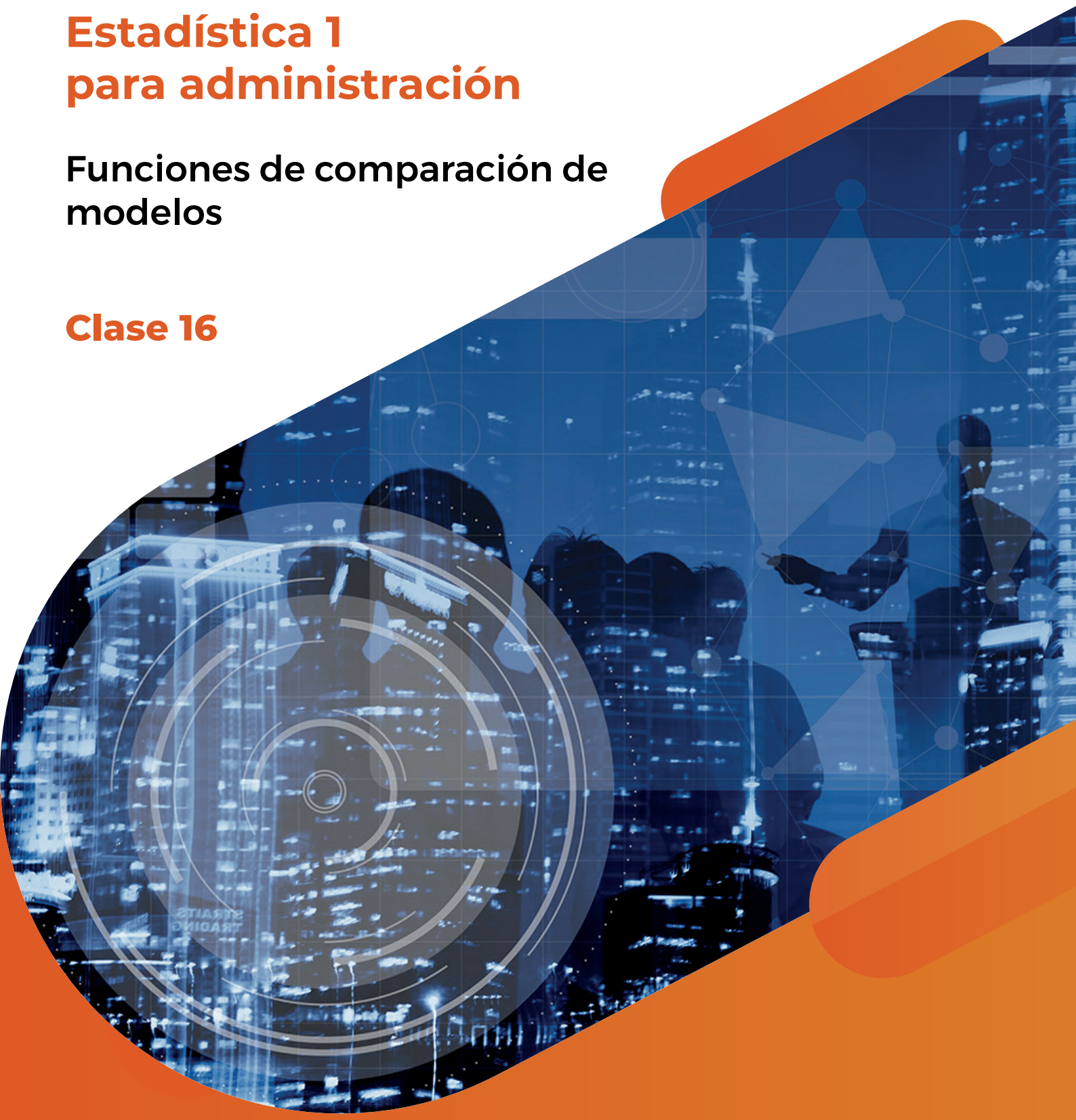
PUCE Virtual
EDUCACIÓN FLEXIBLE

Estadística 1 para administración

Funciones de comparación de modelos

Clase 16



La excelencia no se improvisa

Desarrollo

Introducción de la clase

En esta clase revisaremos dos temas : El primero está relacionado con métricas adicionales para medir y comparar la eficiencia de un modelo de predicción con otros. En la clase anterior habíamos destacado el uso log-likelihood o verosimilitud que mide la plausibilidad de diferentes valores de los parámetros de un modelo dados los datos observados. En base a este concepto analizamos los estimadores AIC y BIC que luego los usamos en la automatización para encontrar el mejor modelo. Ahora introducimos una nueva herramienta ANOVA para medir y comparar uno o más modelos.

El segundo enfoque es utilizar la validación cruzada. Cuando ajustamos un modelo siempre estaremos sujetos a una muestra, pero no podemos asegurar que dicha muestra es la mejor representación de la población, por lo tanto, el uso de estos mecanismos de validación cruzada nos proporciona una visión más amplia del comportamiento del modelo con múltiples muestras independientes**.**

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 16: Funciones de comparación de modelos

Para entender la presente clase debemos profundizar en los siguientes conceptos vistos en clases anteriores. Específicamente, habíamos tratado el concepto del estadístico R^2, como una medida de la bondad del modelo.

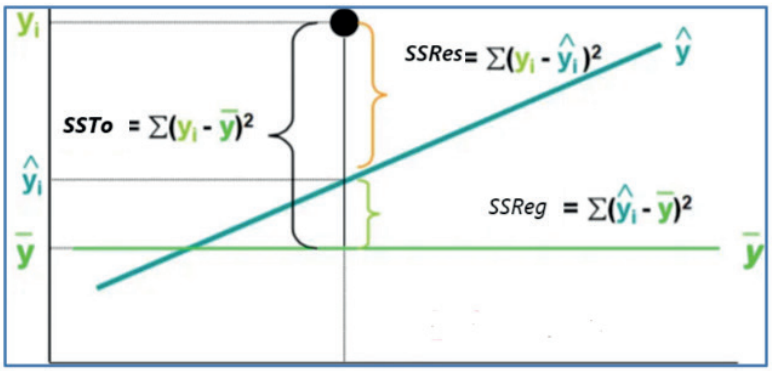
Supongamos por un momento que cada punto y_i estuviera muy cerca de la media $\overline{y}$, esto significaría que cada y_i no dependería de la variación de x_i, y que tampoco habría mucho error aleatorio (ε) en el valor. Dado que esperamos que esto no sea así, entendemos cuánto contribuyen la predicción de x_i y el error aleatorio a y_i. En particular, observemos la distancia entre y_i y la media $\overline{y}$. Escribiremos esta diferencia como SSTo.

En particular, el residuo se define como $y_i - \hat{y}_i$, que es la distancia desde el punto de datos original hasta el valor predicho en la línea de regresión. Se puede considerar como el error restante después de que el modelo haya realizado su trabajo. Esta diferencia se muestra gráficamente en la figura siguiente.

Figura 1

Relación entre SSTo, SSres, SSReg

Creación del autor Alfonso Prado



De la figura podemos deducir que para cada punto i $(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$. El primer término del lado derecho puede entenderse como el valor explicado por el modelo , mientras que el segundo término puede entenderse como un valor que no es explicado por el modelo o el error cometido por el modelo. Si sumamos todas las diferencias y las elevamos al cuadrado (para evitar el neteo de valores positivos y negativos), obtenemos la fórmula de la figura 2a y poniendo la misma en relación con SSto la figura 2b.

Figura 2

Fórmula de regresión (a) y en función de SSto

Creación del autor Alfonso Prado

(a)	(b)
$SSto = SSreg + SSres$	$1 = \frac{SSreg}{SSto} + \frac{SSres}{SSto}$

Pero SSreg/SSto es precisamente la definición de R2 o coeficiente de determinación mencionado en clases anteriores. Aquí vemos porque R^2 puede interpretarse como la fracción de variabilidad en los datos que explica el modelo, por lo tanto, ahora tenemos una definición distinta de que es R^2, en realidad representa cuanto de la variable de

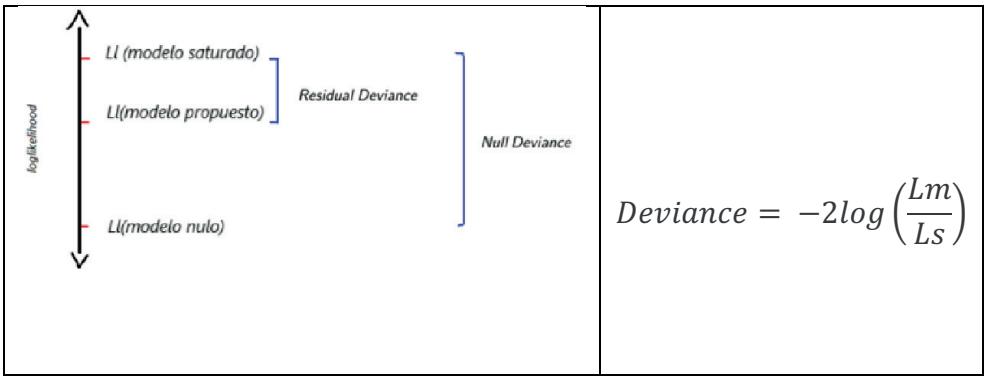
respuesta está explicada por las predictoras y cuanto no. Cuando utilizamos el algoritmo de mínimos cuadrados ordinarios (MCO) lo que éste trata de obtener son los coeficientes (Betas) que minimizan el segundo componente (SSres/SSto).

Por otro lado, en los modelos lineales generalizables (GLM) su métrica principal para explicar la relación entre variables predictoras y de respuesta es conocida como "deviance", (desviación o desviancia en español) es una medida de la bondad del ajuste, cuanto menor sea la desviación, mejor será el ajuste. Más específicamente podemos decir que si, para un modelo glm en particular, denotamos L_m como el máximo likelihood (o verosimilitud) alcanzable bajo este modelo, y a L_s como el likelihood del "modelo saturado" (que se define como el modelo perfecto cuya predicción se ajusta exactamente a las observaciones), entonces el deviance residual se calcularía como el doble negativo del logaritmo de la relación entre L_m y L_s . Y generalizamos diciendo que este modelo también tiene la máxima verosimilitud alcanzable entre todos los modelos posibles dentro del marco glm.

Figura 3

Fórmula de desviación en glm

Creación del autor: Alfonso Prado



Entonces la desviación puede verse como una generalización de la suma de cuadrados residuales de los modelos lineales.

Por otro lado, respecto a anova, habíamos indicado que existen 2 tipos de pruebas en las que usamos este concepto. La prueba conocida como anova de 1 vía cuya función es aov() la cual nos permite contestar la siguiente pregunta: ¿La varianza entre las medias de dos poblaciones es significativamente diferente?", y la segunda que, es de nuestro interés, es

la función anova() que nos permite obtener una tabla de análisis de varianza de uno o más objetos de la clases lm o glm incluyendo una prueba F. A continuación, presentamos la sintaxis de anova.

Figura 4

Sintaxis de la función anova

Fuente: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/anova

```
anova(object, ..., scale = 0, test = "F")
```

En la función anova(), al especificar un solo objeto, se obtiene una tabla de análisis de varianza secuencial para ese ajuste.

Por ejemplo, en clase anteriores habíamos generado un modelo para explicar los valores de viviendas en NY en función la cantidad de unidades habitacionales, los pies cuadrados de la misma y el vecindario (valxSqFt ~TUnits+ GSqFt +Boro). Aplicando anova() obtenemos la siguiente información:

Cada fila de la tabla corresponderá a un término del modelo, más una fila adicional en la parte superior para el modelo nulo (solo considera intersección).

Df (grados de libertad): Los grados de libertad asociados a cada término, normalmente 1 en caso de variable numéricas y numero de niveles -1 en caso de variable nominales.

Deviance: La desviación, una medida del ajuste del modelo a los datos para cada término.

DF residual (grados de libertad residuales): Los grados de libertad residuales tras añadir el término.

Deviance Residual (desviación residual): La desviación residual tras añadir el término.

Figura 5

Anova de 1 modelo logístico, la salida varia en modelos de regresión lineal

Creación de autor: Alfonso Prado

```
Analysis of Deviance Table

Model: binomial, link: logit

Response: exclusivo

Terms added sequentially (first to last)


      Df Deviance Resid. Df Resid. Dev
NULL                    99     136.058
Units  1    0.396        98     135.662
SqFt   1    3.171        97     132.491
Boro   4   68.439        93      64.052
```

Es decir, las reducciones en la suma de cuadrados residuales a medida que se añade cada término de la fórmula (variables predictoras) y se presentan como las filas de la tabla, por último, presenta la suma de cuadrados residuales.

Esa tabla nos permite analizar cuál es el efecto de ir aumentando cada predictora, específicamente, sobre el ejemplo, vemos que al incluir la variable Boro el deviance residual cae de 136.05 a 64 indicado que esta variable ha mejorado la calidad del modelo.

Anova

Si se especifica más de un objeto como argumento en la función anova() ésta permite comparar distintos modelos. La tabla incluye los estadísticos usados en la prueba. Normalmente, la prueba F es la más adecuada, ya que compara la media cuadrática de una fila con la suma de cuadrados residuales del modelo más grande considerado, pero opcionalmente podrían especificarse otros como chi-cuadrado.

La prueba F se construye a partir de dos tipos de cantidades: sumas de cuadrados (SS) y grados de libertad (gl). Estos dos elementos definen un valor cuadrático medio (MS = SS/gl), y obtenemos nuestro estadístico F contrastando el valor de MS asociado con "todo lo demás" que serían los residuos. Esto nos permitirá establecer un marco para comparar modelos diferentes. En otras palabras, lo que queremos hacer es averiguar cómo describir el valor de SS asociado con la diferencia entre dos modelos.

Es importante entender cómo anova trabaja para comprender su salida. La función está hecha para comparar distintos modelos, pero estos deben ser incrementales (algunos autores lo llaman modelos anidados). Por ejemplo, el modelo 1 tendría 2 predictoras, el modelo 2 debe tener las mismas variables del modelo 1 más un incremento de variable(s). Entonces podemos decir que el modelo 1 está anidado dentro del segundo. Para cada

modelo va a calcular los grados de libertad, que en este caso representa la cantidad de variables -1.

En el caso de comparar 2 modelos podemos despejar los SSreg para cada modelo de la siguiente forma $SSreg_{M0}=SSTo-SSres_{M0}$ y $SSreg_{M1}=SSTo-SSres_{M1}$ y su diferencia como $SS\Delta= SSreg_{M0} - SSreg_{M1}$ que sería igual a $SSres_{M0} - SSres_{M1}$. Ahora estos valores tenemos que ponerlos en función de los grados de libertad de cada modelo (también llamado la media de las sumas de cuadrados), entonces para el primer modelo $MS_{m1}= SSRes_{m1}/df_{m1}$ y para el modelo completo $MS\Delta= SS\Delta/df\Delta$. Finalmente obtenemos el estadístico F de la siguiente forma:

Figura 6

Estadístico F (Fratio) para el modelo 1

Creación de autor Alfonso Prado

$$Fratio = MS\Delta/\text{MSm1}$$

Veamos un ejemplo para la comparación de los modelos de viviendas

```
#Creamos los modelos
housingmod1 <- lm(ValuePerSqFt ~ Units  , data=housing )
housingmod2 <- lm(ValuePerSqFt ~ Units +SqFt , data=housing )
housingmod3 <- lm(ValuePerSqFt ~ Units +SqFt + Boro , data=housing )
housingmod4 <- lm(ValuePerSqFt ~ Units +SqFt +Boro + Class , data=housing )
anova(housingmod1,                    housingmod2,housingmod3,housingmod4)
```

```
> anova(housingmod1, housingmod2,housingmod3,housingmod4)
Analysis of Variance Table

Model 1: ValuePerSqFt ~ Units
Model 2: ValuePerSqFt ~ Units + SqFt
Model 3: ValuePerSqFt ~ Units + SqFt + Boro
Model 4: ValuePerSqFt ~ Units + SqFt + Boro + Class
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     98 509249
2     97 470583  1     38666 17.179 7.678e-05 ***
3     93 204960  4    265622 29.502 1.124e-15 ***
4     90 202577  3      2384  0.353    0.7871
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

La tabla tiene una fila para los grados de libertad residuales y la suma de cuadrados de cada modelo. Para todos los modelos, excepto el primero, también se proporciona la diferencia en los grados de libertad con respecto al modelo 1 y la suma de cuadrados obtenida. Esto solo tiene sentido estadístico si los modelos están anidados. Es habitual enumerar los modelos de menor a mayor, pero esto queda a criterio del analista.

Anova es en realidad una prueba de hipótesis cuyo H0 indica que los nuevos coeficientes (con respecto al modelo anterior) son 0 y H1 que al menos un coeficiente es distinto de 0.

Si los nuevos coeficientes son 0, entonces entendemos que las correspondientes variables no tienen valor de predicción y anova recomendará el modelo más sencillo. En principio, si los nuevos coeficientes mejoran el RSS, anova preferirá el modelo más complejo (mayor df), pero si no hay mejora anova seleccionará el modelo más simple.

Prueba de Significancia:

El ANOVA también puede comprobar la significancia de términos individuales o grupos de términos dentro de un modelo examinando el cambio en la desviación estándar al eliminarlos. Esto se realiza comparando el estadístico F calculado, con un valor crítico o valor p obtenido de la distribución F. Si el valor p es inferior al nivel de significancia (normalmente 0,05), se rechaza la hipótesis nula y las diferencias entre las medias de los grupos se consideran estadísticamente significativas.

Prueba de razón de verosimilitud:

El análisis de la desviación estándar está estrechamente relacionado con la prueba de razón de likelihood, una prueba estadística que se utiliza para comparar el ajuste de dos modelos anidados.

Veamos un ejemplo

```
#Si queremos extraer el log-likelihood de cada modelo y los grados de libertad
 podemos usar:
ll1 <- logLik(housingmod1)
> logLik(housingmod1)
'log Lik.' -568.67 (df=3)
ll2 <- logLik(housingmod2)
```

```
> logLik(housingmod2)
'log Lik.' -564.7217 (df=4)
```

#En base a lo anterior podemos obtener la diferencia del log-likelihood

(teststat <- -2 * (as.numeric(ll1)-as.numeric(ll2)))

```
[1] 7.896565
```

#Contrastamos contra la distribución chisq con df igual a la diferencia en grados de libertad de los dos modelos

p.val <- pchisq(teststat, df = 1, lower.tail = FALSE)

```
[1] 0.004952876
```

Con un nivel de significancia de 0,05, podríamos rechazar la hipótesis

#nula. Esto significa que deberíamos usar el modelo complejo en lugar del

#modelo anidado

Podemos hacer lo mismo usando la prueba de razón de likelihood usando el paquete lmtest. Este calcula el log likelihood de cada modelo, su diferencia , los grados de libertad y el nivel de significancia.

lrtest(housingmod1, housingmod2)

```
Likelihood ratio test

Model 1: ValuePerSqFt ~ Units
Model 2: ValuePerSqFt ~ Units + SqFt
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   3 -568.67
2   4 -564.72  1 7.8966   0.004953 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Rechazaremos la hipótesis nula con un nivel de significancia de 0,05. Sin embargo, si fijáramos el nivel de significancia en 0,01, no rechazaríamos la hipótesis nula. Esto indica que el modelo complejo es significativamente más preciso con un alfa de 0,05

Anova para comparar modelos

Explica fundamentos y matemática

Enlace

Validación Cruzada

En clases anteriores habíamos revisado la técnica de validación cruzada como un mecanismo para prevenir el sobreajuste. Pero la validación cruzada sirve principalmente

como una medida de la bondad del modelo mediante el entrenamiento del modelo con distintos subconjuntos de datos.
Para empezar, diremos que la validación cruzada es una técnica de re muestreo, esto quiere decir que partiendo de un dataset grande puedo generar múltiples dataset pequeños.

Para propósito de evitar el sobreajuste habíamos indicado la necesidad de tener 2 subconjuntos conocidos como data de entrenamiento y prueba. Pero para propósito de la evaluación de la bondad del modelo entrenaremos el modelo usando varios subconjuntos de datos y los modelos resultantes se evalúan posteriormente en otros subconjuntos, que no se utilizaron durante el entrenamiento. El promedio del rendimiento que alcanzan los modelos con estos subconjuntos es una estimación más precisa del rendimiento del modelo.

Existen muchas técnicas de re-muestreo, a continuación, algunas de las más comunes.

- ✓ Re muestreo simple: Es el método más sencillo de validación consiste en repartir aleatoriamente las observaciones disponibles en dos grupos, uno se emplea para entrenar al modelo y otro para evaluarlo. La estimación del error es altamente variable dependiendo de qué observaciones se incluyan como conjunto de entrenamiento y cuáles como conjunto de validación.
- ✓ Dejar uno afuera (LOOCV): Es un método iterativo que se inicia empleando como conjunto de entrenamiento todas las observaciones disponibles excepto una, que se excluye para emplearla como validación. Si se emplea una única observación para calcular el error, este varía mucho dependiendo de qué observación se haya seleccionado. El proceso requiere que el modelo sea reajustado y validado tantas veces como observaciones disponibles (n) lo que en algunos casos puede ser muy complicado.

- ✓ Re muestreo aleatorio k-fold Consiste en dividir los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación. Este proceso se repite k veces utilizando un grupo distinto como validación en cada iteración. El proceso genera k estimaciones del error cuyo promedio se emplea como estimación final. En la práctica valores de k entre 5 y 10 son recomendados, lo cual no implica un costo

computacional alto. La principal ventaja de K-fold CV es que consigue una estimación precisa del error de test gracias a un mejor balance entre bias y varianza

Realizar el re muestreo, entrenar el modelo y probar con otro conjunto parecería una ardua labor, pero R lo hace muy fácil a través del uso de la función cv.glm() del paquete boot. Esta función retorna una lista con algunos valores entre los cuales se halla el estadístico delta. Delta es un vector de longitud dos. El primer componente es el error de predicción normal y el segundo componente es un poco más difícil de entender. Recuerde que el modelo se creó con la data de entrenamiento menos 1 fold. Se requiere un ajuste por este motivo y eso es el segundo componente (error de cross validación ajustado por efecto del k-1).

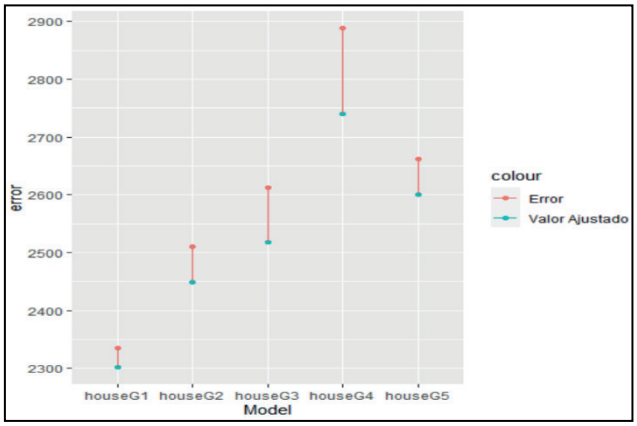
cv.glm toma como argumento el valor k (cantidad de grupos en los que se deben dividir los datos kfold) si el valor es igual al número de observaciones asume que debe usar LOOCV.

En base a la métrica delta preferimos el modelo con el menor rango de validación cruzada.

Figura 7

Comparación de modelos usado validación cruzada

Creación del autor Alfonso Prado



Referencias citadas en la Clase 16

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=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoiX2RpcmVjdCJ9fQ

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

Szretter, M. (2017), Apunte de Regresión Lineal, https://mate.dm.uba.ar/~meszre/apunte_regresion_lineal_szretter.pdf

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

PUCE Virtual

La excelencia no se improvisa

síguenos

CV_ICYS_M