An Introduction to Statistical Learning [Tibshirani-2013]

Métodos de remuestreo [Chapter 5]

Data Scientist
Jhoan Esteban Ruiz Borja Msc

Los métodos de **remuestreo** son una herramienta indispensable en las estadísticas modernas. Implican extraer muestras de un conjunto de entrenamiento y volver a colocar un modelo de interés en cada muestra para obtener información adicional sobre el modelo ajustado. Por ejemplo, para estimar la variabilidad de un ajuste de regresión lineal, podemos extraer repetidamente diferentes muestras de los datos de entrenamiento, ajustar una regresión lineal a cada nueva muestra y luego examinar la medida en que difieren los ajustes resultantes. Este enfoque puede permitirnos obtener información que no estaría disponible si se ajusta el modelo solo una vez utilizando la muestra de entrenamiento original.

Los enfoques de **remuestreo** pueden ser computacionalmente costosos, ya que implican ajustar el mismo método estadístico varias veces utilizando diferentes subconjuntos de los datos de entrenamiento. Sin embargo, debido a los recientes avances en el poder de cómputo, los requisitos computacionales de los métodos de **remuestreo** generalmente no son prohibitivos. En este capítulo, analizamos dos de los métodos de **remuestreo** más utilizados, la **validación cruzada *(Cross-Validation)*** y el **bootstrap**. Ambos métodos son herramientas importantes en la aplicación práctica de muchos procedimientos de aprendizaje estadístico. Por ejemplo, la validación cruzada se puede usar para estimar el error de prueba asociado con un método de aprendizaje estadístico determinado para evaluar su desempeño o para seleccionar el nivel apropiado de flexibilidad. El proceso de evaluar el desempeño de un modelo se conoce como **evaluación del modelo *(model assesment)***, mientras que el proceso de selección del nivel adecuado de flexibilidad para un modelo se conoce como **selección de modelo *(model selection)***. El **bootstrap** se usa en varios contextos, más comúnmente para proporcionar una medida de precisión de una estimación de parámetro o de un método de aprendizaje estadístico dado.

Validación cruzada (Cross-Validation: CV) [5.1]

En el Capítulo 2 discutimos la distinción entre la tasa de error de prueba ***(test error rate)*** y la tasa de error de entrenamiento ***(training error rate)***. El error de prueba es el error promedio que resulta de usar un método de aprendizaje estadístico para predecir la respuesta en una nueva observación, es decir, una medida que no se usó para entrenar el método. Dado un conjunto de datos, el uso de un método de aprendizaje estadístico particular está garantizado si produce un error de prueba bajo. El error de prueba se puede calcular fácilmente si un conjunto de prueba designado está disponible. Desafortunadamente, este no suele ser el caso. En contraste, el error de entrenamiento se puede calcular fácilmente aplicando el método de aprendizaje estadístico a las observaciones utilizadas en su entrenamiento. Pero como vimos

en el Capítulo 2, la tasa de error de entrenamiento a menudo es bastante diferente de la tasa de error de prueba, y en particular la primera puede subestimar dramáticamente la segunda.

En ausencia de un conjunto de pruebas designado muy grande que se pueda usar para estimar directamente la tasa de error de la prueba, se pueden usar varias técnicas para estimar esta cantidad usando los datos de entrenamiento disponibles. Algunos métodos hacen un ajuste matemático a la tasa de error de entrenamiento para estimar la tasa de error de prueba. Dichos enfoques se analizan en el Capítulo 6. En esta sección, en cambio, consideramos una clase de métodos que estiman la tasa de error de la prueba al presentar un subconjunto de las observaciones de capacitación del proceso de adaptación y luego aplicar el método de aprendizaje estadístico a las observaciones de prácticas.

En las Secciones 5.1.1–5.1.4, por simplicidad, asumimos que estamos interesados en realizar una regresión con una respuesta cuantitativa. En la Sección 5.1.5 consideramos el caso de clasificación con una respuesta cualitativa. Como veremos, los conceptos clave siguen siendo los mismos, independientemente de si la respuesta es cuantitativa o cualitativa.

El enfoque del conjunto de validación [5.1.1]

Supongamos que nos gustaría estimar el error de prueba asociado con ajustar un método de aprendizaje estadístico particular en un conjunto de observaciones. El enfoque del **conjunto de validación**, que se muestra en la Figura 5.1, es una estrategia muy simple para esta tarea. Implica dividir aleatoriamente el conjunto disponible de observaciones en dos partes, un conjunto de entrenamiento ***(Training)*** y un conjunto de validación ***(Validation)*** o conjunto de espera ***(hold-out)***. El modelo se ajusta al conjunto de entrenamiento, y el modelo ajustado se usa para predecir las respuestas para las observaciones en el conjunto de validación. La tasa de error del conjunto de validación resultante, que generalmente se evalúa utilizando MSE en el caso de una respuesta cuantitativa, proporciona una estimación de la tasa de error de la prueba.

Ilustramos el enfoque del conjunto de validación en el conjunto de datos **Auto**. Recuerde del Capítulo 3 que parece haber una relación no lineal entre **mpg** y **horsepower**, y que un modelo que predice los **mpg** usando **horsepower** y $\boldsymbol{horsepower^2}$ da mejores resultados que un modelo que usa solo un término lineal. Es natural preguntarse si un ajuste cúbico o de orden superior puede proporcionar resultados aún mejores. Respondemos a esta pregunta en el Capítulo 3 observando los $\boldsymbol{p}$-valores asociados con un término cúbico y los términos polinomiales de orden superior en una regresión lineal. Pero también podríamos responder a esta pregunta utilizando el método de validación. Dividimos aleatoriamente las 392 observaciones en dos conjuntos, un conjunto de entrenamiento que contiene 196 de los puntos de datos y un conjunto de validación que contiene las 196 observaciones restantes. Los índices de error del conjunto de validación que resultan de ajustar varios modelos de regresión en la muestra de entrenamiento y evaluar su desempeño en la muestra de validación, utilizando MSE como medida del error del conjunto de validación, se muestran en el panel de la izquierda de la Figura 5.2. El conjunto de validación MSE para el ajuste cuadrático es considerablemente más pequeño que para el ajuste lineal. Sin embargo, el

conjunto de validación MSE para el ajuste cúbico es en realidad un poco más grande que para el ajuste cuadrático. Esto implica que incluir un término cúbico en la regresión no conduce a una mejor predicción que simplemente usar un término cuadrático.

FIGURE 5.1. *A schematic display of the validation set approach. A set of n observations are randomly split into a training set (shown in blue, containing observations 7, 22, and 13, among others) and a validation set (shown in beige, and containing observation 91, among others). The statistical learning method is fit on the training set, and its performance is evaluated on the validation set.*

Recuerde que para crear el panel de la izquierda de la Figura 5.2, dividimos aleatoriamente el conjunto de datos en dos partes, un conjunto de entrenamiento y un conjunto de validación. Si repetimos el proceso de dividir aleatoriamente el conjunto de muestras en dos partes, obtendremos una estimación algo diferente para el MSE de prueba. Como ilustración, el panel de la derecha de la Figura 5.2 muestra diez curvas de MSE de conjuntos de validación diferentes del conjunto de datos **Auto**, producidas utilizando diez divisiones aleatorias diferentes de las observaciones en los conjuntos de entrenamiento y validación. Las diez curvas indican que el modelo con un término cuadrático tiene un conjunto de validación mucho más pequeño que el modelo con solo un término lineal. Además, las diez curvas indican que no hay mucho beneficio en la inclusión de términos polinomiales de orden superior o cúbico en el modelo. Pero vale la pena señalar que cada una de las diez curvas resulta en una estimación de MSE de prueba diferente para cada uno de los diez modelos de regresión considerados. Y no hay consenso entre las curvas en cuanto a qué modelo da como resultado el conjunto de validación más pequeño MSE. Basándonos en la variabilidad entre estas curvas, todo lo que podemos concluir con confianza es que el ajuste lineal no es adecuado para estos datos.

El enfoque del conjunto de validación es conceptualmente simple y fácil de implementar. Pero tiene dos inconvenientes potenciales:

1. Como se muestra en el panel de la derecha de la Figura 5.2, la estimación de la validación de la tasa de error de la prueba puede ser muy variable, dependiendo de las observaciones que se incluyen en el conjunto de entrenamiento y las observaciones que se incluyen en el conjunto de validación.
2. En el enfoque de validación, solo un subconjunto de las observaciones, aquellas que están incluidas en el conjunto de entrenamiento en lugar de en el conjunto de validación, se utilizan para ajustar el modelo. Como los métodos estadísticos tienden a tener un peor desempeño cuando se entrenan con menos observaciones, esto sugiere que la tasa de error del conjunto de validación puede tender a sobreestimar la tasa de error de prueba para el ajuste del modelo en todo el conjunto de datos.

En las siguientes subsecciones, presentaremos la validación cruzada, un refinamiento del enfoque del conjunto de validación que aborda estos dos problemas.

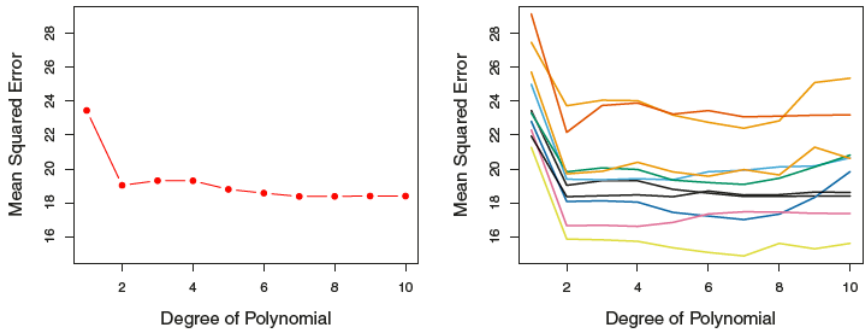

FIGURE 5.2. *The validation set approach was used on the* `Auto` *data set in order to estimate the test error that results from predicting* `mpg` *using polynomial functions of* `horsepower`. Left: *Validation error estimates for a single split into training and validation data sets.* Right: *The validation method was repeated ten times, each time using a different random split of the observations into a training set and a validation set. This illustrates the variability in the estimated test MSE that results from this approach.*

Validación cruzada de dejar uno fuera [5.1.2]

La validación cruzada de Leave-one-out (LOOCV) está estrechamente relacionada con el enfoque del conjunto de validación de la Sección 5.1.1, pero intenta solucionar los inconvenientes de ese método.

Al igual que el enfoque del conjunto de validación, LOOCV implica dividir el conjunto de observaciones en dos partes. Sin embargo, en lugar de crear dos subconjuntos de tamaño comparable, se utiliza una sola observación $(\boldsymbol{x_1}, \boldsymbol{y_1})$ para el conjunto de validación, y las observaciones restantes $\{(\boldsymbol{x_2}, \boldsymbol{y_2}), \ldots, (\boldsymbol{x_n}, \boldsymbol{y_n})\}$ conforman el conjunto de entrenamiento. El método de aprendizaje estadístico se ajusta a las $\boldsymbol{n-1}$ observaciones de entrenamiento, y se hace una predicción $\hat{\boldsymbol{y}}_{\boldsymbol{1}}$ para la observación excluida, utilizando su valor $\boldsymbol{x_1}$. Dado que $(\boldsymbol{x_1}, \boldsymbol{y_1})$ no se usaron en el proceso de ajuste, $\boldsymbol{MSE_1} = (\boldsymbol{y_1} - \hat{\boldsymbol{y}}_{\boldsymbol{1}})^{\boldsymbol{2}}$ proporciona una estimación aproximadamente imparcial para el error de prueba. Pero, aunque $\boldsymbol{MSE_1}$ no es imparcial para el error de prueba, es una estimación pobre porque es muy variable, ya que se basa en una sola observación $(\boldsymbol{x_1}, \boldsymbol{y_1})$.

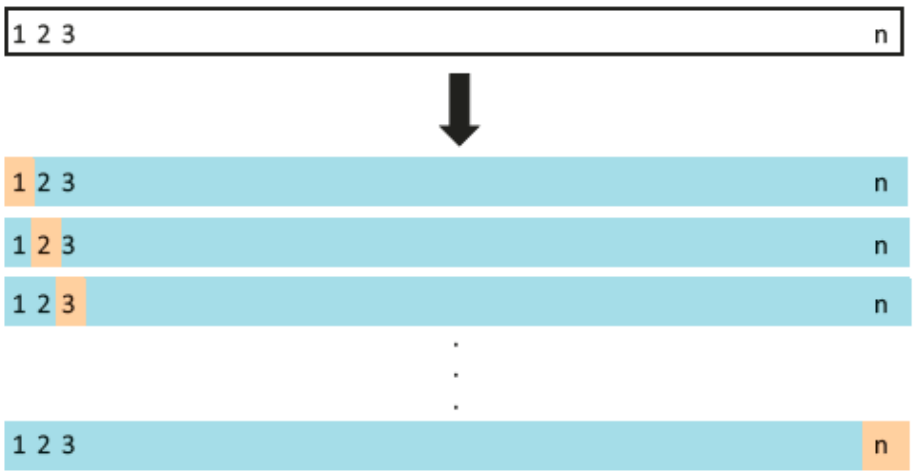

Figura 5.3: Una visualización esquemática de LOOCV. Un conjunto de n puntos de datos se divide repetidamente en un conjunto de entrenamiento (mostrado en azul) que contiene todas las observaciones menos una, y un conjunto de validación que contiene solo esa observación (que se muestra en color beige). El error de la prueba se calcula promediando los n valores de MSE resultantes. El primer conjunto de entrenamiento contiene todos menos la observación 1, el segundo conjunto de entrenamiento contiene todos menos la observación 2, y así sucesivamente.

Podemos repetir el procedimiento seleccionando (x_2, y_2) para los datos de validación, entrenando el procedimiento de aprendizaje estadístico en las $n-1$ observaciones $\{(x_1, y_1), (x_3, y_3), \ldots, (x_n, y_n)\}$, y calcular $MSE_2 = (y_2 - \hat{y}_2)^2$. La repetición de este enfoque n veces produce n errores al cuadrado, $MSE_1, \ldots, MSE_n$.

La estimación de LOOCV para el MSE de prueba es el promedio de estas n estimaciones de error de prueba:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} MSE_i \qquad (5.1)$$

En la Figura 5.3 se ilustra un esquema del enfoque LOOCV.

LOOCV tiene un par de ventajas importantes sobre el enfoque de conjunto de validación. Primero, tiene mucho menos sesgo. En LOOCV, ajustamos repetidamente el método de aprendizaje estadístico utilizando conjuntos de entrenamiento que contienen $n-1$ observaciones, casi tantos como están en el conjunto de datos completo. Esto contrasta con el enfoque del conjunto de validación, en el que el conjunto de capacitación suele ser aproximadamente la mitad del tamaño del conjunto de datos original. En consecuencia, el enfoque LOOCV tiende a no sobreestimar la tasa de error de prueba tanto como lo hace el enfoque de conjunto de validación. En segundo lugar, a diferencia del enfoque de validación que dará resultados diferentes cuando se aplique repetidamente debido a la aleatoriedad en las divisiones del conjunto de entrenamiento/validación, la ejecución de LOOCV varias veces siempre producirá los mismos resultados: no hay aleatoriedad en las divisiones del conjunto de entrenamiento/validación.

Utilizamos LOOCV en el conjunto de datos **Auto** para obtener una estimación del conjunto de prueba MSE que resulta de ajustar un modelo de regresión lineal para predecir el **mpg** usando las funciones polinómicas de **horsepower**. Los resultados se muestran en el panel de la izquierda de la Figura 5.4.

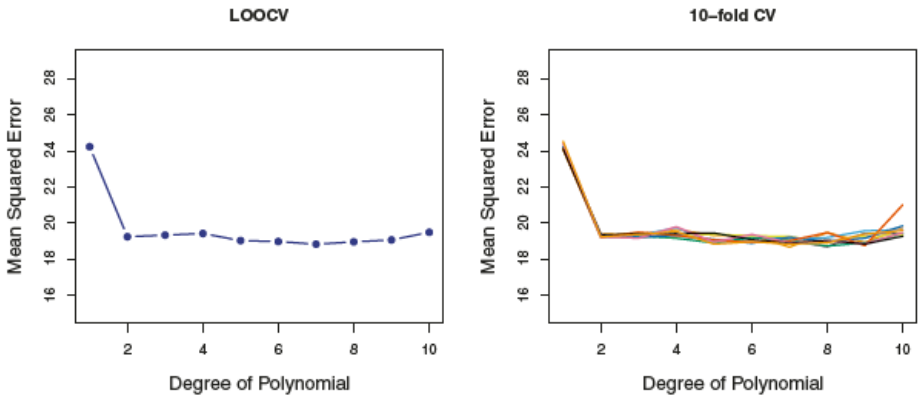

Figura 5.4: La validación cruzada se usó en el conjunto de datos **Auto** para estimar el error de prueba que resulta de la predicción de **mpg** usando funciones polinómicas de potencia. Izquierda: La curva de error LOOCV. Derecha: 10 veces la CV se ejecutó nueve veces por separado, cada una con una división aleatoria diferente de los datos en diez partes. La figura muestra las nueve curvas de error CV ligeramente diferentes.

LOOCV tiene el potencial de ser costoso de implementar, ya que el modelo debe ajustarse $\boldsymbol{n}$ veces. Esto puede llevar mucho tiempo si $\boldsymbol{n}$ es grande, y si cada modelo individual es lento para adaptarse. Con regresión lineal o polinomial de mínimos cuadrados, un atajo sorprendente hace que el costo de LOOCV sea el mismo que el de un solo modelo. La siguiente fórmula sostiene:

$$\boldsymbol{CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2} \qquad (\mathbf{5.2})$$

Donde $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ es el valor ajustado del $\boldsymbol{i}$-esimo ajuste de mínimos cuadrados original, y $\boldsymbol{h_i}$ es el apalancamiento definido en (3.37) en la página 98. Esto es como el MSE ordinario, excepto que el $\boldsymbol{i}$-esimo residual se divide por $\boldsymbol{1 - h_i}$. El apalancamiento se encuentra entre $\boldsymbol{1/n}$ y $\boldsymbol{1}$, y refleja la cantidad que una observación influye en su propio ajuste. Por lo tanto, los residuos para los puntos de alto apalancamiento se inflan en esta fórmula exactamente por la cantidad correcta para que se mantenga esta igualdad.

LOOCV es un método muy general y puede usarse con cualquier tipo de modelado predictivo. Por ejemplo, podríamos usarlo con regresión logística o análisis discriminante lineal, o cualquiera de los métodos descritos en capítulos posteriores. La fórmula mágica (5.2) no se cumple en general, en cuyo caso el modelo tiene que ser reacondicionado $\boldsymbol{n}$ veces.

K-Fold Validación Cruzada [5.1.3]

Una alternativa a LOOCV es la CV de $\boldsymbol{k}$-fold. Este enfoque implica dividir aleatoriamente el conjunto de observaciones en $\boldsymbol{k}$ grupos, o pliegues, de aproximadamente el mismo tamaño.

El primer pliegue se trata como un conjunto de validación, y el método se ajusta a los $\boldsymbol{k-1}$ restantes. El error cuadrático medio, $\boldsymbol{MSE_1}$, se calcula luego sobre las observaciones en el pliegue retenido. Este procedimiento se repite $\boldsymbol{k}$ veces; cada vez, un grupo diferente de observaciones se trata como un conjunto de validación. Este proceso da como resultado $\boldsymbol{k}$ estimaciones del error de prueba, $\boldsymbol{MSE_1}, \boldsymbol{MSE_2}, \ldots, \boldsymbol{MSE_k}$. La estimación de la CV de $\boldsymbol{k}$-fold se calcula promediando estos valores,

$$\boldsymbol{CV_{(n)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i} \qquad (\mathbf{5.3})$$

La figura 5.5 ilustra el enfoque de la CV de $\boldsymbol{k}$-fold.

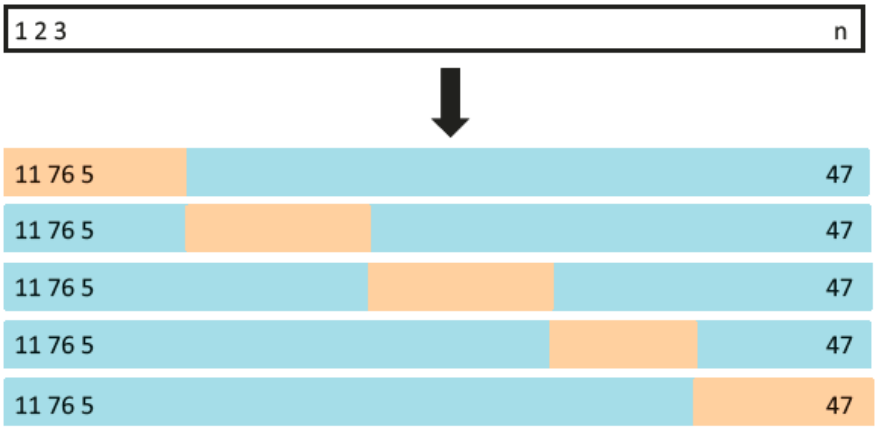

Figura 5.4: Una visualización esquemática de la CV de 5-fold. Un conjunto de $\boldsymbol{n}$ observaciones se divide aleatoriamente en cinco grupos no superpuestos. Cada una de estas quintas actúa como un conjunto de validación (mostrado en color beige) y el resto como un conjunto de entrenamiento (mostrado en azul). El error de la prueba se estima promediando las cinco estimaciones MSE resultantes.

No es difícil ver que LOOCV es un caso especial de la CV de $\boldsymbol{k}$-fold en el que $\boldsymbol{k}$ se establece como igual a $\boldsymbol{n}$. En la práctica, uno normalmente realiza la CV de $\boldsymbol{k}$-fold usando $\boldsymbol{k = 5}$ o $\boldsymbol{k = 10}$. ¿Cuál es la ventaja de usar $\boldsymbol{k = 5}$ o $\boldsymbol{k = 10}$ en lugar de $\boldsymbol{k = n}$? La ventaja más obvia es computacional. LOOCV requiere ajustar el método de aprendizaje estadístico $\boldsymbol{n}$ veces. Esto tiene el potencial de ser computacionalmente caro (a excepción de los modelos lineales ajustados por mínimos cuadrados, en cuyo caso se puede usar la fórmula (5.2)). Pero la validación cruzada es un enfoque muy general que se puede aplicar a casi cualquier método de aprendizaje estadístico. Algunos métodos de aprendizaje estadístico tienen procedimientos de ajuste intensivos computacionalmente, por lo que la ejecución de LOOCV puede plantear problemas computacionales, especialmente si $\boldsymbol{n}$ es extremadamente grande. Por el contrario, realizar una CV de 10-fold requiere ajustar el procedimiento de aprendizaje solo diez veces, lo que puede ser mucho más factible. Como vemos en la Sección 5.1.4, también puede haber otras ventajas no computacionales para realizar la CV de 5-fold a 10-fold, lo que implica la compensación de sesgo-varianza.

El panel de la derecha de la Figura 5.4 muestra nueve diferentes estimaciones de la CV de 10-fold para el conjunto de datos **Auto**, cada una resultante de una división aleatoria diferente de las observaciones en diez pliegues. Como podemos ver en la figura, existe cierta variabilidad en las estimaciones de CV como resultado de la variabilidad en cómo se dividen las observaciones en diez pliegues. Pero esta variabilidad suele ser mucho más baja que la variabilidad en las estimaciones de error de prueba que resultan del enfoque de conjunto de validación (panel derecho de la Figura 5.2).

Cuando examinamos datos reales, no conocemos el verdadero MSE de prueba, por lo que es difícil determinar la precisión de la estimación de validación cruzada. Sin embargo, si examinamos los datos simulados, podemos calcular el verdadero MSE de prueba y, por lo tanto, podemos evaluar la precisión de nuestros resultados de validación cruzada. En la Figura 5.6, trazamos las estimaciones de validación cruzada y las tasas de error de prueba reales que resultan de la aplicación de splines de suavizado a los conjuntos de datos simulados ilustrados en las Figuras 2.9–2.11 del Capítulo 2. La MSE de prueba verdadera se muestra en azul. Las líneas continuas negras discontinuas y naranjas muestran las estimaciones estimadas de la CV de LOOCV y 10-fold. En las tres parcelas, las dos estimaciones de validación cruzada son muy similares. En el panel de la derecha de la Figura 5.6, el MSE de prueba real y las curvas de validación cruzada son casi idénticas. En el panel central de la Figura 5.6, los dos conjuntos de curvas son similares en los grados más bajos de flexibilidad, mientras que las curvas CV sobrestiman el MSE del conjunto de prueba para obtener grados más altos de flexibilidad. En el panel de la izquierda de la Figura 5.6, las curvas CV tienen la forma general correcta, pero subestiman el verdadero MSE de prueba.

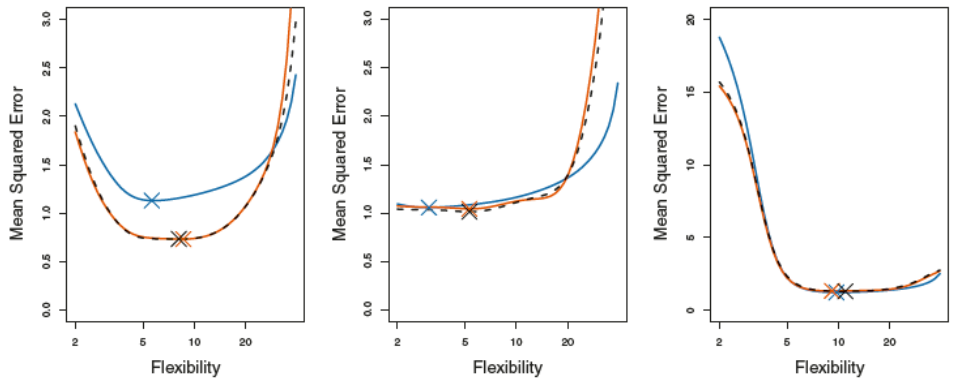

Figura 5.6: Prueba MSE verdadera y estimada para los conjuntos de datos simulados en las Figuras 2.9 (izquierda), 2.10 (centro) y 2.11 (derecha). La prueba MSE verdadera se muestra en azul, la estimación LOOCV se muestra como una línea discontinua negra y la estimación de la CV de 10-fold se muestra en naranja. Las cruces indican el mínimo de cada una de las curvas MSE.

Cuando realizamos la validación cruzada, nuestro objetivo podría ser determinar qué tan bien puede esperarse que un procedimiento de aprendizaje estadístico dado se realice en datos independientes; en este caso, la estimación real de la prueba MSE es de interés. Pero en otros momentos solo nos interesa la ubicación del punto mínimo en la curva de MSE de prueba estimada. Esto se debe a que podríamos estar realizando una validación cruzada en varios métodos de aprendizaje estadístico, o en un solo método con diferentes niveles de

flexibilidad, a fin de identificar el método que resulte en el error de prueba más bajo. Para este propósito, la ubicación del punto mínimo en la curva MSE de prueba estimada es importante, pero el valor real de la MSE de prueba estimada no lo es. Encontramos en la Figura 5.6 que a pesar del hecho de que a veces subestiman el verdadero MSE de prueba, todas las curvas de CV se aproximan a la identificación del nivel correcto de flexibilidad, es decir, el nivel de flexibilidad correspondiente al MSE de prueba más pequeño.

Compensación de variación de sesgo para la validación cruzada k-Fold [5.1.4]

Mencionamos en la Sección 5.1.3 que la CV de $\boldsymbol{k}$-fold con $\boldsymbol{k < n}$ tiene una ventaja computacional para LOOCV. Pero dejando de lado los problemas informáticos, una ventaja menos obvia pero potencialmente más importante de la CV de $\boldsymbol{k}$-fold es que a menudo proporciona estimaciones más precisas de la tasa de error de prueba que LOOCV. Esto tiene que ver con una compensación sesgo-desviación.

Se mencionó en la Sección 5.1.1 que el enfoque del conjunto de validación puede llevar a sobreestimar la tasa de error de la prueba, ya que en este enfoque el conjunto de capacitación utilizado para ajustar el método de aprendizaje estadístico contiene solo la mitad de las observaciones de todo el conjunto de datos. Usando esta lógica, no es difícil ver que LOOCV proporcionará estimaciones aproximadamente imparciales del error de la prueba, ya que cada conjunto de entrenamiento contiene $\boldsymbol{n - 1}$ observaciones, que es casi la cantidad de observaciones en el conjunto de datos completo. Y al realizar la CV de $\boldsymbol{k}$-fold para $\boldsymbol{k = 5}$ o $\boldsymbol{k = 10}$ conducirá a un nivel intermedio de sesgo, ya que cada conjunto de entrenamiento contiene $\frac{(\boldsymbol{k-1})\boldsymbol{n}}{\boldsymbol{k}}$ observaciones: menos que en el enfoque LOOCV, pero sustancialmente más que en el planteamiento del conjunto de validación. Por lo tanto, desde la perspectiva de la reducción del sesgo, está claro que LOOCV debe preferirse la CV de $\boldsymbol{k}$-fold.

Sin embargo, sabemos que el sesgo no es la única fuente de preocupación en un procedimiento de estimación; también debemos considerar la varianza del procedimiento. Resulta que LOOCV tiene una varianza más alta que la CV de $\boldsymbol{k}$-fold con $\boldsymbol{k < n}$. ¿Por qué es este el caso? Cuando realizamos LOOCV, estamos en efecto promediando los resultados de $\boldsymbol{n}$ modelos ajustados, cada uno de los cuales se entrena en un conjunto de observaciones casi idéntico; por lo tanto, estos resultados están altamente correlacionados entre sí. En contraste, cuando realizamos la CV de $\boldsymbol{k}$-fold con $\boldsymbol{k < n}$, estamos promediando las salidas de $\boldsymbol{k}$ modelos ajustados que están algo menos relacionados entre sí, ya que la superposición entre los conjuntos de entrenamiento en cada modelo es menor. Dado que la media de muchas cantidades altamente correlacionadas tiene una varianza mayor que la media de muchas cantidades que no están tan correlacionadas, la estimación del error de prueba que resulta de LOOCV tiende a tener una varianza mayor que la estimación del error de prueba que resulta de la CV de $\boldsymbol{k}$-fold.

Para resumir, hay una compensación de sesgo-desviación asociada con la elección de $\boldsymbol{k}$ en la validación cruzada $\boldsymbol{k}$-fold. Por lo general, dadas estas consideraciones, uno realiza una

validación cruzada de k-fold utilizando $k = 5$ o $k = 10$, ya que se ha demostrado empíricamente que estos valores producen estimaciones de tasa de error de prueba que no sufren sesgos excesivamente altos ni una varianza muy alta.

Validación cruzada en problemas de clasificación [5.1.5]

Hasta ahora, en este capítulo, hemos ilustrado el uso de la validación cruzada en la configuración de regresión donde el resultado Y es cuantitativo, y por lo tanto, hemos utilizado MSE para cuantificar el error de prueba. Pero la validación cruzada también puede ser un enfoque muy útil en la configuración de clasificación cuando Y es cualitativo. En esta configuración, la validación cruzada funciona tal como se describió anteriormente en este capítulo, excepto que en lugar de usar MSE para cuantificar el error de la prueba, en cambio usamos el número de observaciones mal clasificadas. Por ejemplo, en la configuración de clasificación, la tasa de error LOOCV toma la forma

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} Err_i \qquad (5.4)$$

Donde $Err_i = I(y_i \neq \hat{y}_i)$. La tasa de error de la CV k-fold y las tasas de error del conjunto de validación se definen de manera análoga.

Como ejemplo, ajustamos varios modelos de regresión logística en los datos de clasificación bidimensional que se muestran en la Figura 2.13. En el panel superior izquierdo de la Figura 5.7, la línea sólida negra muestra el límite de decisión estimado resultante de ajustar un modelo de regresión logística estándar a este conjunto de datos. Dado que se trata de datos simulados, podemos calcular la verdadera tasa de error de prueba, que toma un valor de 0.201 y, por lo tanto, es sustancialmente mayor que la tasa de error de Bayes de 0.133. Es evidente que la regresión logística no tiene la flexibilidad suficiente para modelar el límite de decisión de Bayes en esta configuración. Podemos extender fácilmente la regresión logística para obtener un límite de decisión no lineal mediante el uso de funciones polinomiales de los predictores, como hicimos en la configuración de regresión en la Sección 3.3.2. Por ejemplo, podemos ajustar un modelo de regresión logística cuadrática, dado por

$$Log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 \qquad (5.5)$$

El panel superior derecho de la Figura 5.7 muestra el límite de decisión resultante, que ahora está curvado. Sin embargo, la tasa de error de prueba ha mejorado solo ligeramente, hasta 0.197. Una mejora mucho mayor es evidente en el panel inferior izquierdo de la Figura 5.7, en el que hemos ajustado un modelo de regresión logística que involucra polinomios cúbicos de los predictores. Ahora la tasa de error de prueba ha disminuido a 0.160. Ir a un polinomio quártico (abajo a la derecha) aumenta ligeramente el error de la prueba.

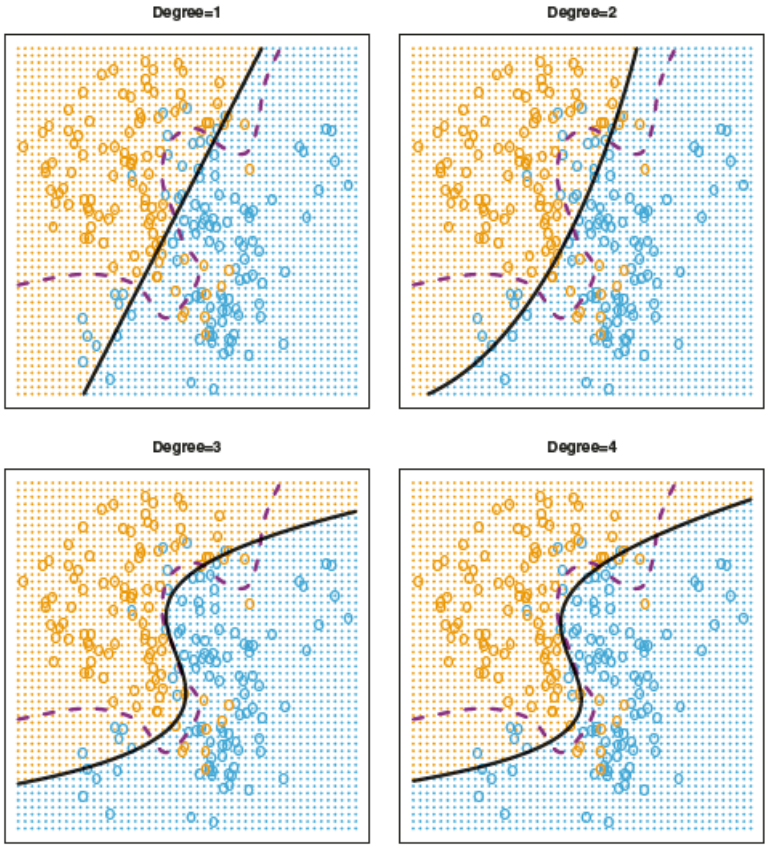

Figura 5.7: La regresión logística se ajusta a los datos de clasificación bidimensionales mostrados en la Figura 2.13. El límite de decisión de Bayes se representa mediante una línea discontinua de color púrpura. Los límites de decisión estimados de las regresiones logísticas lineales, cuadráticas, cúbicas y quárticas (grados 1-4) se muestran en negro. Las tasas de error de prueba para los cuatro ajustes de regresión logística son, respectivamente, 0.201, 0.197, 0.160 y 0.162, mientras que la tasa de error de Bayes es 0.133.

En la práctica, para datos reales, el límite de decisión de Bayes y las tasas de error de prueba son desconocidos. Entonces, ¿cómo podríamos decidir entre los cuatro modelos de regresión logística que se muestran en la Figura 5.7? Podemos utilizar la validación cruzada para tomar esta decisión. El panel de la izquierda de la Figura 5.8 muestra en negro las tasas de error de la CV 10-fold que resultan de ajustar diez modelos de regresión logística a los datos, usando funciones polinomiales de los predictores hasta el décimo orden. Los verdaderos errores de prueba se muestran en marrón, y los errores de entrenamiento se muestran en azul. Como hemos visto anteriormente, el error de entrenamiento tiende a disminuir a medida que aumenta la flexibilidad del ajuste. (La figura indica que, aunque la tasa de errores de entrenamiento no disminuye de manera monótona, tiende a disminuir en general a medida que aumenta la complejidad del modelo). En contraste, el error de prueba muestra una forma de U característica. La tasa de error de la CV de 10-fold proporciona una buena aproximación a la tasa de error de prueba. Si bien subestima un poco la tasa de error, alcanza un mínimo cuando se usan polinomios de cuarto orden, que está muy cerca del mínimo de la curva de prueba, que ocurre cuando se usan polinomios de tercer orden. De hecho, el uso de polinomios de cuarto orden probablemente conduciría a un buen rendimiento del conjunto

de pruebas, ya que la verdadera tasa de error de prueba es aproximadamente la misma para los polinomios de tercer, cuarto, quinto y sexto orden.

El panel de la derecha de la Figura 5.8 muestra las mismas tres curvas que utilizan el enfoque de ***KNN*** para la clasificación, en función del valor de ***K*** (que en este contexto indica el número de vecinos utilizados en el clasificador de ***KNN***, en lugar del número de la CV de ***k***-fold utilizadas). Nuevamente, la tasa de errores de entrenamiento disminuye a medida que el método se vuelve más flexible, por lo que vemos que la tasa de errores de entrenamiento no se puede usar para seleccionar el valor óptimo para ***K***. Aunque la curva de error de validación cruzada subestima ligeramente la tasa de errores de prueba, toma un mínimo muy cercano al mejor valor para ***K***.

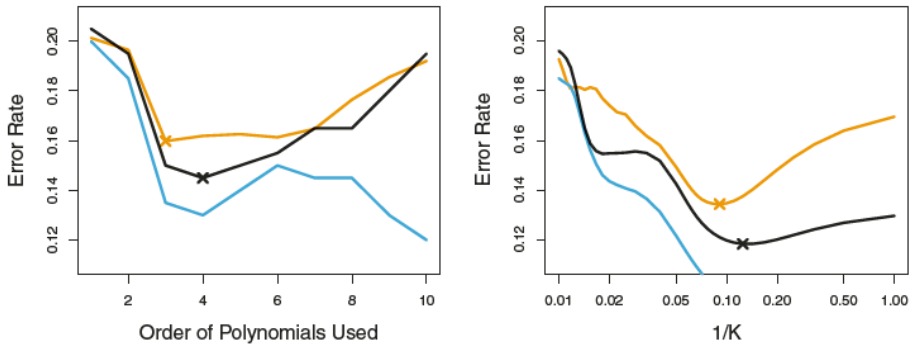

Figura 5.8: Error de prueba (marrón), error de entrenamiento (azul) y error de la CV de 10-fold (negro) en los datos de clasificación bidimensionales que se muestran en la Figura 5.7. Izquierda: regresión logística utilizando funciones polinomiales de los predictores. El orden de los polinomios utilizados se muestra en el eje x. A la derecha: el clasificador KNN con diferentes valores de K, el número de vecinos utilizados en el clasificador KNN.

El bootstrap [5.2]

El **bootstrap** es una herramienta estadística ampliamente aplicable y extremadamente poderosa que se puede usar para cuantificar la incertidumbre asociada con un estimador dado o un método de aprendizaje estadístico. Como un ejemplo simple, el **bootstrap** puede usarse para estimar los errores estándar de los coeficientes de un ajuste de regresión lineal. En el caso específico de la regresión lineal, esto no es particularmente útil, ya que en el Capítulo 3 vimos que el software estadístico estándar, como R, genera dichos errores estándar automáticamente. Sin embargo, el poder del programa de arranque reside en el hecho de que se puede aplicar fácilmente a una amplia gama de métodos de aprendizaje estadístico, incluidos algunos para los cuales, por lo demás, es difícil obtener una medida de variabilidad y el software estadístico no la obtiene automáticamente.

En esta sección, ilustramos el **bootstrap** en un ejemplo de juguete en el que deseamos determinar la mejor asignación de inversión bajo un modelo simple. En la Sección 5.3 exploramos el uso de la rutina de carga para evaluar la variabilidad asociada con los coeficientes de regresión en un ajuste de modelo lineal.

Supongamos que deseamos invertir una suma fija de dinero en dos activos financieros que producen rendimientos de $\boldsymbol{X}$ e $\boldsymbol{Y}$, respectivamente, donde $\boldsymbol{X}$ e $\boldsymbol{Y}$ son cantidades aleatorias. Invertiremos una fracción $\boldsymbol{\alpha}$ de nuestro dinero en $\boldsymbol{X}$ he invertiremos el $\boldsymbol{1-\alpha}$ restante en $\boldsymbol{Y}$. Dado que existe una variabilidad asociada con los rendimientos de estos dos activos, deseamos elegir $\boldsymbol{1-\alpha}$ para minimizar el riesgo total o la variación de nuestra inversión. En otras palabras, queremos minimizar $\boldsymbol{Var(\alpha X + (1-\alpha)Y)}$. Se puede demostrar que el valor que minimiza el riesgo está dado por

$$\boldsymbol{\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}}} \qquad \boldsymbol{(5.6)}$$

Donde $\boldsymbol{\sigma_X^2 = Var(X)}$, $\boldsymbol{\sigma_Y^2 = Var(Y)}$, y $\boldsymbol{\sigma_{XY} = Cov(X,Y)}$.

En realidad, las cantidades $\boldsymbol{\sigma_X^2}$, $\boldsymbol{\sigma_Y^2}$ y $\boldsymbol{\sigma_{XY}}$ son desconocidas. Podemos calcular estimaciones para estas cantidades, $\boldsymbol{\hat{\sigma}_X^2}$, $\boldsymbol{\hat{\sigma}_Y^2}$ y $\boldsymbol{\hat{\sigma}_{XY}}$, usando un conjunto de datos que contiene mediciones pasadas para $\boldsymbol{X}$ e $\boldsymbol{Y}$. Luego podemos estimar el valor de $\boldsymbol{\alpha}$ que minimiza la variación de nuestra inversión usando

$$\boldsymbol{\hat{\alpha} = \frac{\hat{\sigma}_Y^2 - \hat{\sigma}_{XY}}{\hat{\sigma}_X^2 + \hat{\sigma}_Y^2 - 2\hat{\sigma}_{XY}}} \qquad \boldsymbol{(5.7)}$$

La Figura 5.9 ilustra este enfoque para estimar $\boldsymbol{\alpha}$ en un conjunto de datos simulados. En cada panel, simulamos 100 pares de rendimientos para las inversiones $\boldsymbol{X}$ e $\boldsymbol{Y}$. Utilizamos estos rendimientos para estimar $\boldsymbol{\sigma_X^2}$, $\boldsymbol{\sigma_Y^2}$ y $\boldsymbol{\sigma_{XY}}$, que luego sustituimos en (5.7) para obtener estimaciones de $\boldsymbol{\alpha}$. El valor de $\boldsymbol{\hat{\alpha}}$ resultante de cada conjunto de datos simulados varía de 0.532 a 0.657.

Es natural desear cuantificar la precisión de nuestra estimación de $\boldsymbol{\alpha}$. Para estimar la desviación estándar de $\boldsymbol{\hat{\alpha}}$, repetimos el proceso de simular 100 observaciones pareadas de $\boldsymbol{X}$ e $\boldsymbol{Y}$, y estimar $\boldsymbol{\alpha}$ usando (5.7), 1,000 veces. Por lo tanto, obtuvimos 1,000 estimaciones para $\boldsymbol{\alpha}$, que podemos llamar $\boldsymbol{\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{1000}}$. El panel de la izquierda de la Figura 5.10 muestra un histograma de las estimaciones resultantes. Para estas simulaciones, los parámetros se establecieron en $\boldsymbol{\sigma_X^2 = 1}$, $\boldsymbol{\sigma_Y^2 = 1.25}$ y $\boldsymbol{\sigma_{XY} = 0.5}$, por lo que sabemos que el valor verdadero de $\boldsymbol{\alpha}$ es 0.6. Indicamos este valor utilizando una línea vertical sólida en el histograma. La media sobre todas las 1,000 estimaciones para $\boldsymbol{\alpha}$ es

$$\boldsymbol{\bar{\alpha} = \frac{1}{1000}\sum_{r=1}^{1000} \hat{\alpha}_r = 0.5996}$$

Muy cerca de $\boldsymbol{\alpha = 0.6}$, y la desviación estándar de las estimaciones es

$$\boldsymbol{\sqrt{\frac{1}{1000-1}\sum_{r=1}^{1000}(\hat{\alpha}_r - \bar{\alpha})^2}} = 0.083$$

Esto nos da una muy buena idea de la precisión de $\hat{\alpha}$: $SE(\hat{\alpha}) \approx 0.083$. En términos generales, para una muestra aleatoria de la población, esperaríamos que $\hat{\alpha}$ difiera de α en aproximadamente 0.08, en promedio.

Sin embargo, en la práctica, el procedimiento para estimar la $SE(\hat{\alpha})$ descrito anteriormente no se puede aplicar, porque para datos reales no podemos generar nuevas muestras de la población original. Sin embargo, el enfoque bootstrap nos permite usar una computadora para emular el proceso de obtención de nuevos conjuntos de muestras, de modo que podamos estimar la variabilidad de $\hat{\alpha}$ sin generar muestras adicionales. En lugar de obtener repetidamente conjuntos de datos independientes de la población, en cambio obtenemos conjuntos de datos distintos mediante el muestreo repetido de observaciones del conjunto de datos original.

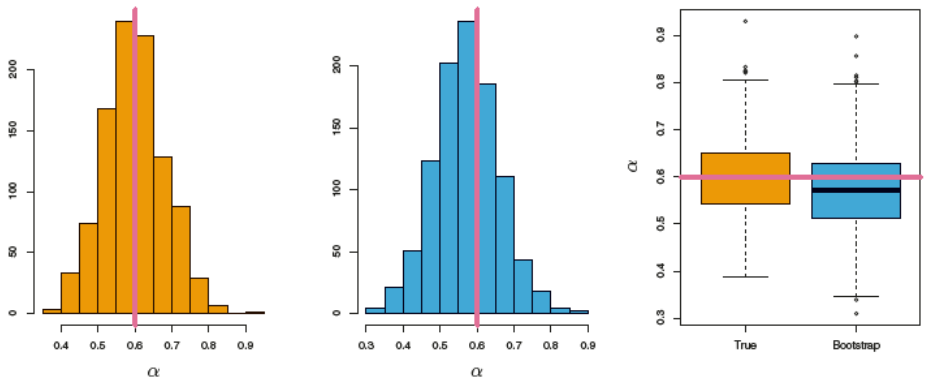

Figura 5.10: Izquierda: un histograma de las estimaciones de α obtenidas al generar 1,000 conjuntos de datos simulados a partir de la población real. Centro: un histograma de las estimaciones de α obtenidas de 1,000 muestras de arranque de un solo conjunto de datos. Derecha: las estimaciones de α mostradas en los paneles izquierdo y central se muestran como diagramas de caja. En cada panel, la línea rosa indica el valor verdadero de α.

Este enfoque se ilustra en la Figura 5.11 en un conjunto de datos simple, que llamamos Z, que contiene solo $n = 3$ observaciones. Seleccionamos aleatoriamente n observaciones del conjunto de datos para generar un conjunto de datos **bootstrap**, Z^{*1}. El muestreo se realiza con reemplazo, lo que significa que la misma observación puede ocurrir más de una vez en el conjunto de datos de arranque. En este ejemplo, Z^{*1} contiene la tercera observación dos veces, la primera observación una vez y ninguna instancia de la segunda observación. Tenga en cuenta que si una observación está contenida en Z^{*1}, se incluyen tanto sus valores X como Y. Podemos usar Z^{*1} para producir una nueva estimación **bootstrap** para α, que llamamos $\hat{\alpha}^{*1}$. Este procedimiento se repite B veces para un gran valor de B, para producir B diferentes conjuntos de datos de arranque, $Z^{*1}, Z^{*2}, \ldots, Z^{*B}$, y B correspondientes estimaciones α, $\hat{\alpha}^{*1}, \hat{\alpha}^{*2}, \ldots, \hat{\alpha}^{*B}$. Podemos calcular el error estándar de estas estimaciones de arranque usando la fórmula

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{1-B}\sum_{r=1}^{B}\left(\hat{\alpha}^{*r} - \frac{1}{B}\sum_{r'=1}^{B}\hat{\alpha}^{*r'}\right)^2} \qquad (5.8)$$

Esto sirve como una estimación del error estándar de $\hat{\alpha}$ estimado a partir del conjunto de datos original.

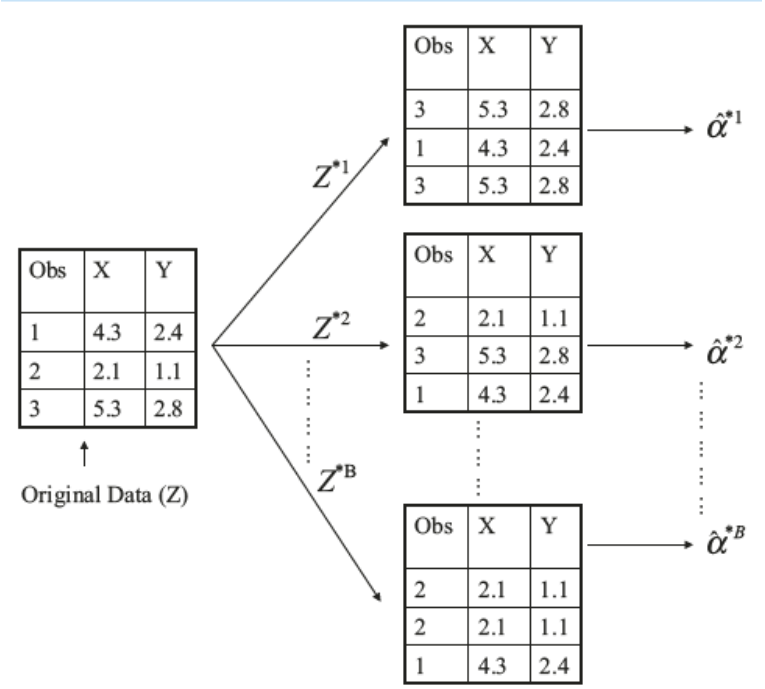

Figura 5.11: Una ilustración gráfica del enfoque **bootstrap** en una muestra pequeña que contiene $n = 3$ observaciones. Cada conjunto de datos de arranque contiene n observaciones, muestreadas con reemplazo del conjunto de datos original. Cada conjunto de datos de arranque se utiliza para obtener una estimación de α.

El enfoque de **bootstrap** se ilustra en el panel central de la Figura 5.10, que muestra un histograma de 1,000 estimaciones de **bootstrap** de α, cada una calculada utilizando un conjunto de datos de bootstrap distinto. Este panel se construyó sobre la base de un único conjunto de datos y, por lo tanto, podría crearse utilizando datos reales.

Tenga en cuenta que el histograma es muy similar al panel de la izquierda, que muestra el histograma idealizado de las estimaciones de α obtenidas al generar 1,000 conjuntos de datos simulados a partir de la población real. En particular, la estimación de **bootstrap** $SE(\hat{\alpha})$ de (5.8) es 0.087, muy cerca de la estimación de 0.083 obtenida utilizando 1.000 conjuntos de datos simulados. El panel de la derecha muestra la información en los paneles central e izquierdo de una manera diferente, a través de diagramas de caja de las estimaciones para α obtenidas mediante la generación de 1,000 conjuntos de datos simulados a partir de la población real y utilizando el enfoque **bootstrap**. Nuevamente, los diagramas de caja son bastante similares entre sí, lo que indica que el enfoque **bootstrap** se puede usar para estimar efectivamente la variabilidad asociada con $\hat{\alpha}$.