2.1. Jacknife

Esta técnica fue propuesta por [3] y consiste en la siguiente observación.

Se puede probar que muchos de los estimadores tiene la propiedad que

$$\text{Sesgo}\,(T_n) = \frac{a}{n} + \frac{b}{n^2} + O\left(\frac{1}{n^3}\right) \tag{2.1}$$

para algún a and b.

Es sesgado

Por ejemplo $\sigma^2 = \mathrm{Var}\,(X_i)$ y sea $\boxed{\widehat{\sigma}^2_n = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2.}$ Entonces,

$$\mathbb{E}\left(\widehat{\sigma}^2_n\right) = \boxed{\frac{n-1}{n}\sigma^2}$$

por lo tanto

$$\boxed{\text{Sesgo} = -\frac{\sigma^2}{n}}$$

Por lo tanto en este caso $a = -\sigma^2$ y $b = 0$.

Defina $\boxed{T_{(-i)}}$ como el estimador $\boxed{T_n}$ pero eliminando el i-ésimo término.

Es claro que en este contexto, se tiene que

$$\text{Sesgo}\left(T_{(-i)}\right) = \frac{a}{n-1} + \frac{b}{(n-1)^2} + O\left(\frac{1}{(n-1)^3}\right) \tag{2.2}$$

Laboratorio 2.1.1

Una forma fácil de construir los $T_{(-i)}$ es primero replicando la matriz de datos múltiple veces usando el producto de kronecker

```
n <- length(x)
jackdf <- kronecker(matrix(1, 1, n), x)

kable(jackdf[1:10, 1:10])
```

15	15	15	15	15	15	15	15	15	15
60	60	60	60	60	60	60	60	60	60
45	45	45	45	45	45	45	45	45	45
10	10	10	10	10	10	10	10	10	10
30	30	30	30	30	30	30	30	30	30
60	60	60	60	60	60	60	60	60	60
45	45	45	45	45	45	45	45	45	45
10	10	10	10	10	10	10	10	10	10
25	25	25	25	25	25	25	25	25	25
15	15	15	15	15	15	15	15	15	15

Y luego se elimina la diagonal

```
diag(jackdf) <- NA

kable(jackdf[1:10, 1:10])
```

~~NA~~	15	15	15	15	15	15	15	15	15
60	~~NA~~	60	60	60	60	60	60	60	60
45	45	~~NA~~	45	45	45	45	45	45	45
10	10	10	~~NA~~	10	10	10	10	10	10
30	30	30	30	~~NA~~	30	30	30	30	30
60	60	60	60	60	~~NA~~	60	60	60	60
45	45	45	45	45	45	~~NA~~	45	45	45
10	10	10	10	10	10	10	NA	10	10
25	25	25	25	25	25	25	25	NA	25
15	15	15	15	15	15	15	15	15	NA

Cada columna contiene toda la muestra excepto el i-ésimo elemento.

Solo basta estimar la media de cada columna:

```
T_i <- apply(jackdf, 2, var, na.rm=TRUE)


kable(T_i[1:10])
```

x
429.7098
428.1905
429.6023
429.3756
430.1087
428.1905
429.6023
429.3756
430.0764
429.7098

$T_{-(1)}$

$T_{(2)}$

$\vdots$

$T_{(-n)}$

Definamos el sesgo *jackife* como

$$b_{jack} = (n-1)(\overline{T}_n - T_n)$$

donde

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T_{(-i)}$$

Laboratorio 2.1.2

En nuestro caso tendríamos lo siguiente:

```
(bjack <- (n-1)*(mean(T_i) - Tn))

## [1] 0
```

es decir, que los `T_i` generan estimadores de `T_n` que contienen el mismo sesgo.

Observe que b_{jack} tiene la siguiente propiedad

$$\frac{1}{n}\sum T_{(-i)}$$

$$\begin{aligned}
\mathbb{E}\left(b_{\text{jack}}\right) &= (n-1)\left(\mathbb{E}\left[\overline{T}_n\right] - \mathbb{E}\left[T_n\right]\right) \\
&= (n-1)\left(\mathbb{E}\left[\overline{T}_n\right] - \theta + \theta - \mathbb{E}\left[T_n\right]\right) \\
&= (n-1)\left(\text{Sesgo}\left(\overline{T}_n\right) - \text{Sesgo}\left(T_n\right)\right) \\
&= (n-1)\left[\left(\frac{1}{n-1} - \frac{1}{n}\right)a + \left(\frac{1}{(n-1)^2} - \frac{1}{n^2}\right)b + O\left(\frac{1}{n^3}\right)\right] \\
&= \frac{a}{n} + \boxed{\frac{(2n-1)b}{n^2(n-1)}} + O\left(\frac{1}{n^2}\right) \\
&= \text{Sesgo}\left(T_n\right) + O\left(\frac{1}{n^2}\right)
\end{aligned}$$

$$\approx \frac{b}{n^2}$$

Nota 2.1.3

Es decir, en general, el estimador b_{jack} aproxima correctamente $\text{Sesgo}\left(T_n\right)$ hasta con un error del n^{-2}.

Podemos usar los T_i para generar muestras adicionales para estimar el parámetro θ.

En este caso defina el siguiente estimador:

$$\widetilde{T}_i = nT_n - (n-1)T_{(-i)}.$$

Nota 2.1.4

A $\widetilde{T}_i$ se le llaman **pseudo-valores** y representa el aporte o peso que tiene la variable X_i para estimar T_n.

Tarea 2.1.5

Usado un cálculo similar para el b_{jack} pruebe que

$$\text{Sesgo}\left(T_{\text{jack}}\right) = -\frac{b}{n(n-1)} + O\left(\frac{1}{n^2}\right) = O\left(\frac{1}{n^2}\right).$$

¿Qué conclusión se obtiene de este cálculo?

Laboratorio 2.1.6

Los pseudo-valores se estiman de forma directa como,

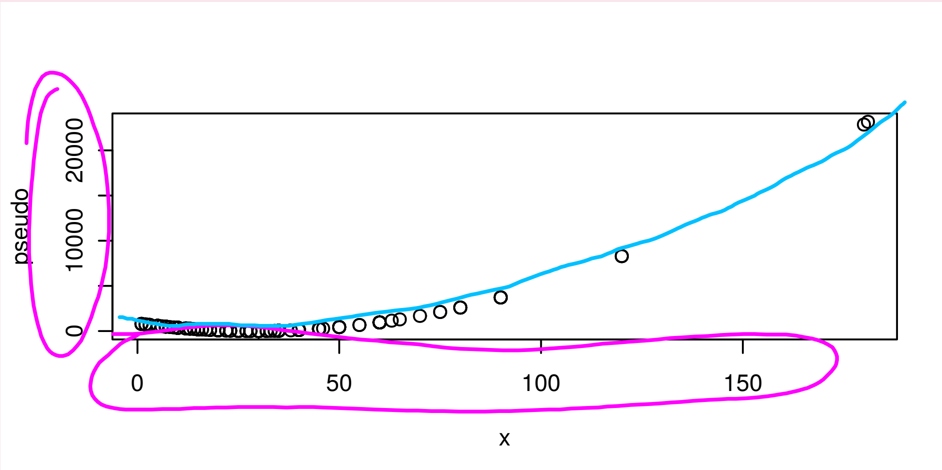
```
pseudo <- n * Tn - (n - 1) * T_i

pseudo[1:10]

##  [1] 199.02972209 957.16225222 252.64417993 365.79679037  -0.06666345
##  [6] 957.16225222 252.64417993 365.79679037  16.09799519 199.02972209
```

Lo importante acá es notar la similitud que tiene con los datos reales,

```r
plot(x = x, y = pseudo)
```



Con estos pseudo-valores, es posible estimar la media y la varianza de T_n con sus respectivos estimadores:

$$T_{\text{jack}} = \frac{1}{n}\sum_{i=1}^{n} \widetilde{T}_i$$

donde

$$v_{jack} = \frac{\sum_{i=1}^{n}\left(\widetilde{T}_i - \frac{1}{n}\sum_{i=1}^{n}\widetilde{T}_i\right)^2}{n(n-1)}.$$

Nota 2.1.7

Sin embargo, se puede demostrar fácilmente que se pueden usar pseudovalores para construir una prueba normal de hipótesis. Dado que cada pseudovalor es independiente e idénticamente distribuido (iid), se deduce que su promedio se ajusta a una distribución normal a medida que el tamaño de la muestra aumenta. El promedio de los pseudovalores es solo T_{jack} y el valor esperado de ese promedio, debido a la construcción a la imparcialidad del estimador, es el parámetro bajo investigación, θ. Por lo tanto, tenemos que

$$\boxed{\frac{\sqrt{n}\left(T_{jack} - \theta\right)}{\sqrt{v_{jack}}} \rightarrow N(0,1).}$$

Laboratorio 2.1.8

```
(Tjack <- mean(pseudo))

## [1] 429.2484
```

```
(Vjack <- var(pseudo, na.rm = TRUE))

## [1] 2701991
```

```
(sdjack <- sqrt(Vjack))

## [1] 1643.774
```

```
(z <- qnorm(1 - 0.05 / 2))

## [1] 1.959964
```

```
c(Tjack - z * sdjack / sqrt(n),
Tjack + z * sdjack / sqrt(n))

## [1] 285.1679 573.3289
```