

# Regresión Lineal

Fundamentos de Análisis de Datos  
DSLAB

Máster en Data Science. ETSII.

Móstoles, Madrid

26 de noviembre de 2021



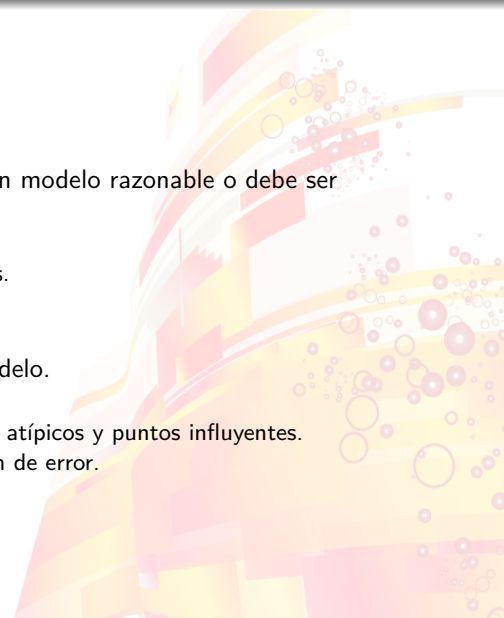
Universidad  
Rey Juan Carlos

El análisis de la regresión es una técnica estadística que se utiliza para investigar y **modelizar la relación entre variables**. Los modelos matemáticos que se aplican a las ciencias experimentales deben incluir un componente de azar, de aleatoriedad, que representa aquello que no alcanzamos a dominar:

$$\text{Observación} = \text{modelo} + \text{error}$$

En un modelo de regresión lineal múltiple, en el que la variable respuesta  $y$  está en función de varios regresores,  $x_1, x_2, \dots, x_k$ , la ecuación fundamental es

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

- 
- 1 Estimación de parámetros.
  - 2 Adecuación del modelo. ¿Es un modelo razonable o debe ser modificado?
    - Selección de variables.
    - Transformación de variables.
    - Análisis de coeficientes.
    - Multicolinealidad.
  - 3 Evaluación y diagnosis del modelo.
    - Análisis de residuos.
    - Detección y tratamiento de atípicos y puntos influyentes.
    - Falta de ajuste y estimación de error.

# Regresión lineal simple

Supóngase que  $y$  es una variable aleatoria y que  $x$  es una variable controlada por el analista. El modelo de la regresión lineal simple es

$$y = \beta_0 + \beta_1 x + \epsilon$$

donde los parámetros  $\beta_0$  y  $\beta_1$  son los coeficientes de regresión. El parámetro  $\beta_0$  es la ordenada en el origen (*intercept*) y  $\beta_1$  es la pendiente (*slope*) de la recta de regresión y representa el cambio en la media de la distribución de  $y$  producido por un cambio unitario en  $x$ . Finalmente,  $\epsilon$  es la componente de error aleatorio.

Se obtienen estimaciones de los parámetros buscando minimizar

$$RSS = \sum_{i=0}^n e_i = \sum_{i=0}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

Es decir, se minimiza la suma de los cuadrados de los residuos (RSS, del inglés *residual sum of squares*). Los residuos representan la diferencia entre el valor observado  $y_i$  y el correspondiente valor ajustado por la regresión,  $\hat{y}_i$ .

Las propiedades de los coeficientes de un modelo de regresión ajustado por mínimos cuadrados:

- El estimador  $\beta_1$  puede expresarse como combinación lineal de los  $y_i$ .
- Los estimadores son insesgados
- Teorema de Gauss–Markov. Para el modelo de regresión  $y = \beta_0 + \beta_1 x + \epsilon$ , con  $E[\epsilon] = 0$ ,  $Var(\epsilon) = \sigma^2$  y los errores incorrelados, se tiene que los estimadores  $\beta_0$  y  $\beta_1$  son insesgados y de mínima varianza.

Un modelo de regresión lineal por mínimos cuadrados es adecuado si cumple las siguientes condiciones sobre sus residuos:

- Media cero.
- Varianza constante.
- Son incorrelados entre si.

Un modelo se considera adecuado si, y sólo si cumple las anteriores condiciones. En caso contrario, se deben aplicar las medidas correctivas adecuadas en cada caso.



# Regresión lineal múltiple

El modelo lineal entre una variable aleatoria respuesta  $y$ , y  $k$  variables regresoras (explicativas), viene definido por

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

Los parámetros  $\beta_j$ ,  $j = 1, 2, \dots, k$  son los coeficientes de regresión. El parámetro  $\beta_j$  representa el cambio esperado en la respuesta y por unidad de cambio en  $x_j$  cuando el resto de las variables regresoras permanece constante.

Utilizando notación matricial, la regresión lineal múltiple se escribe

$$y = X\beta + \epsilon$$

donde

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix},$$

además,  $\text{rango}(X) = k + 1$ , número de parámetros de la regresión.

Se trata de encontrar el vector  $\hat{\beta}$ , que minimice

$$RSS = \sum_{i=0}^n e_i = \epsilon' \epsilon = (y - X\beta)'(y - X\beta)$$

Teniendo en cuenta que  $\left. \frac{\partial RSS}{\partial \beta} \right|_{\hat{\beta}} = 0$ , obtenemos las ecuaciones normales

$$X'X\hat{\beta} = X'y$$

De manera directa, las estimaciones de los coeficientes son

$$\hat{\beta} = (X'X)^{-1}X'y$$

Una vez estimados los parámetros del modelo es conveniente preguntarse si el modelo es globalmente apropiado para nuestros datos y si todos los regresores son importantes en el modelo.

Para responder a estas preguntas se realizan test de hipótesis que requieren que los **errores sean independientes y sigan una distribución normal de media cero y varianza constante**. Esto es,  $\epsilon \sim N(0, \sigma^2 I)$  e independientes, en cuyo caso la variable respuesta es también normal  $y \sim N(X\beta, \sigma^2 I)$  lo que permite utilizar las distribuciones asociadas a los parámetros.

Se trata de determinar si existe relación lineal entre la respuesta  $y$  y cualquiera de las variables regresoras. Puede considerarse como un test para estudiar la adecuación global del modelo. Las hipótesis son

$$H_0 : \beta_1 = \dots = \beta_k = 0,$$

frente a,

$$H_1 : \beta_j \neq 0, \text{ para al menos un } j$$

Rechazar la hipótesis nula implica que al menos uno de los regresores contribuye de manera significativa al modelo. Se rechazará  $H_0$  si  $F_0 > F_{\alpha, k, n-k-1}$ .

Una vez determinado que al menos uno de los coeficientes de regresión es importante (significativo), lo siguiente es **determinar cuáles son los importantes**.

Al incluir un nuevo regresor la varianza del valor ajustado  $\hat{y}$  aumenta, con lo que se ha de tener cuidado para incluir sólo regresores que sean realmente importantes a la hora de explicar la respuesta. Por otro lado, incluir regresores adicionales que no ayudan explicar la respuesta puede hacer que aumente la media de cuadrados de la regresión, con lo que disminuye la utilidad del modelo.

El contraste de hipótesis para la significación de un coeficiente de regresión individual,  $\beta_j$  es

$$H_0 : \beta_j = 0,$$

frente a,

$$H_1 : \beta_j \neq 0,$$

Si no se rechaza  $H_0$ , podemos eliminar el regresor  $x_j$  del modelo. Sin embargo, es preciso actuar con cuidado ya que se trata de un contraste parcial (test marginal) puesto que el coeficiente  $\hat{\beta}_j$  depende de todas las otras variables regresoras  $x_i$ ,  $i \neq j$ . La hipótesis nula será rechazada si  $|t_0| > t_{\alpha/2, n-k-1}$ .



Otra forma de evaluar la adecuación global de un modelo es el coeficiente de determinación

$$R^2 = \frac{\sum_{i=0}^n e_i}{\sum_{i=0}^n (y_i - \hat{y}_i)}$$

que puede verse como la cantidad de varianza explicada. Hay que tener en cuenta que cuando se incluye un regresor nuevo en el modelo, el  $R^2$  aumenta sin importar la cuantía de la aportación del regresor.

Por ello, habitualmente se utiliza

$$R_{adj}^2 = 1 - (1 - R^2) \frac{n - 1}{n - k - 1}$$

que penaliza el número de regresores incluidos en el modelo. El  $R_{adj}^2$  puede ser negativo y su valor siempre será menor o igual que el  $R^2$ .

Habitualmente no es fácil comparar coeficientes de regresión porque la magnitud de  $\hat{\beta}_j$  refleja las unidades de medida del regresor  $x_j$ . En general, las unidades del coeficiente de regresión  $\beta_j$  son (unidades de  $y$ )/(unidades de  $x_j$ ). Por ello es de gran ayuda trabajar con **regresores estandarizados** que producen **coeficientes de regresión adimensionales**, denominados coeficientes de regresión estandarizados. Para calcularlos se utilizan básicamente dos técnicas:

- Escala Normal Unidad:  $z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$
- Escala Longitud Unidad:  $z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{S_{ij}}}$

Las estimaciones de los coeficientes de regresión por ambos métodos es idéntica y producen el mismo conjunto de coeficientes de regresión adimensionales.

Los errores en el signo esperado de los coeficientes de regresión se suele asociar a uno de los siguientes casos:

- El **rango** de alguno de los regresores es demasiado pequeño. Es decir, si los niveles elegidos para  $x$  quedan muy juntos, la varianza de  $\hat{\beta}_j$  será relativamente grande y, en algunos casos, puede suceder que el estimador del coeficiente de regresión resulte negativo.
- Faltan **regresores importantes** en el modelo. En estos casos, el signo no es realmente el equivocado. Ello se debe a la naturaleza parcial de los coeficientes de regresión. Miden el efecto de la variable dado que los otros regresores están en el modelo.
- Presencia de **multicolinealidad**. Incrementa la varianza de los coeficientes de regresión y ello puede provocar que uno o más coeficientes de regresión tengan el signo equivocado
- Se han producido **errores de cálculo**.

# Diagnosis del modelo

Las principales hipótesis que hemos realizado en el estudio del modelo de regresión son

- 1 La relación entre la variable respuesta  $y$  y los regresores  $x_j$  es lineal, al menos aproximadamente.
- 2 Los errores tienen media cero,  $E[\epsilon] = 0$ .
- 3 Los errores tienen varianza constante,  $Var(\epsilon) = \sigma^2$ .
- 4 Los errores son incorrelados.
- 5 Los errores siguen una distribución normal.

Es preciso constatar la validez de estas suposiciones, ya que si alguna de ellas no se verifica podría conducir a un modelo inestable. Para detectar violaciones de estas hipótesis nos apoyaremos, en primer lugar, en el análisis de los residuos.

Los residuos se obtienen como diferencia entre los valores observados y los correspondientes valores ajustados

$$e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$$

Así mismo es una **medida de la variabilidad de la variable respuesta** no explicada por el modelo de regresión. También es conveniente pensar en ellos como en observaciones del modelo de errores.

Los residuos tienen propiedades importantes.

- 1 La media de los residuos es cero:  $\bar{e} = \frac{1}{n} \sum_{i=0}^n e_i = 0$
- 2 Una estimación aproximada de la varianza es  $MS_{Res} = \frac{1}{n-k-1} \sum_{i=0}^n e_i^2$ , que se conoce como **error cuadrático medio**, y tiene  $n - k - 1$  grados de libertad puesto que los  $n$  residuos no son independientes.

Estos métodos son útiles para encontrar valores extremos (outliers)

- **Residuos Estandarizados:**  $d_i = \frac{e_i}{\sqrt{MS_{Res}}}$ . Tienen media 0 y varianza aproximadamente 1.
- **Residuos Estudentizados:**  $r_i = \frac{e_i}{\sqrt{(1-h_{ii})MS_{Res}}}$ . Residuos divididos por la desviación típica exacta del  $i$ -ésimo residuo. Un punto con valores altos de  $e_i$  y de  $h_{ii}$  tiene potencialmente influencia alta en el ajuste por mínimos cuadrados.
- **Residuos PRESS:**  $e_{(i)} = \frac{e_i}{1-h_{ii}}$ . Son los utilizados en el cálculo de *Prediction Error Sum of Squares*. Son los residuos al estimar el modelo de regresión eliminando la observación donde medimos el residuo. Si  $y_i$  es realmente una observación extrema, el modelo de regresión que utiliza todas las observaciones puede estar demasiado influenciado por esta observación. Esto podría producir que  $\hat{y}_i$  sea muy similar a  $y_i$  y que el residuo ordinario  $e_i$  fuera pequeño.

El análisis gráfico de los residuos es un camino muy efectivo para investigar la adecuación del ajuste del modelo de regresión y chequear las hipótesis de partida.

- **Gráfico de Probabilidad Normal:** Es un gráfico diseñado de tal modo que la función de distribución de la normal se dibuja como una línea recta. El ajuste de los residuos a esta recta nos informará de si se cumple o no la hipótesis de normalidad.
- **Gráfico de Residuos frente a los valores ajustados  $\hat{y}_i$ :** Si los residuos de este gráfico se distribuyen más o menos uniformemente sobre una banda horizontal, puede decirse que el modelo no tiene defectos obvios.



- **Gráfico de Residuos frente al Regresor:** Su interpretación es similar a la del caso anterior. Su utilidad, por tanto, también debe ajustarse a las mismas condiciones ya expuestas. Puede ser de ayuda incluir gráficas de residuos frente a regresores no incluidos inicialmente en el modelo.
- **Gráfico de los Residuos respecto al Índice:** Este gráfico puede indicar que los errores en un periodo de tiempo están correlados con otros en periodos diferentes. Este fenómeno de la correlación entre errores en diferentes periodos de tiempo se denomina autocorrelación, que es una violación seria de las suposiciones básicas de la regresión.

Una limitación del gráfico de residuos frente al regresor es que tal vez no muestre completamente el efecto marginal correcto (o completo) de un regresor, dados los otros regresores en el modelo.

Un gráfico de regresión parcial es una variante del gráfico de residuos frente a la predicción  $\hat{y}_i$  y representa una mejora en el estudio de la relación marginal de un regresor dada la presencia de las demás variables en el modelo.

En este gráfico se representan los residuos de la regresión entre  $y$  y  $x_i$  por un lado (abscisas) y entre  $x_i$  y  $x_j$  por otro (ordenadas) si es que se está estudiando la relación parcial entre  $y$  y  $x_i$ . Si, por ejemplo, el regresor  $x_i$  **entra linealmente en el modelo, el gráfico de regresión parcial mostraría una relación lineal.**

Los residuos que son considerablemente más grandes que otros, señalan observaciones extremas en potencia. Se trata de **observaciones que no concuerdan con el resto de los datos** y pueden tener efectos en el modelo de regresión.

Los datos atípicos deben estudiarse con cuidado para determinar su procedencia. En ocasiones, se trata de valores producto de errores. En otras ocasiones se encuentra que la observación atípica es inusual pero perfectamente asumible en el contexto de los datos.

El efecto de las observaciones extremas en el modelo puede chequearse eliminando estos puntos y reajustando la ecuación de regresión. Podemos estudiar la sensibilidad a estos puntos de los coeficientes de regresión o de los estadísticos asociados.

En ocasiones encontramos que una observación o un pequeño subconjunto de ellas **ejerce una influencia grande** (desproporcionada) en el ajuste del modelo de regresión: en la estimación de los coeficientes y en sus propiedades.

Si estos puntos influyentes son calificados como “malos”, deben ser eliminados de la muestra. Si, por el contrario, nada extraño ocurre con estos puntos, su estudio puede darnos algunas propiedades clave del modelo.

La matriz

$$H = X(X'X)^{-1}X'$$

juega un papel importante en identificar las observaciones influyentes. Los elementos de la diagonal principal de la matriz  $H$ , se pueden escribir

$$h_{ii} = x_i'(X'X)^{-1}x_i$$

son una medida estandarizada de la distancia de la  $i$ -ésima observación al centro (centroide) del conjunto de datos. Cualquier observación que verifique

$$h_{ii} > 2\frac{k+1}{n}$$

está suficientemente alejada del resto de los datos para ser considerada una **observación de alto nivel** (*leverage point*).

La distancia de Cook mide **la influencia de la  $i$ -ésima observación si ésta es eliminada de la muestra**. Se define como

$$D_i = \frac{r_i^2}{k+1} \frac{h_{ii}}{1-h_{ii}}$$

donde el primer factor refleja el ajuste del modelo a la  $i$ -ésima observación y el segundo factor mide lo lejos que está ese punto del resto de los datos. Cada uno de los factores, o los dos, pueden contribuir a que el valor de  $D_i$  sea grande.

Es un estadístico que mide, en desviaciones típicas, el cambio en los coeficientes de regresión  $\hat{\beta}_j$  si se elimina la  $i$ -ésima observación. Se calcula de la siguiente manera

$$Dfbetas_{j,i} = \frac{r_{j,i}}{\sqrt{r_j r_i}} \frac{t_i}{\sqrt{1-h_{ii}}}$$

donde  $t_i$  es el residuo R-student,  $r'_j$  representa la fila  $j$ -ésima de la matriz  $R = (X'X)^{-1}X'$ . De modo que  $Dfbetas$  mide tanto el nivel como el efecto de un residuo grande. Se considera que si

$$|Dfbetas_{j,i}| > \frac{2}{\sqrt{n}}$$

entonces la **observación  $i$  es influyente en la estimación de  $\beta_j$** .

Es el número de desviaciones típicas que el valor ajustado  $\hat{y}_{(i)}$  cambia si se elimina la observación  $i$ . Se calcula de la siguiente manera

$$Dffits_i = \sqrt{\frac{h_{ii}}{1-h_{ii}}} t_i$$

donde  $t_i$  es el residuo R-student. **Tanto si una observación es extrema como si tiene un alto nivel**, el Dffits puede ser grande. Se considera que si

$$|Dffits_i| > 2\sqrt{\frac{k+1}{n}}$$

entonces la observación  $i$  puede ser un punto de influencia notable.



Para expresar el papel de la  $i$ -ésima observación en la precisión de la estimación, se define

$$COVRATIO_i = \frac{|(X'_{(i)}X_{(i)})^{-1}S_{(i)}^2|}{|(X'X)^{-1}MS_{Res}|}$$

De tal forma que si  $COVRATIO_i > 1$ , **la observación  $i$  mejora la precisión de la estimación**, mientras que si  $COVRATIO_i < 1$ , la inclusión de punto  $i$ -ésimo reduce la precisión.

Para muestras de tamaño grande se recomienda el siguiente criterio: si  $COVRATIO_i > 1 + 3\frac{(k+1)}{n}$  o  $COVRATIO_i < 1 - 3\frac{(k+1)}{n}$ , la observación  $i$ -ésima debería ser considerada influyente.

## Selección de variables y construcción del modelo

En la mayoría de los problemas prácticos se dispone de numerosos candidatos a regresores, candidatos que deberían incluir todos los factores que influyen en la respuesta. Encontrar un subconjunto de regresores apropiado es conocido como el **problema de la selección de variables**.

- nos gustaría que el modelo incluyera el máximo número de regresores para que la información contenida en estos pueda influir en la predicción del valor de  $y$ .
- queremos que el modelo incluya cuantos menos regresores mejor debido a que la varianza de la predicción  $\hat{y}$  crece con el número de regresores empleados en el modelo.

Además, cuantos más regresores haya en el modelo, más caro será recoger los datos y el mantenimiento del mismo. El proceso para encontrar un modelo que sea un compromiso entre los dos objetivos anteriores es en lo que consiste seleccionar la **mejor ecuación de regresión**.

La motivación para la selección de variables puede resumirse de la siguiente manera:

- La eliminación de variables del modelo puede **mejorar la precisión de los estimadores** de los parámetros de las variables que quedan en el mismo. También se mejora la varianza de la predicción.
- Eliminar variables introduce, potencialmente, sesgo en los estimadores de los coeficientes que permanecen en el modelo y en la variable respuesta. Sin embargo, el sesgo introducido es menor que la reducción de la varianza.
- Es peligroso retener variables que no sean significativas en el modelo, es decir variables con coeficientes menores que sus correspondientes errores estándar en el modelo completo, puesto que se incrementa la varianza de los estimadores de los parámetros y de la predicción.

Para encontrar el subconjunto de variables que se utilizará en la ecuación final, es natural considerar modelos ajustados con varias combinaciones de los candidatos a regresores.

- Mejor subconjunto: Si tenemos  $k$  candidatos a regresores, en total hay que estimar y examinar  $2^k$  ecuaciones. El número de ecuaciones a examinar crece rápidamente con  $k$ . Se suele minimizar algún estadístico como el  $AIC$ , el  $BIC$ , el  $R^2_{adj}$  o el  $RMSE$ . Dependiendo de  $k$ , puede ser imposible calcular todas las combinaciones.
- Métodos secuenciales: Se trata de métodos para evaluar sólo un pequeño número de modelos de regresión parciales, añadiendo o eliminando regresores uno de cada vez.

Las variables se eliminan secuencialmente en el modelo.

- Se fija el umbral máximo  $\delta_{out}$ , dependo de un contraste individual (de la  $t$  o de la  $F$ ).
- Se introducen todas las variables en la ecuación.
- Después se van excluyendo una tras otra.
- En cada etapa se elimina la variable menos influyente, siempre que  $\delta_i < \delta_{out}$ .
- Se detiene cuando no haya ningún  $\delta_i < \delta_{out}$ .

Las variables se introducen secuencialmente en el modelo.

- Se fija el umbral mínimo  $\delta_{in}$ , dependiendo de un contraste individual (de la  $t$  o de la  $F$ ).
- La primera variable que se introduce es la más influyente, siempre que  $\delta_i > \delta_{in}$ .
- A continuación se considera la variable independiente con mayor  $\delta_i$  y que no esté en la ecuación.
- El procedimiento termina cuando ya no quedan variables que cumplan el criterio de entrada.

Este método es una combinación de los procedimientos anteriores.

- Se fija el umbral mínimo  $\delta_{in}$  y máximo  $\delta_{out}$ , en función de un contraste individual (de la  $t$  o de la  $F$ ).
- En cada paso se introduce la variable independiente que no se encuentre ya en la ecuación, con menor  $\delta_i$ , siempre que  $\delta_i > \delta_{in}$  (i.e. hacia adelante).
- Se elimina la variable con  $\delta_i$  más alto, tal que,  $\delta_i < \delta_{out}$  (i.e. hacia atrás).
- El método termina cuando ya no hay más variables candidatas a ser incluidas o eliminadas.



# Multicolinealidad

El uso y la interpretación de un modelo de regresión múltiple a menudo depende explícita o implícitamente de los estimadores de los coeficientes de regresión individuales. Algunos ejemplos de inferencias que se hacen frecuentemente incluyen

- identificar los efectos relativos de las variables regresoras,
- predicción y/o estimación, y
- selección de un conjunto apropiado de variables para el modelo.

Si no existe relación lineal entre los regresores, se dice que éstos son **ortogonales**. Cuando existen dependencias casi lineales entre los regresores, se dice que existe un problema de **multicolinealidad**.

Existen cuatro fuentes primarias de multicolinealidad:

- El **método empleado en la recogida de datos** puede conducir a problemas de multicolinealidad si se muestrea sólo un subespacio de la región de los regresores.
- Las **restricciones** en el modelo o la población. Por ejemplo, si se está investigando el efecto del ingreso familiar y el tamaño de la casa en el consumo eléctrico, los niveles de las dos variables regresoras quedarán aproximadamente sobre una recta.
- La **especificación del modelo**. Por ejemplo, si el rango de  $x$  es pequeño, añadir un término en  $x^2$  puede resultar en multicolinealidad significativa.
- Un **modelo sobredefinido** tiene más variables regresoras que observaciones. Estos modelos se encuentran a veces en investigaciones médicas y de comportamiento, en las que suele haber un número pequeño de sujetos y se recoge información acerca de un número grande de regresores de cada sujeto.

La presencia de multicolinealidad tiene efectos potencialmente serios sobre los estimadores de mínimos cuadrados de los coeficientes de regresión. Esto es, los estimadores tienen varianzas y covarianzas grandes. **Ello implica que muestras distintas tomadas a los mismos niveles de  $x$  podrían proporcionar estimadores muy distintos de los coeficientes de regresión.**

La multicolinealidad está asociada con valores propios de  $X'X$  serán pequeños y, tiende a producir estimadores pobres y demasiado grandes en valor absoluto.

Ello **no implica necesariamente que el modelo ajustado no sea bueno** para realizar predicciones. Si las predicciones se restringen a regiones del espacio donde el modelo sea estable y, proporciona predicciones satisfactorias.

## Examen de la Matriz de Correlación

Una medida muy simple de multicolinealidad es la inspección de los elementos  $r_{ij}$  por encima de la diagonal principal de la **matriz de correlación**  $X'X$ . Si los regresores  $x_i$  y  $x_j$  están próximos a la dependencia lineal, entonces  $|r_{ij}|$  estará cerca de la unidad.

Examinar la correlación simple  $r_{ij}$  entre los regresores ayuda a **detectar la dependencia lineal sólo entre pares de regresores**.

## Factores de Inflación de la Varianza

Los elementos de la diagonal de la matriz  $C = (X'X)^{-1}$  son muy útiles para detectar multicolinealidad. Podemos ver a  $C_{jj}$  como el factor que incrementa la varianza de  $\hat{\beta}_j$  si existen dependencias casi-lineales entre los regresores. Se define

$$VIF_j = C_{jj} = \frac{1}{1R_j^2}$$

como el **factor de inflación de la varianza**.

El  $VIF_j$  es próximo a 1 cuando  $R_j^2$  es próximo a 0, es decir cuando  $x_j$  no depende linealmente del resto de las variables regresoras. VIF superiores a 10 implica serios problemas con la multicolinealidad.

## Análisis de Valores Propios de la matriz $X'X$

Las raíces características o valores propios de  $X'X$ ,  $\lambda_1, \dots, \lambda_p$  puede utilizarse para medir el grado de multicolinealidad de los datos. Si existen una o más dependencias casi-lineales en los datos, entonces una o más de **las raíces características serán pequeñas**.

También se puede examinar el **número de condición** de  $X'X$ , definido por

$$k = \frac{\lambda_{\max}}{\lambda_{\min}}$$

Esta es una medida de la extensión de los valores propios de  $X'X$ . En general, si  $k < 100$  no hay serios problemas con la multicolinealidad. Valores de  $k$  entre 100 y 1000 indican de moderada a fuerte colinealidad y valores de  $k$  por encima de 1000 indican severa multicolinealidad.

Entre las diversas técnicas propuestas para tratar los problemas causados por la multicolinealidad se incluyen recoger datos adicionales, especificación del modelo y el uso de métodos de estimación distintos al de los mínimos cuadrados.

- **Recoger Datos Adicionales:** Incluso aunque fuera posible recoger nuevos datos, puede que nos sea apropiado su uso si los nuevos datos amplían el rango de las variables regresoras fuera de la región de interés actual.
- **Reespecificación del Modelo:** Con frecuencia se dan situaciones en que la multicolinealidad ha sido causada por la elección del modelo, como por ejemplo cuando en la ecuación de regresión se han utilizado regresores altamente correlados.
- **Métodos de regularización:** Ridge, Lasso y Elastic Net.



Los métodos de regularización incorporan **penalizaciones en el ajuste por mínimos cuadrados** con el objetivo de evitar el sobreajuste, reducir varianza de los estimadores de los coeficientes, atenuar el efecto de la multicolinealidad y minimizar la influencia de los regresores menos relevantes. Por lo general, aplicando regularización se consiguen modelos más estables y, por tanto, con mayor capacidad de generalización.

Dado que estos métodos actúan sobre la magnitud de los coeficientes del modelo, todos los regresores deben de estar en la misma escala, por esta razón son necesarios **regresores estandarizados**.

El Método Ridge (regularización de Tikhonov), penaliza la suma de los coeficientes elevados al cuadrado. Tiene el efecto de reducir de forma proporcional el valor de todos los coeficientes del modelo pero **sin que estos lleguen a cero**.

Los coeficientes de la regresión Ridge,  $\hat{\beta}_{\lambda}^R$ , minimizan la expresión

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

El grado de penalización está controlado por el hiperparámetro  $\lambda$ . Cuando  $\lambda = 0$ , la penalización es nula y el resultado es equivalente al de un modelo lineal por mínimos cuadrados. A medida que  $\lambda$  aumenta, mayor es la penalización y menor el valor de los predictores.

El Método Lasso (Least Absolute Shrinkage and Selection Operator), es un modelo basado en la regresión lineal múltiple, en el cual los coeficientes asociados a las variables regresoras pueden contraerse hacia cero. De esta manera se trata de un método más para **selección de variables**, imponiendo penalizaciones sobre los coeficientes de regresión.

Los coeficientes de lasso,  $\hat{\beta}_{\lambda}^L$ , minimizan la expresión

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

Seleccionar un buen valor para  $\lambda$  es crítico, dado que la penalización mediante la norma  $l_1$  fuerza a que algunos coeficientes de la regresión sean exactamente cero cuando el parámetro de regularización  $\lambda$  es suficientemente grande.

Este método de regularización combina las dos anteriores, se añade un nuevo parámetro  $\alpha$  que controla el peso de ambas penalizaciones. Si  $\alpha = 0$ , se elimina la componente de la norma  $L_1$  y se aplica Ridge. Si  $\alpha = 1$ , se elimina la componente correspondiente a la norma  $L_2$  y se aplica Lasso.

$$\frac{\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2}{2n} + \lambda \left( \alpha \sum_{j=1}^p |\beta_j| + \frac{1-\alpha}{2} \sum_{j=1}^p \beta_j^2 \right)$$

Una estrategia frecuentemente utilizada es asignarle casi todo el peso a la penalización  $L_1$  ( $\alpha$  muy próximo a 1) para seleccionar regresores y un poco a la  $L_2$  para disminuir la posible presencia de multicolinealidad.



Falta de adecuación del modelo

El ajuste del modelo de regresión lineal tiene varias suposiciones implícitas

- Los errores del modelo tienen media cero, varianza constante y están incorrelados.
- Los errores del modelo tienen una distribución normal (que se hace para contrastar hipótesis y construir intervalos de confianza).
- La forma del modelo, incluyendo la especificación de los regresores, es correcta.

Cuando alguna de las suposiciones no se cumple, es posible, utilizar procedimientos para construir un modelo de regresión apropiado. Uno de estos métodos es el de transformación de los datos.

La suposición de varianza constante es básica en el análisis de regresión. Un caso corriente en el que esta suposición falla es aquel en el que la variable respuesta sigue una distribución de probabilidad cuya varianza está relacionada con la media.

$$\sigma^2 \propto \text{constante}$$

$$\sigma^2 \propto E[y]$$

$$\sigma^2 \propto E[y](1 - E[y])$$

$$\sigma^2 \propto (E[y])^2$$

$$\sigma^2 \propto (E[y])^3$$

$$\sigma^2 \propto (E[y])^4$$

$$y' = y$$

$$y' = \sqrt{y}$$

$$y' = \sin^{-1} \sqrt{y}$$

$$y' = \ln(y)$$

$$y' = y^{-1/2}$$

$$y' = y^{-1}$$

La falta de linealidad suele contrastarse mediante el test de la falta de ajuste, o mediante la matriz de los diagramas de dispersión, o los gráficos de regresión parcial.

## función linealizable

$$y = \beta_0 x^{\beta_1}$$

$$y = \beta_0 e^{\beta_1 x}$$

$$y = \beta_0 + \beta_1 \log(x)$$

$$y = \frac{x}{\beta_0 x - \beta_1}$$

## transformación

$$y' = \log(y), x' = \log(x)$$

$$y' = \ln(y)$$

$$x' = \log(x)$$

$$y' = \frac{1}{y}, x' = \frac{1}{x}$$

## forma lineal

$$y' = \log(\beta_0) + \beta_1 x'$$

$$y' = \ln(\beta_0) + \beta_1 x$$

$$y' = \beta_0 + \beta_1 x'$$

$$y' = \beta_0 - \beta_1 x'$$



Una familia de transformaciones útil para estabilizar la varianza de una serie, que incluye tanto logaritmos como transformaciones de potencia, es la familia de transformaciones Box-Cox, que dependen del parámetro  $\lambda$  y se definen de la siguiente manera

$$w_t = \begin{cases} \log(y_t), & \lambda = 0; \\ (y_t^\lambda - 1)/\lambda, & \lambda \neq 0. \end{cases}$$

- $\lambda = 1$ : (No hay transformación sustancial)
- $\lambda = \frac{1}{2}$ : (Raíz cuadrada más transformación lineal)
- $\lambda = 0$ : (Logaritmo neperiano)
- $\lambda = -1$ : (Inversa más 1)

## Consideraciones finales

- ¿Es la ecuación razonable? Esto es, ¿tienen sentido los regresores que han quedado en el modelo a la vista del entorno del problema?
- ¿Es utilizable el modelo para el propósito que se planificó? Por ejemplo, un modelo planificado para predicción que contenga un regresor que no es observable en el momento de la predicción, no es un modelo utilizable. Si el coste de recoger datos de un regresor es prohibitivo, resultaría que el modelo tampoco es utilizable.
- ¿Son razonables los coeficientes de regresión? Esto es, las magnitudes y los signos de los coeficientes ¿son realistas? y los errores estándar ¿son relativamente pequeños?
- Los chequeos de diagnóstico para la ecuación del modelo ¿son satisfactorios?