Regresión Lineal Múltiple: Selección de Modelos II

Clase 20

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2019-20

Outline

Motivación

Para este punto, ya podemos estimar nuestro modelo de regresión y verificar si este es globalmente significativo para explicar el comportamiento de nuestra variable Y.

Del mismo modo, ya podemos verificar si las variables son individualmente significativas. Dependiendo de los resultados, la intuición nos dice que podemos dejar unas variables y quitar otras.

La pregunta entonces, ¿cuál modelo es preferible?

Motivación

La prueba de significancia individual está condicionada al grupo actual de variables que se utiliza. Por ejemplo en el modelo,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

La prueba de significancia individual para X_1 (i.e $H_0 : \beta_1 = 0, H_1 : \beta_1 \neq 0$) está probando si el aporte de X_1 es significativo, dado que se tienen las variables X_2 y X_3 en el modelo.

Si el grupo de variables cambia, los resultados de la prueba también puede cambiar. Por tanto no es recomendable eliminar todas las variables que no sean significativas en la regresión de forma simultanea.

Motivación

En ese orden de ideas, llegar al "mejor modelo" es algo complicado y no existe una metodología que lo determine.

Es allí donde entra su labor como analistas! A partir de su conocimiento del problema, se deberá determinar el "mejor modelo".

Esta semana veremos herramientas que ayudarán en la búsqueda del "mejor modelo".

Outline

F Parcial

Suponga que se tienen dos modelos, donde uno es una "ampliación" del otro:

- Modelo 1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

- Modelo 2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

El modelo 1 es una ampliación del modelo 2, dado que tiene las mismas variables además de otras dos.

¿La pregunta es que modelo es preferible?

F Parcial

Observe que en el ejemplo en particular, la pregunta de que modelo es preferibles es equivalente a preguntarse sobre que tan relevante es agregar las variables X_3 y X_4 en el modelo 2. Esto a su vez es igual a tomar el modelo 1 y preguntarse:

$$H_0 : \beta_3 = \beta_4 = 0$$

$$H_1 : \text{Algún } \beta_j \neq 0, j \in \{3, 4\}$$

Esta es una hipotesis conjunta de varios β's, que no se puede desarrollar con lo que hemos visto hasta el momento.

Nota: ¡Esto NO es lo mismo que hacer las pruebas de significancia individual para X_3 y X_4!

F Parcial

Para esto debemos identificar cual es el rol de los dos modelos, el modelo completo y el modelo reducido.

- Modelo completo: Es el modelo mas grande entre los dos, el que tiene todas las variables en este caso (i.e el modelo 1).
- Modelo reducido: Es el modelo mas pequeño, es aquel que al ser ampliado apropiadamente da como resultado el modelo grande (i.e el modelo 2).

Oficialmente, el modelo reducido se obtiene al imponer las restricciones presentes en la hipótesis nula sobre el modelo completo.

F Parcial

La idea de la prueba es simple. Para esto recuerde que la suma de cuadrados de la regresión (SSR) se asocia con la varianza explicada por el modelo, luego:

$$SSR_{completo} - SSR_{reducido}$$

muestra la variación extra que el modelo completo explica, en comparación al modelo reducido.

Si dicha diferencia es "grande", entonces es preferible el modelo completo, pero si dicha diferencia es pequeña, entonces no vale la pena ampliar el modelo reducido.

F Parcial

Para saber si la dierencia es grande o pequeña, debemos incorporar el componente estadístico. Bajo la hipótesis nula, se puede demostrar que:

$$\frac{SSR_{completo} - SSR_{reducido}}{\sigma^2} \sim \chi^2_{glR_{completo} - glR_{reducido}}$$

donde los glR son los respectivos grados de libertad de la regresion.

Como desconocemos a σ^2, debemos estimarlo con el MSE, lo que causa varias modificaciones sobre el estadístico.

F Parcial

De esta forma queda contruida la hipótesis:

$$H_0: \quad \text{Modelo reducido es mejor} \Leftrightarrow \beta_3 = \beta_4 = 0$$
$$H_1: \quad \text{Modelo completo es mejor} \Leftrightarrow \text{Algún } \beta_j \neq 0, j \in \{3, 4\}$$

$$F = \frac{\frac{SSR_{completo} - SSR_{reducido}}{glR_{completo} - glR_{reducido}}}{MSE_{completo}} \sim F_{glR\,completo - glR\,reducido, glE\,completo}$$

$$RH_0 \Leftrightarrow F \geq F_{1-\alpha, glR\,completo - glR\,reducido, glE\,completo}$$

F Parcial

Un detalle de extrema importancia es el hecho de que los modelos deben ser anidados, es decir, el modelo reducido se debe obtener directamente del modelo completo

- Modelo 1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

- Modelo 2

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 X_5 + \epsilon$$

El modelo 2 NO es una reducción del modelo 1, ni viceversa.

Si los modelos NO son anidados, la F parcial NO funciona.

F Parcial

Vimos un caso particular de las pruebas de hipótesis de la F parcial, estas pueden ser mas complejas

Como por ejemplo comparar un modelo con pendientes iguales vs uno con pendientes diferentes

- Modelo 1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- Modelo 2

$$Y = \beta_0 + \beta X_1 + \beta X_2 + \beta X_3 + \epsilon = \beta_0 + \beta(X_1 + X_2 + X_3) + \epsilon$$

F Parcial

El modelo completo es el modelo amplio, el modelo sin restricciones, en este caso seria el modelo 1.

El modelo reducido es el modelo con la restricción impuesta, en este caso seria el modelo 2.

Las hipótesis correspondientes serian:

$$H_0: \quad \text{Modelo reducido es mejor} \Leftrightarrow \beta_1 = \beta_2 = \beta_3 = 0$$
$$H_1: \quad \text{Modelo completo es mejor} \Leftrightarrow \text{Algún par } \beta_j \neq 0, j \in \{1, 2, 3\}$$

En este caso ¿Cómo estimarian el modelo reducido?

Despues de tener ambos modelos, la prueba se lleva a cabo con la F parcial de la misma manera que en el ejemplo anterior

Outline

Medidas de ajuste del modelo

Para el caso en que no se tienen modelos anidados, es imposible realizar la prueba estadística correspondiente.

Por tanto se debe hacer la selección desde un punto de vista puramente predictivo.

Para esto se definen algunas medidas que nos indican la calidad del ajuste del modelo. El modelo que tenga un mejor indicador, se cataloga como el mejor a elegir.

Medidas de ajuste del modelo

La primera medida de ajuste es el R^2 que se define como:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

que muestra la proporción de variación de Y que es explicada por el modelo de regresión. Al ser una proporción su valor esta entre 0 y 1.

Un R^2 mayor indica un mejor modelo, pero no necesariamente. Se puede demostrar que el R^2 siempre aumenta a medida que se agregan mas variables, luego el modelo con todas las variables posibles es el que tendrá el R^2 más alto, pero no necesariamente es el más apropiado.

Por lo tanto, si bien el R^2 mide la calidad del ajuste, no es un buen comparador de modelos.

Medidas de ajuste del modelo

La segunda medida de ajuste es el R^2_{ADJ} que se define como:

$$R^2_{ADJ} = 1 - \frac{MSE}{MST} = 1 - \frac{\frac{SSE}{glE}}{\frac{SST}{glT}}$$

el R^2_{ADJ} es una modificación del R^2para castigar por el número de variables utilizadas en la regresión. Esto se evidencia en los cocientes que tienen involucrados los grados de libertad.

No tiene una interpretación en si mismo y NO es una proporción, de hecho puede tomar valores negativos. El R^2_{ADJ} simplemente es un puntaje comparador: el modelo que tenga un R^2_{ADJ} MAYOR, es el modelo más apropiado.

Nota: Observe que esto equivale a seleccionar el modelo con un menor MSE.

Medidas de ajuste del modelo

La tercera medida de ajuste es el *criterio de información de Akaike (AIC)* que se define como:

$$AIC = -2ln(L) + 2k \propto nlog(SSE) - nlog(n) + 2k$$

donde L es la función de verosimilitud y k es el número de variables consideradas en el modelo.

Uno de los problemas del R^2_{ADJ} es que es una variable aleatoria con mucha varianza mientras que el AIC es más estable.

El AIC No tiene una interpretación en si mismo, simplemente es un puntaje comparador: el modelo que tenga un AIC MENOR, es el modelo más apropiado.

Outline

Estrategias de selección

Si se tienen k variables independientes, en total se pueden conformar $\binom{k}{1}$ modelos de una variable, $\binom{k}{2}$ modelos de dos variables, $\binom{k}{3}$ modelos detres variables, etc. Sumando nos da un total de 2^k modelos posibles que se pueden conformar. ¡Eso es por lo general muy complejo de evaluar!

Existen métodos heurísticos que permiten hallar un buen modelo (Forward, Backward, Stepwise) desde un punto de vista de calidad de predicción. No necesariamente se llegará a un modelo que tenga algún sentido conceptual.

Estos algoritmos NO deben usarse ciegamente, lo mejor es siempre utilizar el concepto de un analista.

Estrategias de selección

Forward Regression

Se inicia con un modelo sin variables y en cada iteración se introduce la variable que genera una mayor disminución del AIC con respecto al modelo actual. El algoritmo termina cuando el AIC ya no pueda reducirse agregando más variables.

Backward Regression

Se inicia con un modelo con todas las variables y en cada iteración se quita la variable que genera una mayor disminución del AIC con respecto al modelo actual. El algoritmo termina cuando el AIC ya no pueda reducirse al quitar más variables

Stepwise Regression

Se inicia con un modelo sin variables y en cada iteración se introduce o remueve la variable que genera una mayor disminución del AIC con respecto al modelo actual. El algoritmo termina cuando el AIC ya no pueda reducirse agregando o quitando más variables.

Outline

Aplicación

Machine Learning Techniques for PM10 Levels Forecast in Bogotá

Publisher: IEEE

4 Author(s) Nicolás Mejía Martínez ; Laura Melissa Montes ; Ivan Mura ; Juan Felipe Franco **View All Authors**

Abstract:
Air quality in Bogotá, Colombia, has become of increasing concern. Especially, the levels of PM_{10} are alarming, because of their relation to health risks. A forecast system for PM_{10} levels is beneficial for developing preventive policies of environmental authorities. This paper proposes different forecasting models of particulate matter obtained with three machine learning techniques. A dataset from 8 air quality monitoring stations including PM_{10} and environmental measurements was constructed. Three selection methods of relevant variables for prediction were assessed: selecting variables with the assistance of an expert group, and using two automatic selection methods. Having three sets of potential variables to use as an input, three different forecasting methods were implemented: logistic regression, classification trees and random forest. Finally, a validation and comparison of results are made, to conclude about the best forecast model to be implemented for the city.

Published in: 2018 ICAI Workshops (ICAIW)

Date of Conference: 1-3 Nov. 2018

INSPEC Accession Number: 18291694

Date Added to IEEE *Xplore*: 03 December 2018

DOI: 10.1109/ICAIW.2018.8554995

Aplicación

Machine Learning Techniques for PM_{10} Levels Forecast in Bogotá

Nicolás Mejía Martínez, Laura Melissa Montes, Ivan Mura, Juan Felipe Franco
Universidad de los Andes, Bogotá, Colombia
{n.mejia10, lm.montes10, i.mura, jffranco}@uniandes.edu.co

***Abstract*—Air quality in Bogotá, Colombia, has become of increasing concern. Especially, the levels of PM_{10} are alarming, because of their relation to health risks. A forecast system for PM_{10} levels is beneficial for developing preventive policies of environmental authorities. This paper proposes different forecasting models of particulate matter obtained with three machine learning techniques. A dataset from 8 air quality monitoring stations including PM_{10} and environmental measurements was constructed. Three selection methods of relevant variables for prediction were assessed: selecting variables with the assistance of an expert group, and using two automatic selection methods. Having three sets of potential variables to use as an input, three different forecasting methods were implemented: logistic regression, classification trees and random forest. Finally, a validation and comparison of results are made, to conclude about the best forecast model to be implemented for the city.**

***Keywords*—Air quality forecast, PM_{10}, predictive models, logistic regression, classification and regression trees, random forest.**

I. Introduction

cantly the costs of preventive policies of air pollution to local governments.

Particulate matter (PM) is in Latin America the pollutant that most affect population [3]. The most dangerous particles are those with a diameter of 10 microns or less (PM_{10}). Therefore, PM_{10} is considered in this work for elaborating a forecast model that can be used to raise preventive alarms. We evaluate a forecasting system of PM_{10} elaborated with three data mining techniques: logistic regression, classification trees (CART) and random forest (RF) for the city of Bogotá. We evaluate the input data required for our prediction models, according to several selection criteria. Then, we implement the techniques for the predictions, according to a proposed methodology. Afterwards, the results are analyzed in order to identify the best forecast method and the strengths and weaknesses of each method. The predictions of the three data mining techniques are finally evaluated to compare the different approaches and draw conclusions on forecast system that could be implemented.

Outline

Ejemplo

Con el fin de determinar la relación entre la calificación de su desempeño laboral (y) y las calificaciones en cuatro exámenes, el departamento de personal de cierta empresa industrial realizó un estudio en el que participaron 12 sujetos. Para esto se estimaron 2 modelos:

$$Modelo_1: \quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$
$$Modelo_2: \quad Y = \beta_0 + \beta_1 X_1$$

Del primero modelo, se obtiene un R^2 de 0.79 y del segundo modelo un R^2 de 0.67.

a. ¿Que modelo es preferible?
b. Si ahora se posee información sobre una nueva variable X_5 y se estima un nuevo modelo $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_5 X_5$ Utilizando el AIC como criterio de selección ¿cuál de los tres modelos seleccionaría? Tenga en cuenta que el R^2 del nuevo modelo es 0.73 y su SST es 1000