Universidad Nacional Autónoma
de México

Facultad de Ciencias

TAREA No. 3

Métodos de remuestreo: Cross-validation & Bootstrap

Aguirre Armada Guillermo

Figueroa Torres Ivan Emiliano

Luna Gutiérrez Yanely

Ortiz Silva Ana Beatriz

PROFESOR DE ASIGNATURA:
Guillermina Eslava

PROFESOR DE ADJUNTO:
Sofía Guzman

12 de noviembre de 2020

CIUDAD UNIVERSITARIA, CD. MX.

1 Ejercicio 8

Realizaremos cross validation en un conjunto de datos simulados.

(a) Genere un conjunto de datos simulados de la siguiente manera:
Fijamos una semilla para replicar los datos y generamos las observaciones de acuerdo al código que se da en el libro:

```
set .seed (1)
x=rnorm (100)
y=x-2* x^2+ rnorm (100)
```

En este conjunto de datos, ¿Quién es n y quién es p? Escribe el modelo utilizado para generar los datos en forma de ecuación.

Tenemos el número de observaciones n=100, el numero de variables es p=2 y la ecuación del modelo es

$$Y = X - 2X^2 + \epsilon$$

Donde ϵ se distribuye normal.

(b) Crea una gráfica de dispersión de Y respecto a X. Comenta que piensas.

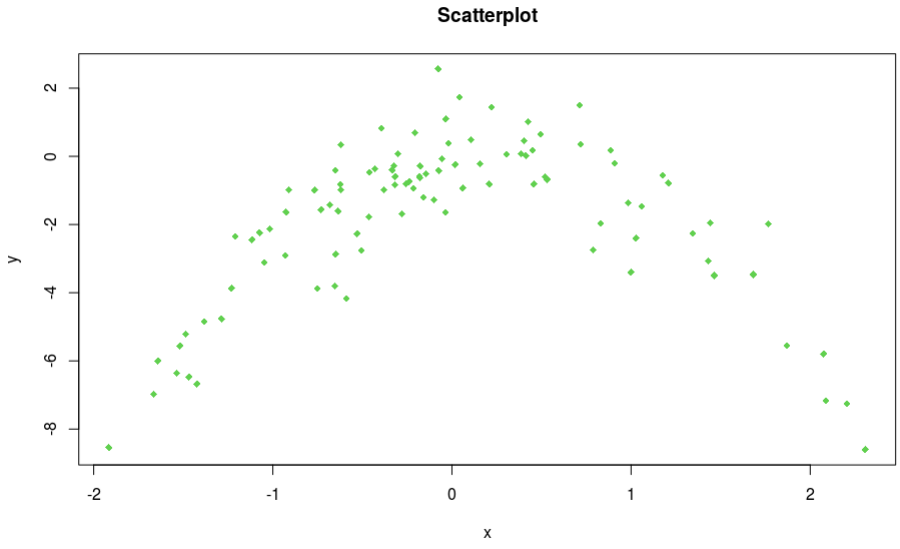

Figure 1: **Gráfica de dispersión.** $Y = X - 2X^2 + \epsilon$

Podemos observar que tiene una forma curva, esto se debe a la construción pues proviene de un polinómio de grado 2.

(c) Elige una semilla aleatoria y calcula los errores LOOCV que resulten de ajustar los siguientes 4 modelos usando mínimos cuadrados.

i. $Y = \beta_0 + \beta_1 X + \epsilon$
ii. $Y = \beta_0 + \beta_1 X + \beta_2 X2 + \epsilon$
iii. $Y = \beta_0 + \beta_1 X + \beta_2 X2 + \beta_3 X3 + \epsilon$
iv. $Y = \beta_0 + \beta_1 X + \beta_2 X2 + \beta_3 X3 + \beta_4 X4 + \epsilon$

NOTA: en el ejercicio 8 en lugar de usar LOOCV utilice K-fold cross-validation con B $>$ 1 repeticiones. Los valores de K y B son de su elección.

modelo	semilla 102		semilla 12	
	500	1000	500	1000
1	5.89	5.89	5.90	5.91
2	**1.09**	**1.09**	**1.09**	**1.09**
3	1.10	1.10	1.10	1.10
4	1.12	1.12	1.12	1.12

Table 1: Tabla de errores de predicción usando C.V. variando B y la semilla.

Realizamos varias pruebas, eligiendo K =5, 20 y 100 notamos que la diferencia entre 5 y 20 era significativa, un cambio de 0.02 en el error de predicción, mientras que el tiempo de cómputo no se elevaba demasiado, cuando probamos $K = 100$ no se observó un cambio significativo en los errores, sin embargo, el tiempo para calcularlo sí aumentó en gran medida, de 4 minutos en k=20 a 20 minutos aproximadamente en k=100, por eso decidimos utilizar $k = 20$.

(d) Repite (C) usando otra semilla aleatoria, y reporta tus resultados. ¿Tus resultados son iguales a los obtenidos en (C)? ¿Por qué?

Es posible distinguir en la Tabla 1 que al utilizar semillas diferentes los resultados cambiaron, esto se debe a que es un algoritmo con aleatoriedad al elegir las observaciones que conforman los K grupos.

(e) ¿Cuál de los modelos en (C) tiene el menor error de predicción? ¿Es este el que esperában? Explica tu respuesta.

Elegimos el modelo 2, pues es el que tiene el error más pequeño. Esto lo podemos observar en la tabla número 1, y en realidad ya lo esperábamos, pues al momento de contruir las observaciones usamos un polinomio de grado 2.

(f) Comenta la significancia estadística de los coeficientes estimados que resultan de ajustar cada modelo en (C) usando mínimos cuadrados. ¿Los resultados concuerdan con las conclusiones extraídas de los resultados de cross validation?

Podemos notar que el término cuadrático es altamente significativo y no sucede esto mismo con el resto de los términos, lo cual tiene sentido con los resultados de la validación cruzada pues, al agregarle más variables sin significancia estadística sólo se incrementó el error del modelo.

Variable	mod 1	mod 2	mod 3	mod 4	Significancia
	P value				
intercepto	0	0	0	0	***
X	.33	.03	.03	.03	*
X^2		**0**	**0**	**0**	***
X^3			.77	.77	
X^4				.64	

Table 2: **Resumen de los 5 modelos.**

2 Ejercicio 9

Para este ejercicio usaremos la base `Boston` del paquete `MASS`.

(a) Basandonos en estos datos proponemos un estimador para la media poblacional de `medv` (esta variable representa el valor medio de las casas ocupadas en las comunidades que conforman la base). Llamaremos a este estimador $\hat{\mu}$.

$$\hat{\mu} = \bar{x} = 22.532$$

(b) Proponemos un estimador para el error estándar de $\hat{\mu}$.

Sacamos la desviación estándar muestral y la dividimos entre la raiz cuadrada del número de observaciones para sacar el estimador del error estándar.

$$SE_{\hat{\mu}} = \sqrt{\frac{\sigma^2}{n}} = 0.4088$$

(c) Ahora estimaremos el error estándar de $\hat{\mu}$, pero esta vez usando bootstrap.

Simulaciones	bias	$\text{SE}_{\hat{(\mu)}}$
1,000	0.0022	0.3947
10,000	-0.0029	0.4104
100,000	0.0003	0.4091
1,000,000	-6.91e-06	0.4083

Table 3: Estimación del error estándar por método *Bootstrap* variando el número de simulación

Observamos que cuanto mayores son las simulaciones más cerca está el $SE_{\hat{\mu}}$ estimado por bootstrap del $SE_{\hat{\mu}}$ estimado en b). Existe una diferencia significativa en las simulaciones 1,000 y 10,000, esto lo notamos en el bias de la tabla, en la primera tenemos un estimador menor al de b) y en la segunda ya lo sobrepasa pero a medida de que se aumenta las simulaciones disminuye el bias.

(d) Basándonos en la estimación hecha en (c) proponemos un intervalo del 95% para $\hat{\mu}$.

Usamos la fórmula

$$[\hat{\mu} - 2SE_{\hat{\mu}}, \hat{\mu} + 2SE_{\hat{\mu}}]$$

para calcular el intervalo a un 95% de confianza.

Método	Intervalo de confianza
B 1,000	[21.717 , 23.347]
B 10,000	[21.711 , 23.353]
B 100,000	[21.714 , 23.350]
B 1,000,000	[21.716 , 23.349]
t.test	[21.729 , 23.336]

Table 4: Intervalos de confianza según los estimadores sacados por *Bootstrap* comparandolo con el *t.test*

Notamos que los intervalos que más se ajustan a los obtenidos en la prueba t son los calculados por los estimadores por 1,000 y 1,000,000 simulaciones.

(e) Propondremos ahora como estimador de la mediana poblacional de `medv` a la mediana muestral, es decir, el segundo cuartil muestral, al que llamaremos $\hat{\mu}_{med}$.

$$\hat{\mu}_{med} = 21.2$$

(f) Ahora queremos obtener el error estándar de $\hat{\mu}_{med}$, sin embargo, a diferencia de $\hat{\mu}$, en este caso no tenemos una fórmula sencilla que nos permita calcularlo, por lo que lo estimaremos usando bootstrap.

Para esto probamos con diferentes valores de **B** (número de muestras formadas) y comparamos los resultados en la Tabla 5

	B = 1000	B = 10,000	B = 100,000
$\hat{\mu}_{med}$	21.204	21.191	21.189
$se(\hat{\mu}_{med})$	0.376	0.380	0.378

Table 5: **Estimaciones hechas con bootstrap** *para la mediana de los datos y su error estándar.*

En esta tabla notamos que los valores estimados de la mediana son muy cercanos para los tres valores de B, y el valor del error estándar también varia muy poco, sin embargo, para B = 10,000 aumenta un poco y luego disminuye para B = 100,000 por lo que no podríamos decir que aún con cien mil muestras el valor no es muy estable.

(g) Ahora estimaremos el cuantil al 10% de `medv` usando la función `quantile` de `R`. Llamaremos a este estimador $\hat{\mu}_{0.1}$.

$$\hat{\mu}_{0.1} = 12.75$$

(h) Usando bootstrap estimamos el error estándar de $\hat{\mu}_{0.1}$. Los resultados se muestran en la Tabla 6

	B = 1000	B = 10,000	B = 100,000
$\hat{\mu}_{0.1}$	12.754	12.758	12.758
$se(\hat{\mu}_{med})$	0.495	0.503	0.503

Table 6: **Estimaciones hechas con bootstrap** *para el cuantil al 10% de los datos y su error estándar.*

Redondeando a tres decimales, el valor del cuantil al 10% queda como 12.758 para B=10,000 y B=100,000. Igualmente, el valor del error estándar es el mismo, 0.503 para las dos estimaciones, por lo que podríamos decir que 10,000 muestras obtenidas con bootstrap son suficientes para estimar el error estándar de $\hat{\mu}_{0.1}$.