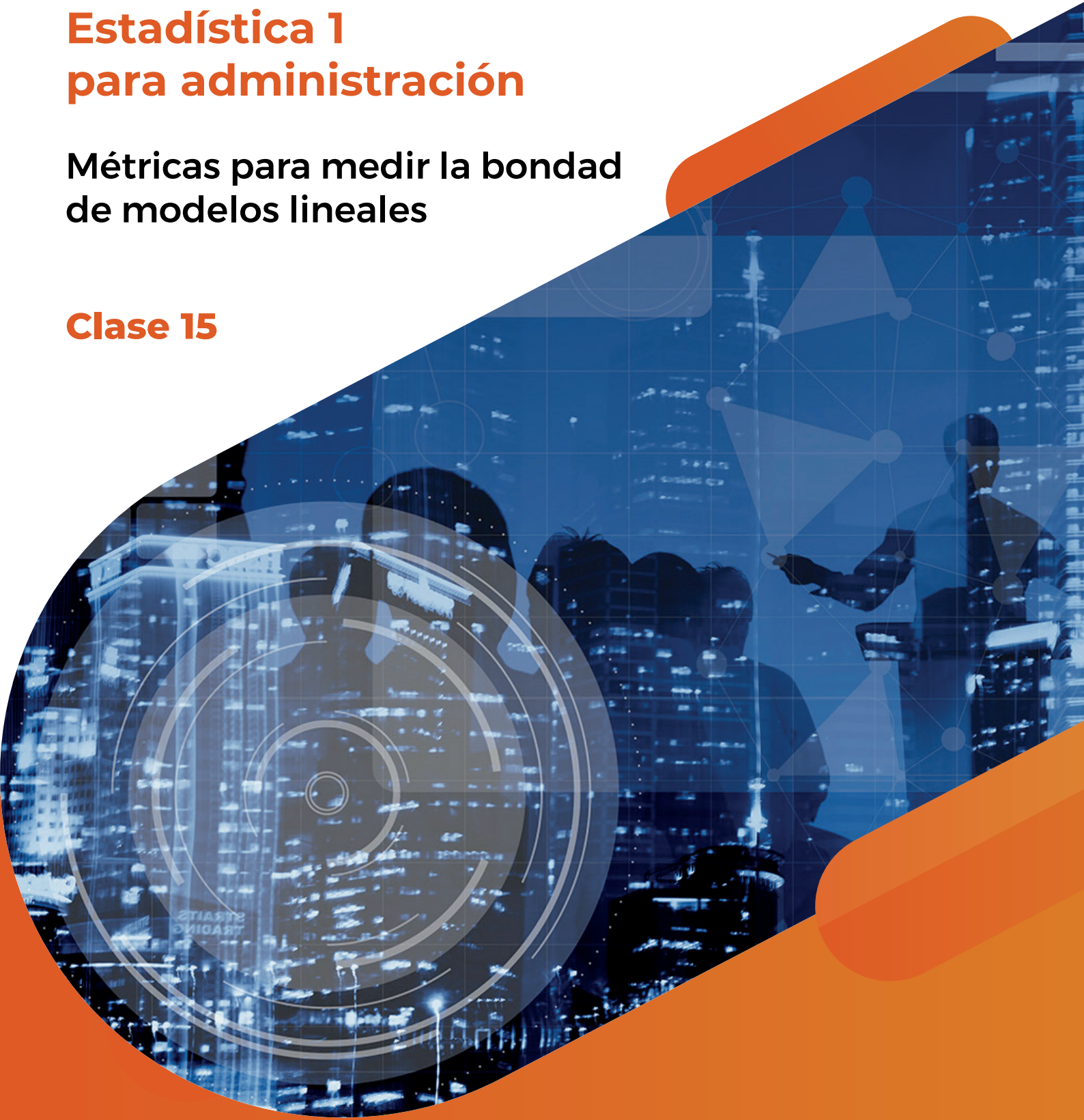


# Estadística 1 para administración

Métricas para medir la bondad  
de modelos lineales

Clase 15



## Desarrollo

### Introducción de la clase

En esta clase abordaremos 2 temas

El primero está relacionado con nuevas métricas para medir la eficiencia en modelos de regresión lineal. En clases anteriores habíamos destacado el uso del estadístico  $R^2$  como una forma de medir la eficiencia basado en los residuos, sin embargo, esta no es la única métrica (ni siquiera la mejor), las métricas que veremos en esta clase se basan en la probabilidad de distribución de los resultados, para lo cual introduciremos una nueva función de regresión.

En la segunda parte veremos cómo en base a estas métricas básicas se han construido otras métricas que toman el nombre de estimadores que tratan de balancear el ajuste del modelo con la complejidad, por lo que nos proveen un análisis más completo, esta sección analizaremos algunos de estos estimadores.

### Resultado o resultados de aprendizaje que será abordado con el contenido de la clase

Presentar informes estadísticos interpretando los resultados dentro del contexto del problema.

## Clase 15: Métricas para medir la bondad de modelos lineales

### $R^2$ y $R^2_a$

De acuerdo a Szretter (2017) “Una vez que se tienen todas las variables, es de interés contar con un criterio numérico para resumir la bondad del ajuste, que un modelo lineal con un cierto subconjunto de covariables da a la variable dependiente observada. A partir de este criterio se podrán ranquear los modelos y elegir un conjunto de unos pocos buenos candidatos para estudiar luego en detalle.”

En general en la regresión podemos establecer 3 valores para cada observación:

$\hat{y}$  = el valor de la predicción de acuerdo al modelo

$\bar{y}$  = el valor medio o esperado para las observaciones

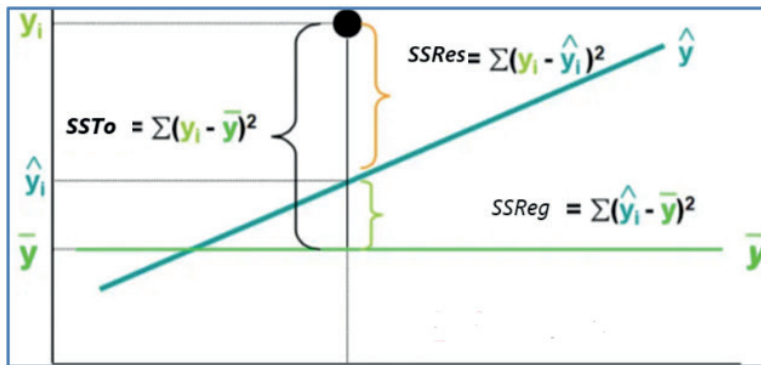
$y$  = el valor de la observación.

La figura 1 clarifica estos conceptos.

Figura 1

Relación entre SST, RSS, SSR

Tomado de Webster A. (2017)



Donde:

**SSTo** = Suma de cuadrados totales, es la suma de los cuadrados de las desviaciones de todas las observaciones,  $y_i$ , de su media.

**SSReg** = Suma de cuadrados de la regresión, le indica cuánto de la variación en la variable dependiente explicó su modelo

**SSRes** = Suma de cuadrados de los residuos, es el valor que no es explicado por el modelo

Cuando tratamos de comparar modelos la primera aproximación sería mediante el  $R^2$  obtenido para cada uno de los distintos modelos con distinta cantidad de covariables. La fórmula del coeficiente de determinación se muestra en la figura siguiente.

Figura 2

Coeficiente de determinación para un modelo con p covariables. Prado A. (2025)

Adaptación de la fórmula de Webster A. (2000)

$$R^2 = 1 - \left( \frac{SSres}{SSTo} \right)$$

En la medida en la que el modelo sea más eficiente, la suma de cuadrado de los residuos va a ser más pequeña, llegando a 0 cuando el modelo es perfecto, en cuyo caso obtendríamos un  $R^2$  sería igual a 1.

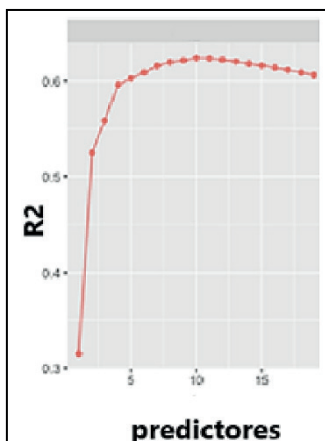
Resulta que comparar modelos usando el criterio de elegir aquél cuyo  $R^2$  sea lo más grande posible equivale a elegir aquel que tenga la menor suma de cuadrados de residuos SSRes (ya que la suma de cuadrados total SSTo no depende de las covariables del modelo ajustado y por eso permanece constante).

Para mejorar el  $R^2$  obtenido los analistas tratan de incluir una mayor cantidad de covariables, sean estas apropiadas para ajustar los datos o no. Es por esto, por lo que el criterio no es identificar el modelo con mayor  $R^2$  (ese será siempre el modelo con todas las covariables disponibles) sino encontrar el punto a partir del cual no tiene sentido agregar más variables ya que estas no inciden en un aumento importante del  $R^2$ . Para encontrar este punto óptimo puede ser útil trazar un gráfico del  $R^2$  en función de la cantidad de predictores. La siguiente figura explica el concepto.

Figura 3

$R^2$  obtenido en función de la cantidad de predictores

Creación del autor Alfonso Prado.



### **R<sup>2</sup> Ajustado**

Para entender esta métrica, es importante indicar los grados de libertad de cada una de las sumas de cuadrados mencionadas anteriormente, esto nos ayudará a calcular la media de las sumas de cuadrados, es una forma de hacer un promedio que pueda ser comparable de una suma de cuadrados a otra. La siguiente tabla presenta los valores.

Tabla 1

Grados de libertad para las sumas de cuadrados

Creación del autor Alfonso Prado

Suma de cuadrados	Grados de libertad
SSreg	p-1
SSres	n-p
SSto	n-1

Como el  $R^2$  no toma en cuenta el número de variables en el modelo de regresión, un criterio de decisión mucho más objetivo es calcular y comparar modelos por medio del  $R^2a$ , para esto partimos de la fórmula de la figura 2 usando la media de las sumas de cuadrados. Entonces notamos que los grados de libertad del denominador no varía con respecto a la cantidad de coeficientes, mientras que el numerador si será afectado.

Por lo tanto, el  $R^2a$  se calculará con la siguiente fórmula

Figura 4

Fórmula de  $R^2a$

Creación del autor Alfonso Prado

$$R^2a = 1 - \left( \frac{SSres/(n-p)}{SSto/(n-1)} \right)$$

Esta sería la fórmula para un modelo sin intercepto donde p representa el número de predictores. Para modelos con intercept los grados de libertad del numerador se calcularía como n-p-1.

De acuerdo a Szretter (2017), entonces buscamos el subconjunto de p – 1 covariables que maximicen el  $R^2a$ , o un subconjunto de muchas menos covariables para las cuales  $R^2a$  el incremento sea tan pequeño que no justifique la inclusión de las covariables adicionales.

Veamos un ejemplo en la predicción de precios de vivienda

#Creamos 5 modelos basado en distintas variables

```

house1 <- lm(valxSqFt ~ TUnits:GSqFt + Boro, data=housing)
house2 <- lm(valxSqFt ~ TUnits+ GSqFt +Boro , data=housing)
house3 <- lm(valxSqFt ~ TUnits* GSqFt +Boro , data=housing)
house4 <- lm(valxSqFt ~ class*Boro , data=housing)

```

#De los cuales obtenemos la siguiente información

```
[1] "house1"
```

```
> summary(house1)$adj.r.squared
```

```
[1] 0.5823141
```

```
> length(house1$coefficients)
```

```
[1] 6
```

```
[1] "house2"
```

```
> summary(house2)$adj.r.squared
```

```
[1] 0.5999892
```

```
> length(house2$coefficients)
```

```
[1] 7
```

```
[1] "house3"
```

```
> summary(house3)$adj.r.squared
```

```
[1] 0.6028864
```

```
> length(house3$coefficients)
```

```
[1] 8
```

```
[1] "house4"
```

```
> summary(house4)$adj.r.squared
```

```
[1] 0.6093757
```

```
> length(house4$coefficients)
```

```
[1] 20
```

Del ejercicio anterior podemos notar que en la medida la cantidad de variables involucradas en cada modelo va aumentando la métrica de  $R^2$  aumenta, pero ya no es proporcional, en el modelo 4 tenemos 20 coeficientes y el incremento del  $R^2$  es insignificante.

## Likelihood

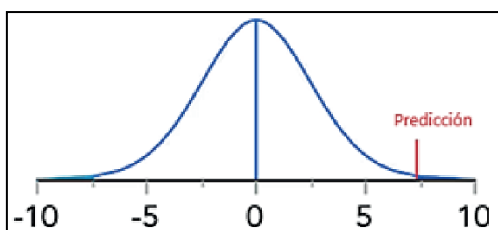
El término likelihood a veces se lo traduce como probabilidad o verosimilitud. En general se refiere a la probabilidad de que una muestra provenga de una distribución dada. Recordando el concepto de la función PDF: Dada una distribución  $\theta$  que tiene una media  $=\mu$  y var  $=\sigma^2$ , la probabilidad de un valor específico es dada por la función de densidad de dicha distribución.

El likelihood se calcula como el producto de densidades de las observaciones. Cuando se hace una predicción se esperaría que el valor predicho caiga dentro de la distribución de la variable  $Y$ , en la medida en la que la probabilidad difiera de la distribución  $\theta$  entendemos que el modelo no es bueno. Por ejemplo, suponga que usted tiene una variable de respuesta  $Y$  que sigue una distribución gaussiana  $\theta$  como en la figura 5, suponga que usted está interesado en predecir cierto valor de la variable basado en determinadas características y al hacerlo encuentra que la salida  $Y$  tendría una probabilidad muy baja, obviamente deducimos que el modelo ha fallado en predecir adecuadamente.

Figura 5

Distribución de la variable  $Y$  y su predicción

Creación del autor Alfonso Prado



Pero, como la probabilidad PDF para un valor dado es generalmente baja, el producto de estas va a ser muy muy bajo, por lo que se estila hacer una transformación logarítmica conocida como el “log-likelihood”. Para esto recordemos una de las propiedades de los logaritmos.

Figura 6

Propiedad de la multiplicación de logaritmos

Creación del autor Alfonso Prado

$$\log_a(b \cdot c) = \log_a b + \log_a c$$

Si se usa logaritmos naturales (base e) obtendríamos las fórmulas del likelihood como mostramos en la siguiente figura, donde  $\pi$  representa la función de producto de las probabilidades.

Figura 7

Fórmulas likelihood como producto (a) y log-likelihood como suma (b). Prado (2025)

Adaptación de la fórmula para la resolución de likelihood Faraway (2014)

(a) $L(\theta) = \prod_{i=1}^n f_i(y_i   \theta)$	(b) $F(\theta) = \sum_{i=1}^n \ln f_i(y_i   \theta)$
--	---

En otras palabras, responde a la pregunta: "¿Cuál es la probabilidad de observar estos datos, dados estos parámetros?"

Estimadores

Existen otros estimadores que nos ayudan a visualizar el mejor modelo basado en el total de las K covariables (el más grande posible) que incluya las covariables importantes de modo que en un modelo completo la estimación de la varianza del error sea insesgada.

**Los principales estimadores son:**

- ✓ AIC : Criterio de Akaike
- ✓ BIC : Criterio Bayesiano de Schwartz
- ✓ CP: Cp de Mallows

Todos intentan resolver este problema introduciendo un término de penalización para el número de parámetros en el modelo.

Desafortunadamente diferentes estimadores pueden recomendar diferentes modelos. Sin embargo, tomados en conjunto estos criterios permiten identificar un conjunto pequeño de modelos de regresión que pueden ser construidos a partir de las variables independientes relevadas. (Szretter, 2017)

## AIC y BIC

Cuando se usa un modelo para representar el proceso que generó los datos, la representación casi nunca será exacta, por lo que se perderá cierta cantidad de información al usar el modelo para representar el proceso.

De acuerdo a Faraway (2020) , AIC y BIC estiman la cantidad relativa de información perdida por un modelo dado, siendo el mejor aquel cuya pérdida de información sea la menor.

Al ajustar modelos, es posible aumentar la probabilidad agregando variables, pero hacerlo puede provocar un sobre ajuste. Los estimadores intentan resolver este problema introduciendo un término de penalización proporcional al número de variables en el modelo.

Entonces dado una familia de modelos  $M_\alpha$  y una cantidad de observaciones, queremos saber cuál es el mejor modelo, donde  $\alpha$  representa el número de predictor(es), ejemplo modelo 1 usa el primer predictor, modelo 2 usa los dos primeros predictores, etc. y donde los posibles predictores se ordenan por correlación con la variable de respuesta.

AIC y BIC parten del likelihood ( $L(\theta)$ ), pero como vemos en la siguiente tabla sus fórmulas son diferentes.

### Tabla 2

Fórmulas matemáticas y funciones programáticas de AIC y BIC

Creación del autor Alfonso Prado

Estimador	Fórmula matemática	Función programática
AIC	$AIC = -2 \log(L) + 2p$	<code>AIC(object, ..., k = 2)</code>
BIC	$BIC = -2 \log(L) + p \log(n)$	<code>BIC(object, ...)</code>

**Donde:**

L=likelihood p=# variables n=# muestras

K= la penalización por parámetro a utilizar; el valor predeterminado k = 2

Estimadores

Explica el uso de AIC y BIC

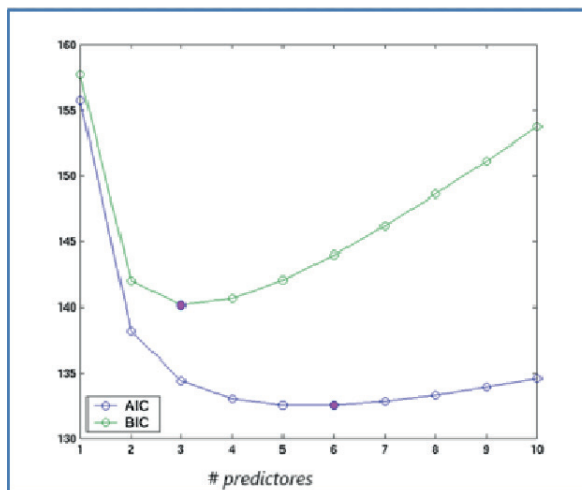
Enlace

Dado una cantidad de modelos con distinta cantidad predictores podemos obtener la siguiente gráfica.

Figura 8

Relación de los estimadores en función de predictores

Creación del autor Alfonso Prado



Continuando con nuestro ejemplo de predicción de precios de vivienda

```
#Creamos 5 modelos basado en distintas variables
housingmod1 <- lm(ValuePerSqFt ~ Units +SqFt +Boro , data=housing )
housingmod2 <- lm(ValuePerSqFt ~ Units * SqFt +Boro , data=housing )
housingmod3 <- lm(ValuePerSqFt ~ Units +SqFt *Boro + Class, data=housing )
housingmod4 <- lm(ValuePerSqFt ~ Units +SqFt +Boro + SqFt*Class , data=housing
)
housingmod5 <- lm(ValuePerSqFt ~ Boro + Class , data=housing )
#Comparamos los modelos con los estimadores
AIC(housingmod1, housingmod2,housingmod3,housingmod4,housingmod5)
BIC (housingmod1, housingmod2,housingmod3,housingmod4,housingmod5)
#Obtenemos los siguientes resultados
```

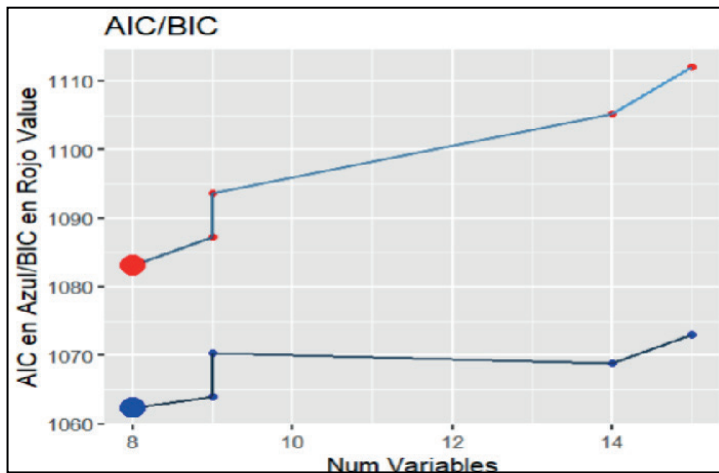
	df	AIC	BIC
<b>housingmod1</b>	8	1062.328	1083.169
<b>housingmod2</b>	9	1063.890	1087.337
<b>housingmod3</b>	15	1072.995	1112.073
<b>housingmod4</b>	14	1068.836	1105.308
<b>housingmod5</b>	9	1070.264	1093.710

```
#Que lo podemos graficar
```

Figura 9

Comparación de 5 modelos con estimadores AIC y BIC

Creación del autor Alfonso Prado



Note que en este caso los 2 estimadores recomiendan el mismo modelo, aunque sus valores son distintos.

De la tabla de los estimadores vale la pena aclarar el termino df. En primer lugar, df significa grados de libertad, que en este caso se traduce a la cantidad de variables predictoras. Si se fija en la creación del modelo 1, este toma 3 variables Units +SqFt +Boro, las 2 primeras son variables continuas y la tercera es una variable categórica que tiene 5 niveles, los cuales han sido automáticamente codificados mediante one-hot, por lo tanto, glm ha creado 5 variables codificadas para la variable Boro. Además, tenemos el intercept del modelo con el cual suma o considera que el modelo tiene 8 variables.

En el modelo 2 tenemos, además de las anteriores, una variable que incluye el efecto combinado de Units:SqFt. Y así sucesivamente con el resto de modelos. Estos grados de libertad vienen a ser la cantidad de variables sobre las cuales se hará la penalización en ambos estimadores.

### Selección automática

De los ejercicios presentados anteriormente entendemos que existen diferentes formas de evaluar la bondad de un modelo, sin embargo, es interesante notar que esto ha requerido una gran cantidad de tiempo de parte del analista, por lo que es importante conocer de mecanismos automáticos que nos permitan ubicar los mejores modelos en una forma oportuna.

**Para esto existen 4 enfoques:**

- ✓ Todos los subconjuntos posibles
- ✓ Eliminación hacia atrás
- ✓ Selección hacia adelante incorporando variables
- ✓ Regresión de a pasos

### **Todos los conjuntos posibles**

Si el dataset tiene  $p$  predictores teóricamente se podría crear  $2^p$  modelos. Mediante este algoritmo crearemos efectivamente esta cantidad de modelos y evaluaremos mediante cualquier métrica descrita anteriormente cuál es el subconjunto óptimo. Por supuesto, esto significa que se realizarán  $2^p$  regresiones, lo cual podría ser problemático para valores de  $p > 10$ .

### **Eliminación hacia atrás**

En este algoritmo, iniciamos con un modelo que contiene todas las covariables, y analizamos los valores  $t$  y  $p$ value para cada coeficiente, si todos los  $p$ values  $< \alpha$ , entonces el modelo completo es el mejor, caso contrario elimina la variable que tenga el mayor  $p$ value, manteniendo las restantes en el modelo. Se evalúan las métricas del modelo. Si el retiro de dicha variable ha provocado un descenso considerable de la métrica entonces el último modelo sería el mejor.

### **Selección hacia adelante**

Es el inverso del anterior, iniciamos con un modelo con una variable y vamos incorporando variables adicionales hasta el punto en el cual las ganancias en el valor de la métrica ya no son significativas. En este caso la métrica que se utiliza es el estadístico  $F$  que nos muestra la relación del modelo con respecto a otro que solo tiene el intercept. Recuerde que este estadístico debe compararse con un  $F_{\text{critico}}$  y cuando esta relación supera el valor  $\alpha$  el proceso se detiene.

### **Paso a paso**

Es una modificación del procedimiento hacia adelante, que elimina una variable en el modelo si ésta pierde significancia cuando se agregan otras variables. El algoritmo es similar al de selección hacia adelante excepto que, a cada paso, después de incorporar una variable, el procedimiento elimina del modelo las variables que ya no tienen significancia.

Es importante mencionar que no siempre el algoritmo hacia adelante y el algoritmo hacia atrás terminarán seleccionando el mismo modelo. Note que la selección se hace en base a AIC.

Veamos un ejemplo de movilidad con mtcars

```
# Algoritmo hacia adelante
data(mtcars)
minimo<- lm(mpg ~1 , data=mtcars) #solo el intercept
summary(minimo)
completo <- c("wt", "cyl", "hp")
step(minimo, direction="forward", scope=formula(completo))

Start: AIC=115.94
mpg ~ 1

Step: AIC=73.22
mpg ~ wt

Step: AIC=63.2
mpg ~ wt + cyl

Step: AIC=62.66
mpg ~ wt + cyl + hp

#Veamos algoritmo hacia atrás
completo <- lm(mpg ~. , data=mtcars)
summary(completo)
step (completo, direction="backward")
```

Start: AIC=70.9

mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb

Step: AIC=68.92

mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb

Step: AIC=66.97

mpg ~ disp + hp + drat + wt + qsec + am + gear + carb

Step: AIC=65.12

mpg ~ disp + hp + drat + wt + qsec + am + gear

Step: AIC=62.16

mpg ~ disp + hp + wt + qsec + am

## Referencias citadas en la Clase 15

Roweida M. Et. Al. (2020) , *Machine Learning with Oversampling and Undersampling Techniques*, Jordan University of Science and Technology

[https://www.researchgate.net/publication/340978368\\_Machine\\_Learning\\_with\\_Oversampling\\_and\\_Undersampling\\_Techniques\\_Overview\\_Study\\_and\\_Experimental\\_Results/download?\\_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoieX2RpcmVjdCJ9fQ](https://www.researchgate.net/publication/340978368_Machine_Learning_with_Oversampling_and_Undersampling_Techniques_Overview_Study_and_Experimental_Results/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoieX2RpcmVjdCJ9fQ)

Faraway J.,(2014), *Linear Models with R*, CHAPMAN & HALL/CRC

Jurafsky D. , . Martin J. (2024) , *Speech and Language Processing*.  
<https://web.stanford.edu/~jurafsky/slp3/>

Hilbe J., (2015) , *Practical Guide to Logistic Regression*, CRC Press

Agresti, A., (2007), *An introduction to categorical data analysis*, John Wiley and Sons publication

Webster A., (2000), *Estadística aplicada a los negocios*, Irwin Professional Publishing

Schmuller J.(2022), *Statistical Analysis with Excel For Dummies*, Wiley

Teck K, (2022) , Practical t-test Power Analysis with R, Practical Assessment, Research & Evaluation , Volumen 27, No 18

Berman H.G., "*F Distribution* , descargado de <https://stattrek.com/probability-distributions/f-distribution>

Rodrigo J. (2016), *Análisis de varianza para comparar múltiples medias*, [https://cienciadedatos.net/documentos/19\\_anova](https://cienciadedatos.net/documentos/19_anova)

Szretter, M. (2017), Apunte de Regresión Lineal, [https://mate.dm.uba.ar/~meszre/apunte\\_regresion\\_lineal\\_szretter.pdf](https://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf)

Estadística Inferencial <https://www.studocu.com/cl/document/universidad-de-valparaiso/estadistica-inferencia>



**La excelencia no se improvisa**

síguenos

