Instituto Cántabro de ESTADÍSTICA

Estadística con R

Selección de Modelo de Regresión Lineal

Regresión Lineal Múltiple: forma matricial del modelo MCO

$$Y = X \cdot \beta + e = \beta_1 X_{1t} + \beta_2 X_{2t} + \ldots + \beta_k X_{kt} + e_t$$

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \ldots \\ Y_n \end{pmatrix} \qquad X = \begin{pmatrix} X_{11} & X_{12} & \ldots & X_{1k} \\ X_{21} & X_{22} & \ldots & X_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ X_{n1} & X_{n2} & \ldots & X_{nk} \end{pmatrix} = \begin{bmatrix} \mathrm{X}_1 & \mathrm{X}_2 & \ldots & \mathrm{X}_k \end{bmatrix}$$

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_K \end{pmatrix} \qquad e = \begin{pmatrix} e_1 \\ e_2 \\ \ldots \\ e_n \end{pmatrix}$$

Solución matricial MCO

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

$$X'X = \begin{pmatrix} \sum_{i=1}^{n} X_{i1}^2 & \sum_{i=1}^{n} X_{i1}X_{i2} & \dots & \sum_{i=1}^{n} X_{i1}X_{ik} \\ \sum_{i=1}^{n} X_{i2}X_{i1} & \sum_{i=1}^{n} X_{i2}^2 & \dots & \sum_{i=1}^{n} X_{i2}X_{ik} \\ \dots\dots & \dots\dots & \dots & \dots\dots \\ \sum_{i=1}^{n} X_{ik}X_{i1} & \sum_{i=1}^{n} X_{ik}X_{i2} & \dots & \sum_{i=1}^{n} X_{ik}^2 \end{pmatrix} \qquad X`Y = \begin{pmatrix} \sum_{i=1}^{n} X_{i1}Y_i \\ \sum_{i=1}^{n} X_{i2}Y_i \\ \dots\dots \\ \sum_{i=1}^{n} X_{ik}Y_i \end{pmatrix}$$

Problema de las estimaciones del modelo lineal múltiple

- Inclusión de una variable innecesaria
- Omisión de una variable relevante
- Multicolinealidad:
  - Matriz (X'X) no invertible porque su determinante es cero ó próximo a cero
  - Ocurre por que existe alguna combinación lineal entre las variables dependientes (X_k)
- Si tenemos un conjunto elevado de explicativas debemos seleccionar de entre todas un subconjunto que garanticen que el modelo esté lo mejor especificado posible.

Procedimientos de selección automática de modelos

- Método backward: se comienza por considerar incluidas en el modelo teórico a todas las variables disponibles y se van eliminando del modelo de una en una según su capacidad explicativa. En concreto, la primera variable que se elimina es aquella que presenta un menor coeficiente de correlación parcial con la variable dependiente-o lo que es equivalente, un menor valor del estadístico t– y así sucesivamente hasta llegar a una situación en la que la eliminación de una variable más suponga un descenso demasiado acusado en el coeficiente de determinación.
- Método forward: se comienza por un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas a la que presente un mayor coeficiente de correlación -en valor absoluto- con la variable dependiente. En los pasos sucesivos se va incorporando al modelo aquella variable que presenta un mayor coeficiente de correlación parcial con la variable dependiente dadas las independientes ya incluidas en el modelo. El procedimiento se detiene cuando el incremento en el coeficiente de determinación debido a la inclusión de una nueva variable explicativa en el modelo ya no es importante.
- Método stepwise: es uno de los más empleados y consiste en una combinación de los dos anteriores. En el primer paso se procede como en el método forward pero a diferencia de éste, en el que cuando una variable entra en el modelo ya no vuelve a salir, en el procedimiento stepwise es posible que la inclusión de una nueva variable haga que otra que ya estaba en el modelo resulte redundante.

Instituto Cántabro de ESTADÍSTICA

leaps

- data("mtcars")
- str(mtcars)
- library(leaps)
  regfit.fwd=regsubsets(mpg~.,data=mtcars,method="forward")
  plot(regfit.fwd)
- summary(regfit.fwd)
- regfit.bwd=regsubsets(mpg~.,data=mtcars,method="backward")
  plot(regfit.bwd)
- summary(regfit.bwd)
- regfit.exh=regsubsets(mpg~.,data=mtcars,method="exhaustive")
  plot(regfit.exh)
- summary(regfit.exh)
- coef(regfit.exh,8)
- data(swiss)