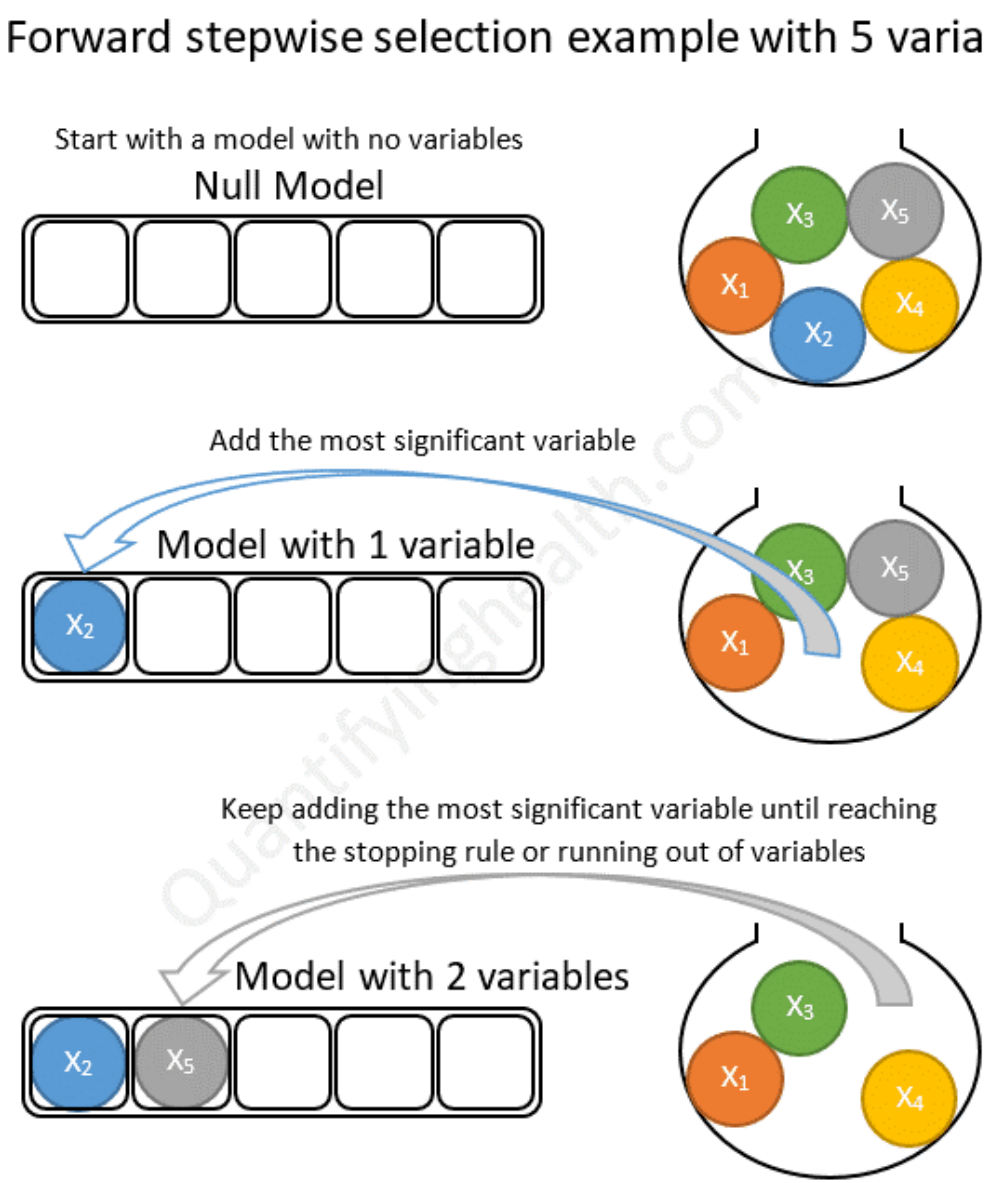


Selección de variables

Óscar Centeno Mora

Preámbulo

- Después de la primera aproximación de la RLM, uno de los momentos cruciales es determinar las variables que deberán entrar a la ecuación de estimación.

- Debemos recordar que un modelo de regresión es ante todo una concepción conceptual para, a través del contraste empírico, tratar de brindar una aproximación al fenómeno de estudio. Por lo que no es introducir variables por tener una cantidad grande de predictores.

- Sin embargo, un modelo debe ser explicado con un óptimo número de variables independientes, y así brindar la mejor calidad y variabilidad explicada.

- En la selección del mejor modelo de una RLM, debemos siempre tener en cuenta el principio de parsimonia.

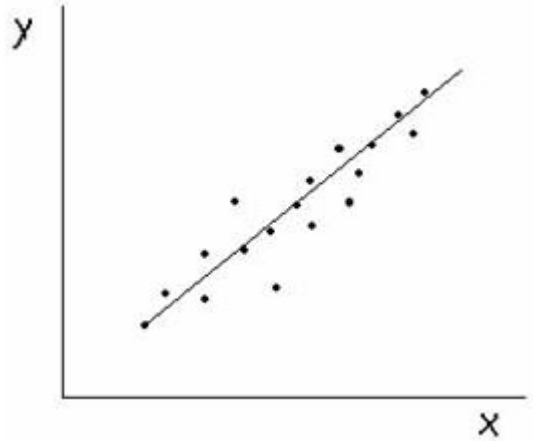
Preámbulo

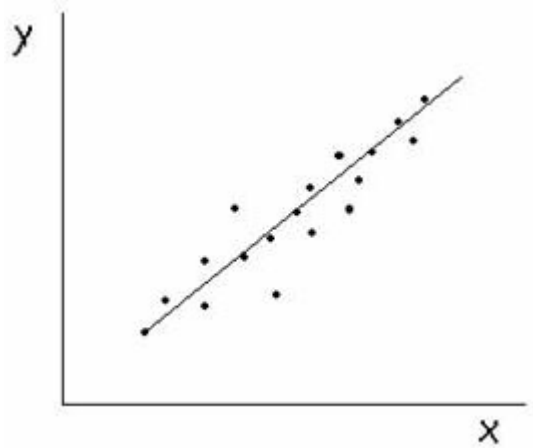
Regresión bivariada

Una variable dependiente (Y)
Una variable independiente (X)

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X1 \\ X2 \\ \cdots \\ X_n \end{bmatrix}$$





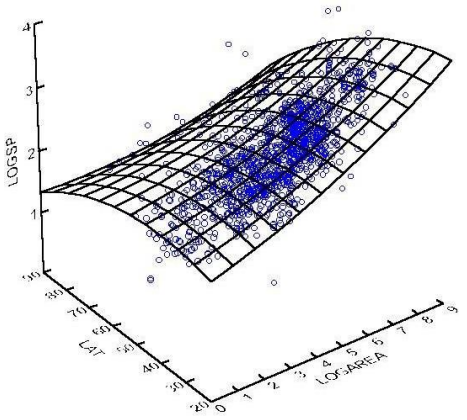
Regresión multivariada

Una variable dependiente
Dos o más variables independientes

$$Y \sim X$$

$$\begin{bmatrix} Y1 \\ Y2 \\ \cdots \\ Y_n \end{bmatrix} \sim \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

¿Cuáles variables?



Preámbulo

El presente capítulo se centra en determinar las herramientas disponibles para determinar los predictores que podrían estar presentes en nuestro modelo de regresión. Veremos algunas herramientas analíticas para poder llegar a una correcta selección de las variables independientes en nuestra regresión.

Índice

Índice

Introducción

- Al iniciar una RLM, lo primero debería ser el análisis descriptivo de todas las variables que concientes la ecuación de mejor. Por TODAS las variables, sí, nos referimos tanto a las variable dependientes como los predictors.
- Luego de un possible análisis descriptive, conocer su distribución, además de posibles valores atípicos, se tendrá un posición más clara de los insumos de trabajo. En este paso se valoran posibles transformaciones en las variables de ambos tipos (dependientes e independientes).
- Vale la pena también aplicar un correlograma, y analizar posibles correlaciones entre las variables independientes. Si existe una alta correlación, es posible que solo se debe introducir una de las variables, o aplicar una regression por componentes o regression de Ridge.
- Siempre el interés estará en brindar un modelo simple, pero a la vez suficientemente explicativo para el contexto de estudio.
- No olvidar la parsimonia.

Índice

El R^2 y el R^2 ajustado

- Habíamos visto en capítulos anteriores el R^2 y el R^2 ajustado

$$R^2 = 1 - \frac{SCE}{SCTot}$$

$$R_a^2 = 1 - \frac{\frac{SCE}{n-p}}{\frac{SCT}{n-1}} = 1 - \frac{CME}{S_Y^2}$$

$$R_a^2 = 1 - \frac{\frac{SCE}{n-p}}{\frac{SCT}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SCE}{SCT} = 1 - \left(\frac{n-1}{n-p}\right)(1-R^2)$$

El R^2 y el R^2 ajustado

- A título MUY personal, creo que el coeficiente de determinación es muy poco útil: solo funciona en modelos con naturaleza lineal, y en comprarción con otros modelos de este tipo.
- Cuando se habla de un R^2 en modelos no lineales, o RLM con otras varaciones, ya este indicador carece de importancia.
- Para datos temporales, el R^2 no posee ningún sentido por la presencia de las raíces unitarias.

Índice

Los métodos backward, forward y stepwise

- Entre los algoritmos para seleccionar variables podemos destacar los siguientes:
  1. Metodos Forward: consiste en la seleccion de variables hacia adelante.
  2. Metodos Backward: se basa en la eliminacion de variables hacia atras.
  3. Metodos Stepwise: Engloban una serie de procedimientos de selección automática de variables signicativas, basados en la inclusion o exclusion de las mismas en el modelo de una manera secuencial.

- La idea de los algoritmos de seleccion de variables es elegir el mejor modelo en forma secuencial pero incluyendo o excluyendo una sola variable predictora en cada paso de acuerdo a ciertos criterios. El proceso secuencial termina cuando se satisface una regla de parada establecida. A continuacion se describen tres de los algoritmos mas usados.

1. Método Forward: (Seleccion hacia adelante). Se parte de un modelo muy sencillo y se van agregando terminos con algún criterio, hasta que no procede añadir ningún termino mas, es decir, en cada etapa se introduce la variable mas signicativa hasta una cierta regla de parada.

2. Método Backward: (Eliminacion hacia atras). Se parte de un modelo muy complejo, que incorpora todos los efectos que pueden inuir en la respuesta, y en cada etapa se elimina la variable menos inuyente, hasta que no procede suprimir ningún termino mas.

3. Método Stepwise: Este procedimiento es una combinacion de los dos anteriores. Comienza como el de introduccion progresiva, pero en cada etapa se plantea si todas las variables introducidas deben de permanecer en el modelo.

Los métodos backward, forward y stepwise

- Cuando se aplica este tipo de procedimientos tenemos que tener en cuenta cual sera la condicion para suprimir o incluir un termino. Para ello podemos considerar dos criterios: criterios de signicacion del termino y criterios de ajuste global.

1. **Criterios de signicacion**: En un metodo backward se suprimira el termino que resulte menos signicativo, y en un metodo forward se añadirá el termino que al añadirlo al modelo resulte mas signicativo. Un criterio de signicacion puede ser la signicacion de cada coeficiente.

2. **Criterios globales**: En vez de usar la signicacion de cada coeciente, podemos basarnos en un criterio global, una medida global de cada modelo, de modo que tenga en cuenta el ajuste y el exceso de parámetros. Escogeremos el modelo cuya medida global sea mejor. Como criterios destacamos el Criterio de Información de Akaike (AIC) y el Criterio de Informacion de Bayes (BIC). Se trata de buscar un modelo cuyo AIC o BIC sea pequeño, ya que en ese caso habrá una verosimilitud muy grande y pocos parámetros.

Índice

Criterios de información

- También, a criterio muy personal, me parece que los criterios de información son mejores indicadores que el coeficiente de determinación R^2 o la significancia. Los dos principales son el AIC, o criterio de Akaike, y el BIC, o criterio de Bayes.
- Dado un conjunto de modelos candidatos para los datos, el modelo preferido es el **que** tiene el valor mínimo en el **AIC**. Por lo tanto **AIC** no solamente recompensa la bondad de ajuste, sino también incluye una penalidad, **que** es una función creciente del número de parámetros estimados. Su fórmula es:

$$\mathrm{AIC} = -2\log - likelihood \; + \; 2p$$

- En **estadística**, el criterio de información bayesiano (**BIC**) o el más general criterio de Schwarz (SBC también, SBIC) es un criterio para la selección de modelos entre un conjunto finito de modelos. Su fórmula es la siguiente:

$$\mathrm{BIC} = -2\log - likelihood \; + \; p\log(n)$$

Criterios de información

- **¿Cuál es mejor?**

- **AIC** es **mejor** para la predicción, ya que es asintóticamente equivalente a la validación cruzada. **BIC** es **mejor** para la explicación, ya que permite una estimación consistente del proceso subyacente de generación de datos.
- En el análisis de la información, se suele tomar ambos criteriores para determiner cuál podría ser el major modelo.
- Admás, el criterio es tomar el AIC y BIC más pequeño (incluyendo la presencia de números negativos)

Índice

Estadístico de Mallow

- El estadístico de Mallow es una medida de buena predicción es el error cuadrático medio total (del modelo con p-1 variables) dividido por la variancia del error (del modelo completo):

$$\Gamma_p = \frac{\sum_{i=1}^{n} E\{\hat{Y}_i - \mu_i)^2}{\sigma^2} = \frac{\sum_{i=1}^{n} \left[(E\{Y_i\} - \mu_i)^2 + \sigma^2\{Y_i\}\right]}{\sigma^2}$$

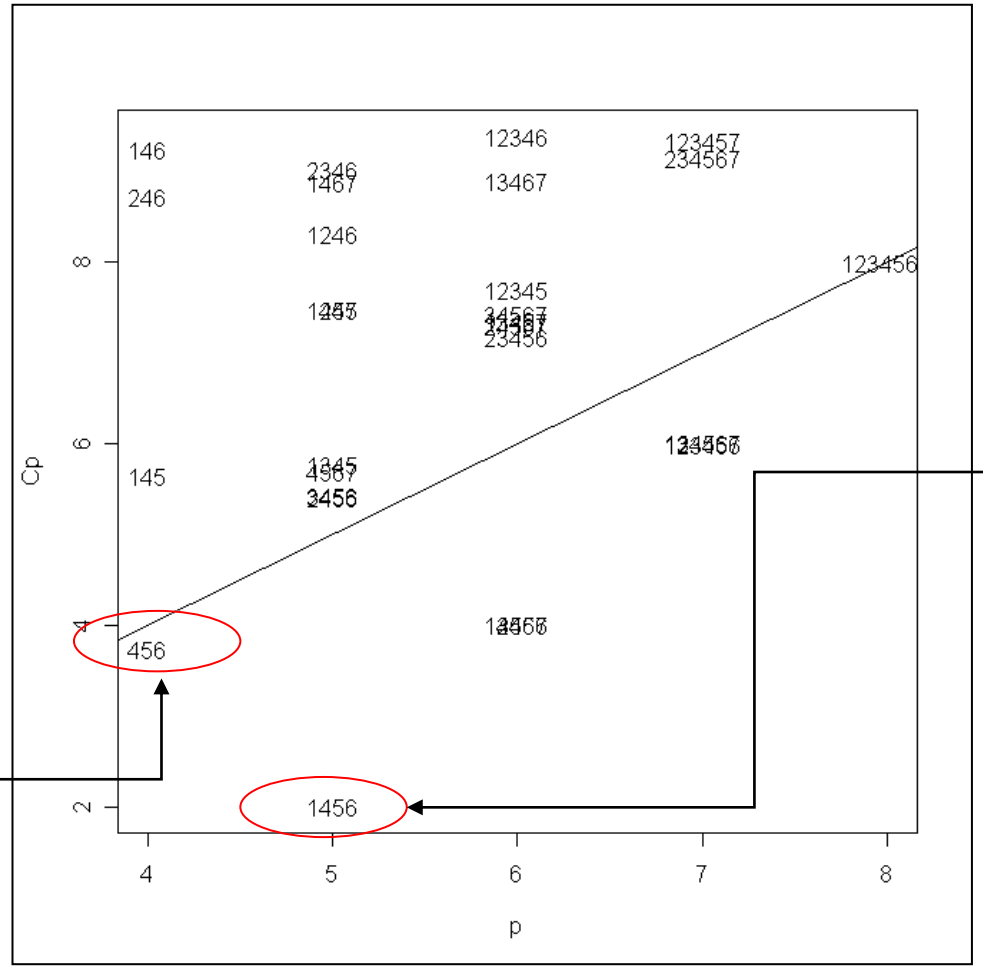
- El estadístico de Mallow C_p es un estimador de Γ_p:

$$C_p = \frac{SC\,\mathrm{Re}\,s_p}{\hat{\sigma}^2} - (n - 2p)$$

Estadístico de Mallow

- Cuando no hay sesgo en el modelo de regresión el valor esperado de C_p es aproximadamente p.
- Cuando se grafican los C_p para todos los posibles modelos contra p, los modelos con poco sesgo tenderán a caer cerca de la línea C_p=p
- Modelos con mucho sesgo tenderán a caer considerablemente encima de esta línea
- Valores de C_p debajo de la línea se interpretan como sin sesgo, quedando debajo de la línea por error de muestreo.

Estadístico de Mallow



Modelo con:

Murder,

HS.Grad y

Frost

Modelo con:

Population,

Murder,

HS.Grad y

Frost

Índice

Otros métodos

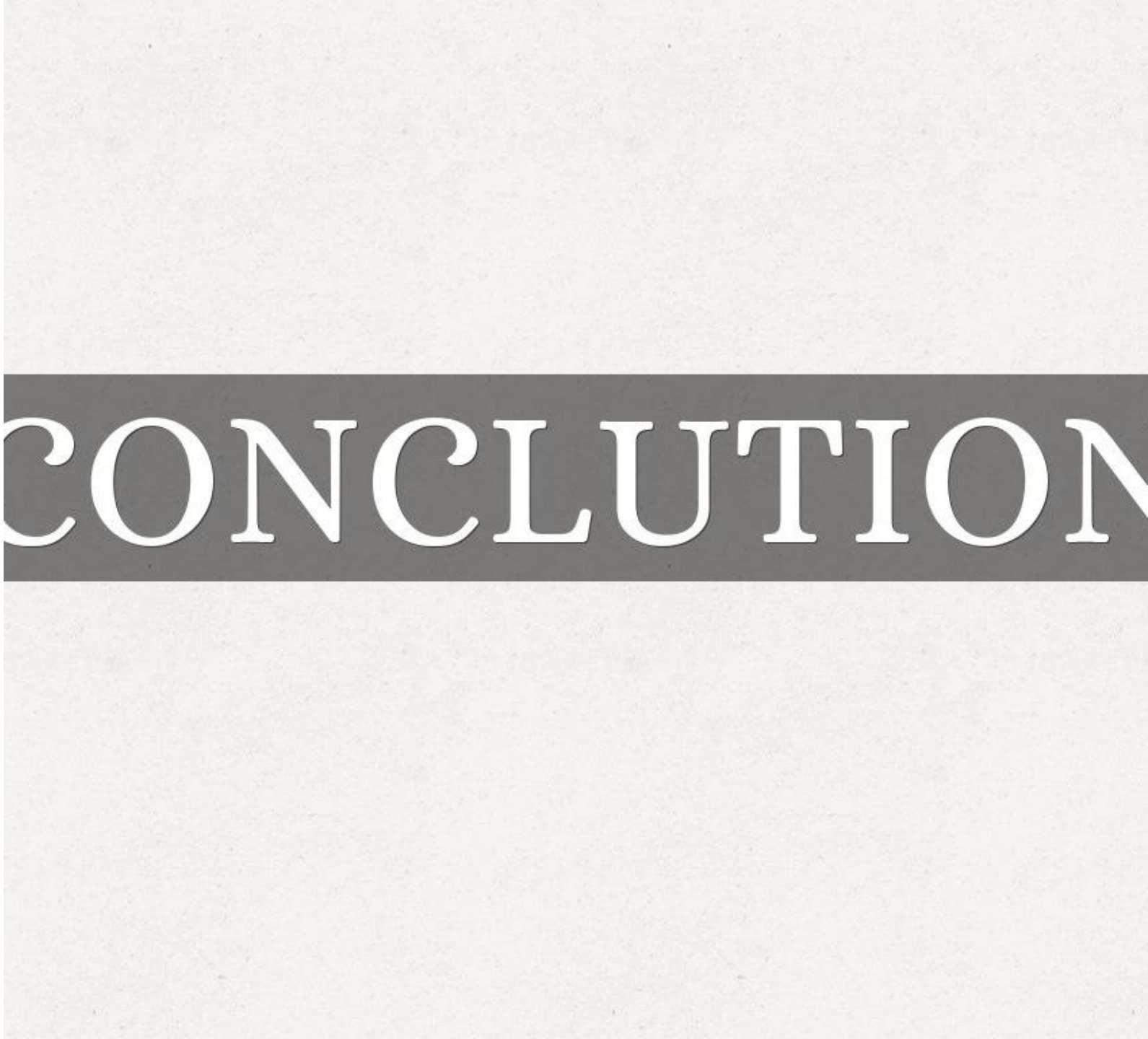
- Dentro de la selección de variables o predictores, existen otros métodos los cuales son:

- Método de mínimos cuadrados penalizados
- Método Lasso
- Método Bridge
- Método SCAD

- No serán cubiertos en este curso, pero aportan un enfoque interesante respecto a cuáles deberían ser las variables que se deberían de incluir.

Conclusión

- El presente capítulo analisó las principals técnicas de selección de variables.
- Antes, es importante un análisis descriptive de los datos.
- Terminada la selección de variables, sería importante ver si el modelo estimado cumple o no con las condiciones.
- Eso será visto en el siguiente capítulo.




The End