Experiencia 4

Camilo Valenzuela Carrasco

Departamento de informática
Universidad Técnica Federico Santa Maria

Estadística Computacional

Regresión Lineal Múltiple

- La regresión lineal es un método matemático que modela la relación entre una variable dependiente Y, y las variables independientes x_i.
- Si sólo se tienen una variable independiente, se tiene una regresión lineal simple.

$$Y = f(x_1) = \alpha x_1 + \beta$$

- Para el caso general tenemos:

$$Y = f(x_1, x_2, \ldots, x_n) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \epsilon = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$$

Regresión Lineal Múltiple

- ϵ_j es el error que se obtiene al tratar predecir y_j dado un vector $\overline{x}$
- Suponemos que los errores ϵ_j se distribuyen normal

$$\epsilon_j \sim \mathcal{N}(0, \sigma^2)$$

Regresión Lineal Múltiple

- Sea el valor aproximado $\hat{y}_j = \sum_{i=1}^{n} \hat{\beta}_i + x_{ij}$ y el valor real es y_j, el error del modelo es $Q = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2$
- Lo que se busca son los valores de β_i que se ajustan más a los datos, o que minimizan el error Q.

$$\hat{\beta} = \min_{\beta} Q(\beta)$$

Regresión Multiple en R

Para hacer una regresión lineal en R se utiliza el comando `lm`

Example

$fit < -lm(Y\ X1 + X2 + X3)$
summary(fit)

Regresión Multiple en R

¿Que es lo que entrega `lm`?

```
Call:
lm(formula = Y ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + sulphates + alcohol, data = data.frame(X,
    Y))

Residuals:
     Min       1Q   Median       3Q      Max
-2.68911 -0.36652 -0.04699  0.45202  2.02498

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.197e+01  2.119e+01   1.036   0.3002
fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357
volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150
residual.sugar        1.633e-02  1.500e-02   1.089   0.2765
chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *
total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density              -1.788e+01  2.163e+01  -0.827   0.4086
pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *
sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Selección de Variables

- Cuando se tienen muchas variables en nuestro modelo, puede ser que algunas de ellas no aporten mucho al modelo, y puedan ser sacadas de éste. El problema radica en encontrar qué variables son las necesarias y cuales de ellas no lo son.
- Para poder encontrar el mejor modelo, se utiliza algún método de selección de variables.
- Veremos 3 Métodos de selcción de variables, todos se realizan por paso y buscan ir mejorando o simplificando, en cada paso el modelo. Estos métodos son: Forward, Backward y Stepwise.

Selección de Variables

- Todas los métodos que veremos se basan en el test de hipótesis:

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0$$

Para ver si se acepta H_0 utilizamos el test de Fisher.
Calculamos el estadístico F, luego buscamos el `p-value` relacionado al estadistico F calculado, y se rechaza H_0 si el `p-value`$< \alpha$ dado.

Backwards

- El primer método que veremos es la eliminación hacia atras (Backwards)
- Comenzamos con el modelo con todas las variables
- Calculamos el `p-value` de cada variable para ver cómo afecta al modelo eliminar alguna.
- Elegimos una variable que es eliminada del modelo, y no puede regresar.
- Para elegir la variable a elminar, utilizaremos la función `drop1`, con el test Fisher (test="F").
- El algoritmo se detiene cuando no hay valores que puedan entrar. (Todos los p-value $< \alpha$).

Forwards

- El segundo método a utilizar es la selección hacia delante (Forward)
- El modelo comienza sin variables independientes.
- Calculamos el `p-value` de cada variable que quiere entrar.
- Elegimos la variable que aporta más y la agregamos al modelo.
- Para elegir la variable a ingresar, utilizaremos la función `add1`, con el test Fisher (test="F").
- El algoritmo se detiene cuando todos los valores que pueden entrar tienen un p-value $> \alpha$

Stepwise

Este algoritmo es la mezcla de los dos anteriores,

- Se comienza con todas las variables
- Comienza eliminando una por una las variables del modelo.
- Cuando no quedan variables por eliminar, trata de ingresar variables al modelo.
- El algoritmo termina cuando no hay variables que puedan ser sacadas o agregadas al modelo, o se cumple una cantidad de pasos máximo.