminar la columna correspondiente a la variable x_j en la matriz de datos.
2) Estimar el modelo de regresión con k-1 variables explicativas y obtener los residuos de esta regresión, que representan la parte de la variable respuesta no explicada por $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{k.}$
3) Representar gráficamente estos residuos frente a la variable eliminada x_j . Este gráfico mide el **efecto marginal** de x_j sobre la variable respuesta y.

- **Gráficos de residuos frente a variables omitidas**

La posible influencia de una **nueva variable no incluida** en el modelo, z, en la variable respuesta, y, puede detectarse representando los residuos respecto de ella.

Si la variable omitida es relevante, veremos una relación lineal entre los residuos y esta variable.

En particular, siempre que las observaciones se hayan recogido en orden cronológico o temporal conviene representar los residuos en función del tiempo.

3. Errores de especificación

Cometemos un **error de especificación** cuando **establecemos mal la dependencia** entre la variable respuesta y las variables explicativas. Esto ocurre si:

- Omitimos variables importantes,
- Introducimos variables innecesarias,
- Suponemos una relación lineal cuando la dependencia no es lineal.

Especificar incorrectamente las variables (omitir o añadir de innecesarias) produce **residuos con esperanza no nula**.

Especificar una relación lineal cuando la existente es no lineal es especialmente grave si se hacen **predicciones fuera del rango de datos**.

Consecuencias de especificar incorrectamente las variables

Excluir una variable afecta a la esperanza y a la varianza de los estimadores:

a) En cuanto a la esperanza, se produce un **sesgo** en los parámetros estimados que depende de la relación entre la variable excluida y las incluidas.

- Si la variable excluida es **incorrelacionada** con las incluidas, entonces el **sesgo** del estimador es **nulo** (**estimador insesgado**).
- El **sesgo aumenta** al **aumentar la correlación** entre la variable excluida y las incluidas.

b) En cuanto a la varianza, ésta es sesgada por exceso (es decir, mayor de lo que debería ser). Puesto que la raíz cuadrada de la varianza es el error estándar del estimador (que interviene en el cociente de los contrastes de significación individuales de los parámetros), esto puede conducir a **no detectar como significativas variables que sí lo son**, es decir:

$$\text{Si } s^2_{\hat{\beta}_j} = s^2_R q_{jj} \text{ es grande} \Rightarrow \left|t_{\exp}\right| = \left|\frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}\right| \text{ es pequeño}$$

por tanto, no se rechaza $H_0: \beta_j = 0$ → x_j es no significativa.

Incluir variables irrelevantes tiene también consecuencias graves, y tanto más desfavorables cuánto mayor sea la dependencia con las ya incluidas. Los estimadores seguirán siendo centrados (insesgados), pero

a) **su varianza aumentará mucho** si incluimos una variable **muy correlacionada** con las restantes,

b) si la variable que se incluye está **incorrelacionada** con las restantes, la varianza de los estimadores **no aumentará**, pero los estimadores **no serán eficientes** porque habremos invertido un grado de libertad en estimar un parámetro innecesario.

Identificación de los errores de especificación

Mediante los gráficos de residuos respecto a:

a) los valores predichos,
b) las variables explicativas x_j
c) nuevas variables potencialmente influyentes,
d) secuencia temporal, si los datos son cronológicos.

- **Ver ejemplo error de especificación**

4. Observaciones influyentes y atípicas

Es muy frecuente que los datos contengan observaciones atípicas o no generadas por el modelo.

Las observaciones atípicas son importantes porque pueden indicar aspectos nuevos del modelo (errores de medición, ausencia de variables relevantes, …) y es **importante identificarlas** porque pueden tener **mucho efecto en la estimación**.

Estudio de la robustez del modelo

Antes de aceptar como válido un modelo es siempre conveniente estudiar si las propiedades básicas del modelo son debidas a todo el conjunto de observaciones o, si por el contrario, estas propiedades **están condicionadas a un pequeño subconjunto de observaciones**.

- **Robustez a priori** o robustez del diseño de recogida de datos.
- **Robustez a posteriori** o robustez de los parámetros estimados, una vez observados los valores de la respuesta.

4.1. Robustez a priori. Los efectos palanca de las observaciones.

El **efecto palanca** (*leverage*) de cada observación es, como vimos en el Tema 3, la **capacidad del punto para atraer** a la ecuación de regresión.

Este efecto depende del valor

$$v_{ii} = \mathbf{x_i}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x_i} = \frac{1}{n}\left(1 + \underbrace{(\tilde{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})'\mathbf{S}_{\mathbf{XX}}^{-1}(\tilde{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}})}_{dist.Mahalanobis}\right)$$

donde

$\tilde{\mathbf{x}}_\mathbf{i} = (x_{1i}, x_{2i}, ..., x_{ki})$ es la observación i-ésima sin el término correspondiente a β_0,

$\bar{\mathbf{x}}$ es el vector de medias de las k variables explicativas (centro de gravedad o centroide),

$\mathbf{S}_{\mathbf{XX}}$ es la matriz de covarianzas de las k variables explicativas.

Recordemos que v_{ii} es una **medida de distancia** entre el punto $\tilde{\mathbf{x}}_\mathbf{i}$ y el centro de gravedad $\bar{\mathbf{x}}$.

Se consideran **puntos palanca** (*leverage points*) aquellos puntos cuyo v_{ii} sea elevado.

Si llamamos

$$\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_{ii} = \frac{\text{tr}(\mathbf{H})}{n} = \frac{k+1}{n}$$

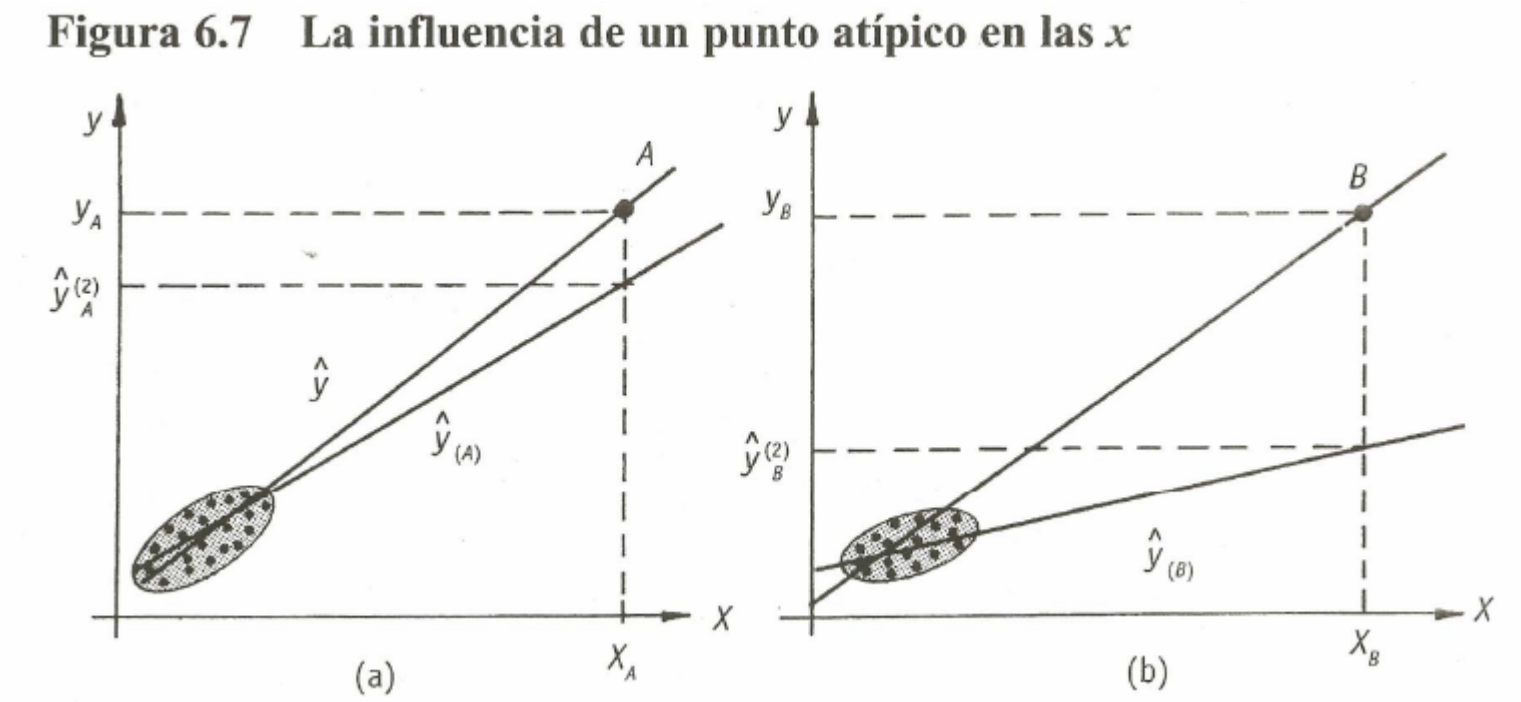
y consideramos el caso en que el número de regresores es $3 \leq k \leq 6$, con distribución normal conjunta, entonces diremos que la observación i-ésima es **extrema o potencialmente influyente** si

$$v_{ii} > 2\,\bar{v} = \frac{2\,(k+1)}{n}$$

4.2. La robustez a posteriori del modelo.

El hecho de que una observación sea **muy influyente a priori, no implica que realmente lo sea**. Lo será si, **al eliminarla, las propiedades del modelo estimado cambian mucho.**

Figura 6.7 La influencia de un punto atípico en las x



(a) El modelo es prácticamente el mismo con y sin el punto A,
(b) El punto B modifica por completo el modelo según si se tiene en cuenta o no.

Un punto o una observación será **influyente** si

a) Modifica el vector $\hat{\boldsymbol{\beta}}$ de parámetros estimados,

b) Modifica el vector $\hat{\mathbf{Y}}$ de predicciones,

c) Hace que la predicción del punto sea muy buena cuando éste se incluye en el modelo, y muy mala cuando se excluye.

Para **medir la influencia de un punto** se utiliza la **distancia de Cook**:

$$D(i) = \frac{r_i^2}{k+1}\left(\frac{v_{ii}}{1-v_{ii}}\right)$$

donde r_i es el residuo estandarizado, es decir:

$$r_i = \frac{e_i}{s_R\sqrt{1-v_{ii}}}$$

Diremos que un punto es **influyente a nivel** α si

$$D(i) \geq F^{\alpha}_{k+1,n-k-1},$$

donde $F^{\alpha}_{k+1,n-k-1}$ es el percentil $(1-\alpha)100\%$ de la ley F de Fisher con $k+1$ y $n-k-1$ grados de libertad.

- **Ver ejemplo puntos influyentes**

4.3. Datos atípicos

Diremos que un dato es **atípico** cuando **no se ha generado por el mismo mecanismo que el resto de las observaciones**.

Por ejemplo, ha ocurrido un error de medida, o esa observación tiene un valor distinto del resto para una variable relevante no incluida en el modelo (generalmente se debe a variables categóricas que no se han tenido en cuenta y que provocan la aparición de distintos grupos).

Para contrastar que un dato es atípico se utiliza su **residuo estudentizado**, t_i.

Bajo la hipótesis nula de que no existen atípicos, los residuos estudentizados tienen una ley t de Student con n-k-2 grados de libertad.

Sea $t_{\max} = \max(t_i)$. Para un nivel de significación α, diremos que la observación correspondiente al máximo residuo estudentizado es **atípica** si

$$\left|t_{\max}\right| = \left|\max_{1\le i\le n}(t_i)\right| > t_{n-k-2}^{1-\alpha/2}$$

donde $t_{n-k-2}^{1-\alpha/2}$ es el percentil $(1\text{-}\alpha/2)100\%$ de la ley t de Student con n-k-2 grados de libertad.

- **Ver ejemplo selección de variables**