Hugo Andres Sabogal Perez 202013538
Natalia Plata Ángel 202013152

Taller de programación en R: Taller 1

Facultad de Economía, Universidad de los Andes

Agosto 16/2024

1) Primer Punto:

1.1) Definan una semilla para trabajar durante el script. Respondan: ¿Por qué es importante definir una semilla?

Solución:

Definir una semilla es fundamental porque permite replicar con exactitud los procesos de generación de datos y números aleatorios. En particular, la función de una semilla es garantizar que, dado un valor específico, se generen siempre los mismos números aleatorios. Esto hace que establecer una semilla sea crucial para la reproducibilidad de los resultados en un código. Para este taller se fijo la semilla con un valor de 55.

1.2) Primero, creen los siguientes 4 vectores: uno que varíe de 1 en 1 desde el 1 hasta el 50. Este corresponde al identificador del individuo. Segundo, un vector de clase *int* llamado *edad* que se distribuya de forma uniforme entre el intervalo 5 a 50. Tercero, un vector que repita el carácter "años" y un vector de 50 nombres propios aleatorios de personas. Todas las cuatro (4) listas deben tener el mismo tamaño.

Solución:

A continuación, se mencionarán las funciones con las que se crearon cada uno de los 4 vectores:

1. **Identificador:** Se utilizo la función base de R *1:50* la cual crea una secuencia de número del 1 al 50.
2. **Edad:** Se utilizo la función *runif()* con mínimo 5 y máximo 50, posteriormente se aproximó al entero más cercano hacia abajo para que la secuencia sea clase int.
3. **Años:** Se utilizo la función *rep() para repetir la cadena de caracteres "años" en cada uno de las 50 entradas del vector.*
4. **Nombres:** Se digitaron manualmente 50 nombres aleatorios.

La solución de este numeral se encuentra entre las líneas 7 y 20 del script.

1.3) Creen un vector en el que cada elemento j sea la concatenación de los elementos j de los vectores del punto anterior. Ordenen y/o agreguen caracteres a cada elemento de la lista para que se consolide una oración con orden semántico que refleje la edad del individuo. Para esto, utilicen la función *pasteo()*

Solución:

Las instrucciones de este punto no requieren argumentación ni interpretación de resultados, ni tampoco responder a una pregunta en forma de texto. Por lo tanto, la solución se limita únicamente al script de R asociado al taller. La solución de este numeral se encuentra entre las líneas 21 y 24 del script.

1.4) Usando un loop realicen un código que presente (print) la concatenación lógica de la edad, nombre y años de cada uno de los individuos dentro de las listas, pero únicamente si el nombre del individuo empieza por una letra distinta de J y la edad sea distinta de un número par. Es decir, el resultado debería ser algo similar a: "Camilo tiene 29 años"

Solución

Las instrucciones de este punto no requieren argumentación ni interpretación de resultados, ni tampoco responder a una pregunta en forma de texto. Por lo tanto, la solución se limita únicamente al script de R asociado al taller. La solución de este numeral se encuentra entre las líneas 25 y 32 del script.

1.5) Programen una función que tome como entrada un vector con valores numéricos y que su output sea el promedio de los valores del vector y la desviación estándar asociada a la misma muestra. Usando esta función respondan: ¿Cuál es la edad promedio de su lista? ¿Cuál es la desviación estándar?

Solución:

Hugo Andres Sabogal Perez 202013538
Natalia Plata Ángel 202013152

Para la solución se utilizaron las funciones base de R *sum()* y *lenght()*. *Con la finalidad de calcular la media y la desviación estándar de la siguiente forma:*

$$media = \frac{\sum_{i=1}^{50} x_i}{n}$$

$$Desviación\ estadndar = \sqrt{\frac{\sum_{i=1}^{50}(x_i - media)^2}{n-1}}$$

La edad promedio de la muestra es de 27,7 años, con una desviación estándar de 13,8 años. Esto indica que la mayoría de las edades en la muestra se agrupan en torno a los 27 años, con una variación promedio de 13 años respecto a la media.

El código referente a este numeral se encuentra desde la línea 33 hasta la 43.

1.6) Programen una función que tome como entrada un vector con valores numéricos y estandarice los valores. Es decir, que los transforme a una normal estándar. Apliquen las funciones que desarrollaron en el literal 1.5) dentro de la función que propongan en este literal.

Para que una secuencia de valores converja a la distribución normal estándar es necesario estandarizar la variable. Para ello se realiza el siguiente calculo:

$$edadstd_i = \frac{edad_i - \overline{edad}}{std(edad_i)}$$

Donde $std(edad_i)$ es la desviación estándar de la edad y $\overline{edad}$ el promedio.

La solución de este numeral se encuentra entre las líneas 44 y 54 del script.

1.7) Apliquen la función programada en el literal 1.6) para crear un vector con la edad estandarizada. Llamen este nuevo vector como $\underline{edadstd_i}$

Las instrucciones de este punto no requieren argumentación ni interpretación de resultados, ni tampoco responder a una pregunta en forma de texto. Por lo tanto, la solución se limita únicamente al script de R asociado al taller. La solución de este numeral se encuentra entre las líneas 53 y 57 del script.

1.8) Por otra parte, generen una lista llamada outcomes_nominales. Esta lista contendrá 3 vectores de 50 observaciones cada uno con los outcomes de interés: salario, índice de salud, experiencia laboral. Para esto, cada una de estas variables tiene que seguir el siguiente proceso generador de datos.

$$(1)\ \boldsymbol{salario_i = 2 + 3edadstd_i + e_i}$$

$$(2)\ \boldsymbol{salud_i = 5 - 3edadstd_{i_i} - edadstd_i^{\,2} + e_i}$$

$$(3)\boldsymbol{exp_i = 2 + e_i}$$

Para todos los procesos e_i corresponde a un error proveniente de una distribución normal con media 0 y varianza 1.

Las instrucciones de este punto no requieren argumentación ni interpretación de resultados, ni tampoco responder a una pregunta en forma de texto. Por lo tanto, la solución se limita únicamente al script de R asociado al taller. La solución de este numeral se encuentra entre las líneas 58 y 65 del script.

1.9) Creen una función que permita convertir un vector en una matriz para la estimación de una regresión lineal simple. Para esto, la función debe tomar como input un vector y debe tener como output una matriz X que concatene los datos de este vector y un vector de 1´s.

Solución:

Para crear una función que retorne la matriz X solicitada se utilizaron las funciones *rep()* y *matrix()* de R base. De modo que al aplicar la función se obtendría una matriz de la siguiente forma:

Hugo Andres Sabogal Perez 202013538
Natalia Plata Ángel 202013152

$$X = \begin{bmatrix} 1 & edadstd_1 \\ 1 & edadstd_2 \\ \vdots & \vdots \\ 1 & edadstd_{50} \end{bmatrix}$$

La solución de este numeral se encuentra entre las líneas 66 y 73 del script.

1.10) A partir de la función anterior consoliden una matriz X con la edad de los individuos estandarizada y un vector de 1´s asociado a una constante.

Las instrucciones de este punto no requieren argumentación ni interpretación de resultados, ni tampoco responder a una pregunta en forma de texto. Por lo tanto, la solución se limita únicamente al script de R asociado al taller. La solución de este numeral se encuentra entre las líneas 74 y 77 del script.

2. Segundo Punto:

2.1) Programen una función que tome como input una matriz X y un vector y_i, posteriormente, el output debe corresponder a una estimación puntual del estimador (β_1) de Mínimos Cuadrados Ordinarios (MCO) para la muestra y a su error estándar asociado (σ_{β}).

Para solucionar este numeral, se calculó el vector de coeficientes estimados con la siguiente operación matricial:

$$\hat{\beta} = (X'X)^{-1}X'y$$

Ahora para calcular los errores estándares de los coeficientes se utilizó la matriz de varianzas y covarianzas definida de la siguiente forma:

$$Var(\hat{\beta}) = \hat{\sigma}^2 \, (X'X)^{-1}$$

Donde

$$\hat{\sigma}^2 = \frac{\varepsilon'\varepsilon}{N - K - 1}$$

Siendo N el numero de observaciones y K el numero de variables explicativas de la regresión. En este caso $N = 50, K = 1$.

Adicionalmente, es importante mencionar que el vector de los residuales ε está definido de la siguiente forma:

$$\varepsilon = Y - \hat{Y}$$

Donde

$$\hat{Y} = X\hat{\beta}$$

Una vez obtenida la matriz de varianzas y covarianzas, se extrajeron los elementos de la diagonal principal, calculándose la raíz cuadrada del segundo elemento de dicha diagonal. Este valor corresponde al error estándar del coeficiente $\hat{\beta}_1$.

La solución de este numeral se encuentra entre las líneas 80 y 90 del script.

2.2) Utilizando un loop, apliquen esta función a los diferentes outcomes en las listas de outcomes_nominales, guarden los coeficientes estimados y los errores estándar en una matriz donde la primera columna corresponde al nombre del outcome, la segunda columna al coeficiente estimado para la constante ($\boldsymbol{\beta_0}$). La tercera debe tener el coeficiente estimado asociado a correlación con la edad ($\boldsymbol{\beta_1}$) y la cuarta al error estándar ($\boldsymbol{\beta_1}$). En esta matriz, cada fila representará una estimación.

A continuación, se presentará la tabla 1, la cual corresponde a la matriz solicitada en la instrucción.

Tabla 1: Matriz resultados regresión

Outcome	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}_{\hat{\beta}_1}$
Salario	1.91	3.04	0.12
Salud	3.93	-2.76	0.19
Experiencia	1.91	0.041	0.12

Fuente: Elaboración propia, 2024.

La solución de este numeral se encuentra entre las líneas 91 y 107 del script.

Hugo Andres Sabogal Perez 202013538
Natalia Plata Ángel 202013152

2.3) Interpreten los coeficientes, en particular, comparen sus resultados con los procesos generadores de datos del punto 1.8) ¿Qué patrones encuentran?

Solución

Interpretación sobre la variable salario.

El intercepto estimado en el modelo es de 1.91, lo que sugiere que, en promedio, el salario de un individuo es de 1.91 unidades cuando no se considera la edad. Al comparar esta estimación con el parámetro poblacional, que es conocido y tiene un valor de 2, se observa una diferencia de 0.09 unidades. Esta diferencia indica un sesgo de 0.09 en la estimación, el cual podría atribuirse al hecho de que el número de observaciones no es suficientemente grande para garantizar que el estimador sea consistente.

El coeficiente de la edad estandarizada sobre el salario es de 3.04, lo que implica que un incremento de una desviación estándar en la edad se asocia con un aumento de 3.04 unidades en el salario. Además, el error estándar del coeficiente es menor que el doble de su valor, lo que proporciona evidencia estadística para afirmar que el coeficiente es significativamente diferente de cero al nivel de significancia del 5%.

Finalmente, al comparar la estimación con el parámetro poblacional, dado que la regresión se realizó sobre datos simulados, sabemos que el parámetro poblacional es 3.0, mientras que la estimación es 3.04. Por lo tanto, se concluye que la estimación es consistente al haber una diferencia mínima entre el parámetro poblacional y la estimación.

Interpretación sobre la variable salud.

El intercepto estimado en el modelo es 3.93, lo que sugiere que, en promedio, el índice de salud de un individuo es de 3.93 unidades cuando no se considera la edad. Al comparar esta estimación con el parámetro poblacional, conocido y con un valor de 5, se observa una diferencia de 1.07 unidades. Esta diferencia indica un sesgo de 1.07 en la estimación, que podría deberse a la omisión del término cuadrático de la edad estandarizada, lo que genera un sesgo por variable omitida y provoca que las estimaciones no coincidan con el parámetro poblacional.

Hugo Andres Sabogal Perez 202013538
Natalia Plata Ángel 202013152

El coeficiente de la edad estandarizada sobre el índice de salud es de -2.76, lo que implica que un incremento de una desviación estándar en la edad se asocia con una disminución de 2.76 unidades en el índice de salud. Además, el error estándar del coeficiente es menor que el doble de su valor absoluto, lo que proporciona evidencia estadística para afirmar que el coeficiente es significativamente diferente de cero al nivel de significancia del 5%.

Finalmente, al comparar la estimación con el parámetro poblacional, dado que la regresión se realizó sobre datos simulados, sabemos que el parámetro poblacional es -3.0, mientras que la estimación es -2.76. Por lo tanto, se concluye que la estimación presenta un sesgo de 0.24, atribuible a la omisión del término cuadrático de la edad.

Interpretación sobre la variable experiencia.

El intercepto estimado en el modelo es 1.91, lo que sugiere que, en promedio, la experiencia de un individuo es de 1.91 unidades cuando no se considera la edad. Al comparar esta estimación con el parámetro poblacional, que es conocido y tiene un valor de 2, se observa una diferencia de 0.09 unidades. Esta diferencia indica un sesgo de 0.09 en la estimación, el cual podría atribuirse al hecho de que el número de observaciones no es suficientemente grande para garantizar que el estimador sea consistente.

El coeficiente de la edad estandarizada sobre el índice de salud es de 0,041, lo que implica que un incremento de una desviación estándar en la edad se asocia con un aumento de 0,041 unidades en el índice de salud. Además, el error estándar del coeficiente es mayor que el doble de su valor absoluto, lo que proporciona evidencia estadística para afirmar que el coeficiente es igual a cero bajo un nivel de significancia del 5%.

Finalmente, al comparar la estimación con el parámetro poblacional, dado que la regresión se realizó sobre datos simulados, sabemos que el parámetro poblacional es 0, mientras que la estimación es 0,041. Por lo tanto, se concluye que la estimación presenta un sesgo de 0,041, dado que este sesgo es muy cercano a cero se considera que la estimación es consistente.