



PEP02 - Práctica

Alumno: Rodrigo Pereira Yañez

Rut: 16.610.470-k

12/0/2024

Primera pregunta (35 puntos).

- Para su ramo de análisis de algoritmos y estructura de datos le solicitaron crear un algoritmo que sea capaz de analizar bases de datos de una empresa con el objetivo de clasificar en percentiles los sueldos de los empleados de manera gráfica y a través de un listado con cada nombre, sueldo y cuartil del empleado. El día que presentaron los resultados, el profesor los organizo en parejas, les entregó los computadores de la Universidad (todos con la misma marca, modelo y año) y diversas bases de datos para analizar con sus algoritmos elaborados. Los resultados de rendimiento de los algoritmos (suyo y de su pareja) fueron los siguientes:

Algoritmo 1 (Creado por usted)		Algoritmo 2 (Creado por su pareja designada)	
Prueba	Tiempo (S)	Prueba	Tiempo (S)
1	3.2	1	2.2
2	3.0	2	2.7
3	2.9	3	2.6
4	3.5	4	2.5
5	2.7	5	2.3
6	3.1	6	2.0
7	3.2	7	2.5
8	2.9	8	2.2
9	3.0	9	2.1
10	2.8	10	2.7
11	3.1	11	2.6
12	3.1	12	2.1
13	3.5	13	2.5
14	3.0	14	2.0
15	2.7	15	2.3
16	2.9	16	2.2
17	3.0	17	2.2
18	3.5	18	2.1
19	3.0	19	2.6
20	2.7	20	2.4
21	2.9	21	2.0
22	3.2	22	2.6
23	3.4	23	2.8
25	3.1	25	2.5

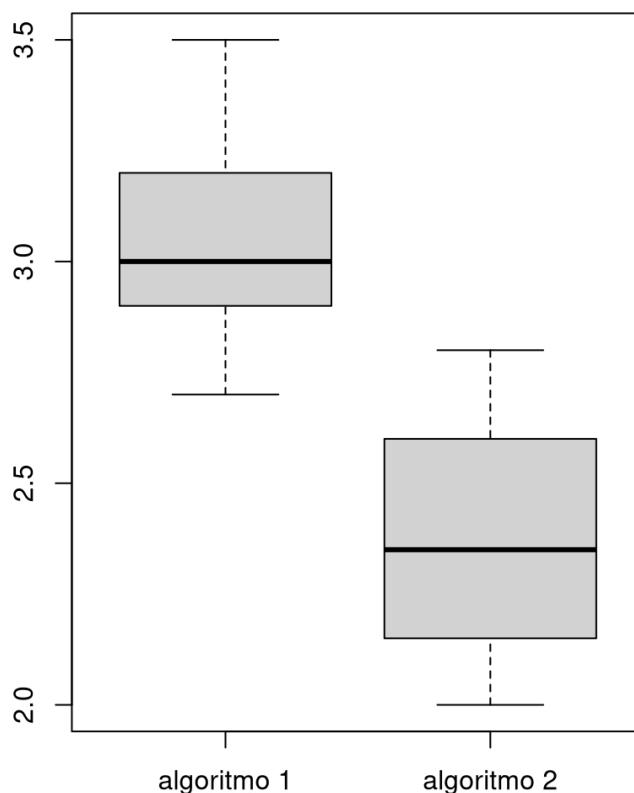
- Pregunta a) Genere en R una representación para los datos entregados, calcule los parámetros básicos (media, desviación estándar y coeficientes de variación) y explique a qué conclusión puede llegar con los resultados obtenidos. (10 puntos).



Respuesta: Los parámetros básicos de los datos de algoritmo 1 y 2 son:

	Media	Desviación Estándar	Coeficientes de variación
Algoritmo 1	3.058333	0.2412227	7.887392
Algoritmo 2	2.362500	0.2498913	10.577409

Representación Gráfica



Conclusión Basados en los datos de la media, podemos concluir que el algoritmo 1 tiene un menor rendimiento (media de tiempo mayor) con respecto al algoritmo 2, ya que la media del rendimiento obtenido es mayor (diferencia de 0.69 seg). Entre algoritmo 1 y 2 tienen igual desviación estándar, por lo tanto, los datos en ambos conjuntos se distribuyen de manera similar en torno a sus respectivas medias. En el caso de los coeficientes de variación, el algoritmo 2 tienen un mayor porcentaje, por lo que se puede concluir que el algoritmo 1 tienen mayor consistencia en los tiempos de rendimiento que



el algoritmo 2. El gráfico boxplot es la mejor forma de representar datos que se alejan de la media y se confirma visualmente (tamaño de cada boxplot) lo concluido en con los coeficientes de variación en el algoritmo 2, con respecto al algoritmo 1 y que las medias entre ambos algoritmos son diferentes.

Código R:

```
pep2_pregunta1.R pep2_pregunta2.R
Source on Save Run Source
```

```
1 library(psych)
2
3 #===========
4 # Primera pregunta (35 puntos)
5 # Para su ramo de análisis de algoritmos y estructura de datos le solicitaron crear un algoritmo que sea
6 # capaz de analizar bases de datos de una empresa con el objetivo de clasificar en percentiles los sueldos de
7 # los empleados de manera gráfica y a través de un listado con cada nombre, sueldo y cuartil del empleado.
8 # El día que presentaron los resultados, el profesor los organiza en parejas, les entrego los computadores de
9 # la Universidad (todos con misma marca, modelo y año) y diversas bases de datos para analizar con sus
10 # algoritmos elaborados. Los resultados de rendimiento de los algoritmos (suyo y de su pareja) fueron los
11 # siguientes:
12 #===========
13 #===========
14 # a) Genere en R una representación grafica para los datos entregados, calcule los parámetros básicos (media,
15 # desviación estándar y coeficientes de variación) y explique a que conclusión puede llegar con los
16 # resultados obtenidos. (10 puntos) |
17 #===========
18 #===========
19
20 alg_1 = c(3.2, 3.0, 2.9, 3.5, 2.7, 3.1, 3.2, 2.9, 3.0, 2.8,
21     3.1, 3.1, 3.5, 3.0, 2.7, 2.9, 3.0, 3.5, 3.0, 2.7,
22     2.9, 3.2, 3.4, 3.1)
23 alg_2 = c(2.2, 2.7, 2.6, 2.5, 2.3, 2.0, 2.5, 2.2, 2.1, 2.7,
24     2.6, 2.1, 2.5, 2.0, 2.3, 2.2, 2.2, 2.1, 2.6, 2.4,
25     2.0, 2.6, 2.8, 2.5)
26 datos = data.frame(alg_1, alg_2)
27
28 boxplot(alg_1,alg_2,names=c("algoritmo 1","algoritmo 2"))
29
30 valores = describe(datos, IQR=T, quant=c(.25,.50,.75))
31 print(valores)
32
33 medias = valores$mean
34 medias
35
36 desviaciones_estandar = valores$sd
37 desviaciones_estandar
38
39 coef_variacion = (desviaciones_estandar/medias)*100
40 coef_variacion
41
```

Consola R:

```
Console Terminal Background Jobs
R 4.3.3 ~/🔗
```

```
>
> #===========
> # Primera pregunta (35 puntos)
> # Para su ramo de análisis de algoritmos y estructura de datos le solicitaron crear un algoritmo que sea
> # capaz de analizar bases de datos de una empresa con el objetivo de clasificar en percentiles los sueldos de
> # los empleados de manera gráfica y a través de un listado con cada nombre, sueldo y cuartil del empleado.
> # El día que presentaron los resultados, el profesor los organiza en parejas, les entrego los computadores de
> # la Universidad (todos con misma marca, modelo y año) y diversas bases de datos para analizar con sus
> # algoritmos elaborados. Los resultados de rendimiento de los algoritmos (suyo y de su pareja) fueron los
> # siguientes:
> #===========
> #===========
> # a) Genere en R una representación grafica para los datos entregados, calcule los parámetros básicos (media,
> # desviación estándar y coeficientes de variación) y explique a que conclusión puede llegar con los
> # resultados obtenidos. (10 puntos).
> #===========
>
> alg_1 = c(3.2, 3.0, 2.9, 3.5, 2.7, 3.1, 3.2, 2.9, 3.0, 2.8,
+ 3.1, 3.1, 3.5, 3.0, 2.7, 2.9, 3.0, 3.5, 3.0, 2.7,
+ 2.9, 3.2, 3.4, 3.1)
> alg_2 = c(2.2, 2.7, 2.6, 2.5, 2.3, 2.0, 2.5, 2.2, 2.1, 2.7,
+ 2.6, 2.1, 2.5, 2.0, 2.3, 2.2, 2.2, 2.1, 2.6, 2.4,
+ 2.0, 2.6, 2.8, 2.5)
> datos = data.frame(alg_1, alg_2)
>
> boxplot(alg_1,alg_2,names=c("algoritmo 1","algoritmo 2"))
>
> valores = describe(datos, IQR=T, quant=c(.25,.50,.75))
> print(valores)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se	IQR	Q0.25	Q0.5	Q0.75
alg_1	1	24	3.06	0.24	3.00	3.05	0.15	2.7	3.5	0.8	0.40	-0.72	0.05	0.30	2.90	3.00	3.2
alg_2	2	24	2.36	0.25	2.35	2.36	0.37	2.0	2.8	0.8	0.03	-1.44	0.05	0.42	2.18	2.35	2.6

```
> medias = valores$mean
> medias
[1] 3.058333 2.362500
>
> desviaciones_estandar = valores$sd
> desviaciones_estandar
[1] 0.2412227 0.2498913
>
> coef_variacion = (desviaciones_estandar/medias)*100
> coef_variacion
[1] 7.887392 10.577409
```



Pregunta b) Evalué la normalidad de los datos entregados para cada algoritmo, con los diversos test revisados en clases, determine que test puede utilizar para comparar ambos algoritmos. Fundamente su respuesta en detalle. (10 puntos)

Respuesta: Basado en los datos entregados del algoritmo 1 y 2, se evalúa la normalidad de estos. Se aplican las siguientes pruebas de contraste:

- Shapiro-wilk,
- lillie (kolmogorov-smirnov),
- anderson-darling,
- cramér-von mises,
- pearson shi-square
- shapiro-francia test.

Donde la hipótesis nula (H_0) es: Los datos tienen una distribución normal. Con los resultados obtenidos de todos los test, con un nivel de significancia de 0,05 y el p-value de c/u es mayor que este. Por lo tanto, se acepta la hipótesis nula y podemos decir con un 95% de confianza que los datos tienen una distribución normal.

Código R:

```
42 # =====
43 # b) Evalué la normalidad de los datos entregados para cada algoritmo, con los diversos test revisados
44 # en clases, determine que test puede utilizar para comparar ambos algoritmos. Fundamente su
45 # respuesta en detalle. (10 puntos)
46 # =====
47 library("nortest")
48
49 # Test alg_1
50 shapiro.test(datos$alg_1)$p.value
51 lillie.test(datos$alg_1)$p.value
52 ad.test(datos$alg_1)$p.value
53 cvm.test(datos$alg_1)$p.value
54 pearson.test(datos$alg_1)$p.value
55 sf.test(datos$alg_1)$p.value
56
57 # Test alg_2
58 shapiro.test(datos$alg_2)$p.value
59 lillie.test(datos$alg_2)$p.value
60 ad.test(datos$alg_2)$p.value
61 cvm.test(datos$alg_2)$p.value
62 pearson.test(datos$alg_2)$p.value
63 sf.test(datos$alg_2)$p.value
64
65 # =====
76:3 # (Untitled) ▾ R Script ▾
```



Consola R:

```
Console Terminal < Background Jobs <
R 4.3.3 · ~/r

> #=====
> # b) Evalúe la normalidad de los datos entregados para cada algoritmo, con los diversos test revisados
> # en clases, determine que test puede utilizar para comparar ambos algoritmos. Fundamente su
> # respuesta en detalle. (10 puntos)
> #=====
> library("nortest")
>
> # Test alg_1
> shapiro.test(datos$alg_1)$p.value
[1] 0.08532966
> lillie.test(datos$alg_1)$p.value
[1] 0.2613568
> ad.test(datos$alg_1)$p.value
[1] 0.1238254
> cvm.test(datos$alg_1)$p.value
[1] 0.1886969
> pearson.test(datos$alg_1)$p.value
[1] 0.2466342
> sf.test(datos$alg_1)$p.value
[1] 0.1451861
>
> # Test alg_2
> shapiro.test(datos$alg_2)$p.value
[1] 0.09139051
> lillie.test(datos$alg_2)$p.value
[1] 0.08097028
> ad.test(datos$alg_2)$p.value
[1] 0.0847468
> cvm.test(datos$alg_2)$p.value
[1] 0.07989909
> pearson.test(datos$alg_2)$p.value
[1] 0.7512117
> sf.test(datos$alg_2)$p.value
[1] 0.1795012
>
```

- Pregunta c) Aplique un test correspondiente a este ejercicio que le permita determinar y comparar los rendimientos de ambos algoritmos, explique los resultados obtenidos y comparé con los resultados obtenidos en la parte a (10 puntos).

Respuesta: Basados en los resultados de las pruebas de contraste, donde se evaluó la normalidad y se concluyó que los datos siguen una distribución normal. Se determina que se debe trabajar con la Estadística Paramétrica. Determinado lo anterior sabemos que son dos muestras (Algoritmo 1 y 2) que miden rendimientos de estos algoritmos y además que son datos independientes, es decir, que no se afectan entre sí en el tiempo. Se decide aplicar el test "Welch Two Sample t-test". Cabe mencionar que para guiar la decisión se tomó como referencia el gráfico de evaluación entregado en clases, Figura 1 en ANEXO.

La hipótesis nula (H_0) para el Welch's t-test establece que: No hay diferencia entre las medias de las dos poblaciones, es decir, las medias son iguales. Con un nivel de significancia de 0,05 se aplica el test y se obtiene un p-value = 7.453e-13 que menor que el nivel de significancia, antes mencionado, por lo que se rechaza la hipótesis nula y podemos decir con un 95% de confianza que las medias del algoritmo 1 y 2 son diferentes, por lo que su rendimiento medio (tiempo) es diferente entre ellos. Lo cual tiene relación cuando en la pregunta a) se evaluó las medias de cada uno y se determinó que el algoritmo 1 tiene un menor rendimiento (media de tiempo mayor) con respecto al algoritmo 2. Y además es lo que se observa en el gráfico boxplot al ver las medias desplazadas.



Código R:

```
65 v =====
66 # c) Aplique un test correspondiente a este ejercicio que le permita determinar y comparar los
67 # rendimientos de ambos algoritmos, explique los resultados obtenidos y compare con los resultados
68 # obtenidos en la parte a (10 puntos).
69 v =====
70 t.test(alg_1, alg_2)
71 |
```

Consola R:

```
R 4.3.3 · ~ ↵
> =====
> # c) Aplique un test correspondiente a este ejercicio que le permita determinar y comparar los
> # rendimientos de ambos algoritmos, explique los resultados obtenidos y compare con los resultados
> # obtenidos en la parte a (10 puntos).
> =====
> t.test(alg_1, alg_2)

Welch Two Sample t-test

data: alg_1 and alg_2
t = 9.8147, df = 45.943, p-value = 7.453e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
0.5531198 0.8385469
sample estimates:
mean of x mean of y
3.058333 2.362500
```

- Pregunta d) Dentro de los dos grupos presentados, determine mediante algún test si los datos presentan o no homoscedasticidad y explique los resultados y conclusiones obtenidas. (5 puntos)

Respuesta: La homoscedasticidad indica que las varianzas en más de una muestra son iguales, por lo tanto, al aplicar el “LeveneTest” cuya hipótesis nula (H_0) es que: Las varianzas entre los diferentes grupos son iguales. Con un nivel de significancia de 0,05 y obtenido un $p\text{-value}=0.3533$ mayor que el nivel de significancia antes mencionado, se establece que la ' H_0 ' es válida, por lo tanto, podemos decir con un 95% de confianza que las varianzas de ambos grupos son iguales. En consecuencia, entre ambos conjunto de datos, si cumplen con los principios de homoscedasticidad y sus varianzas son homogéneas. Esto refuerza la decisión de tomar el camino por la Estadística Paramétrica, ya que una de sus condiciones, aparte de las mencionadas en la respuesta c) es que su Varianza sea Homogénea.

Código R:

```
72 v =====
73 # d) Dentro de los dos grupos presentados, determine mediante algún test si los datos presentan o no
74 # homoscedasticidad y explique los resultados y conclusiones obtenidas. (5 puntos)
75 v =====
76 library(car)
77
78 datos_test = data.frame(grupos=c(rep("alg_1", 24), rep("alg_2", 24)), valores=c(alg_1, alg_2))
79 leveneTest(valores ~ grupos, data=datos_test)
80 |
```

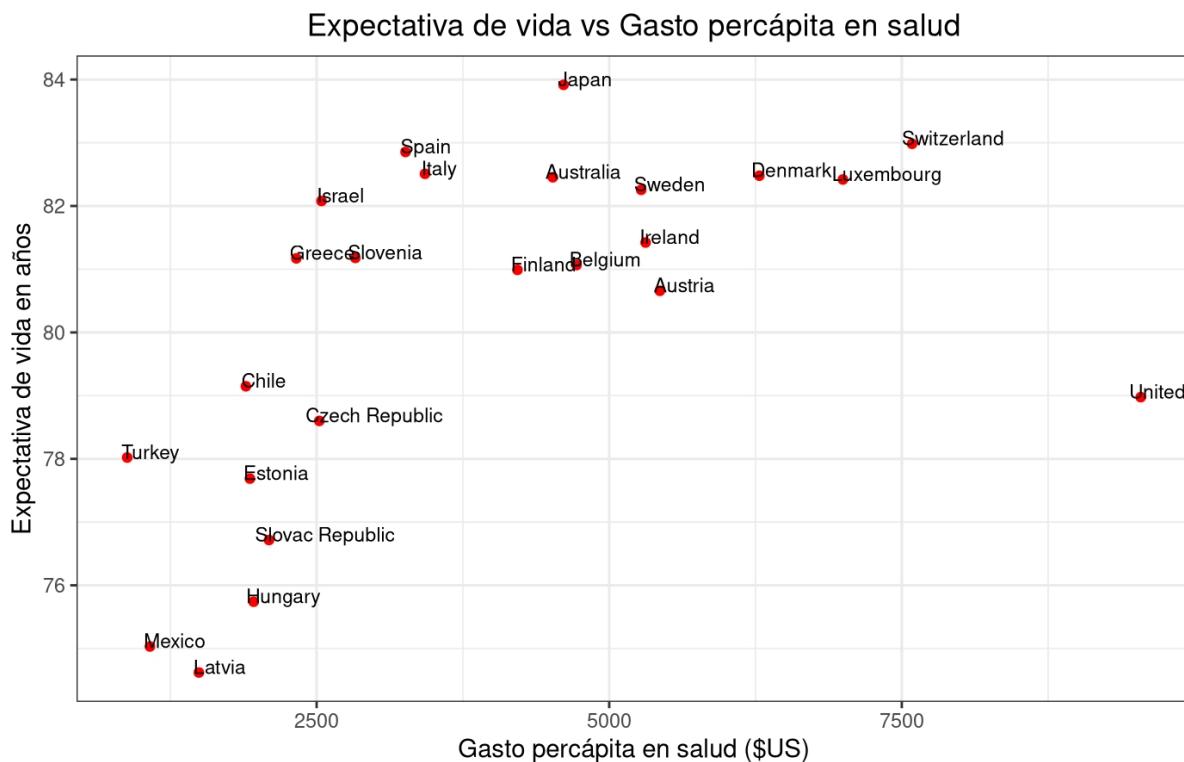
Consola R:

```
R 4.3.3 · ~ ↵
> datos_test = data.frame(grupos=c(rep("alg_1", 24), rep("alg_2", 24)), valores=c(alg_1, alg_2))
> leveneTest(valores ~ grupos, data=datos_test)
Levene's Test for Homogeneity of Variance (center = median)
          Df F value Pr(>F)
group     1 0.8792 0.3533
46
```



Segunda pregunta (25 puntos).

- 2) En este ejercicio, usted analizará un conjunto de datos que reflejan la relación que existe entre la expectativa de vida en años de las personas en algunos países del mundo y el gasto per cápita en salud (Datos correspondientes al año 2015 recolectados → <https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure>). Los datos se han organizado y se encuentran en el archivo “.csv” disponible junto al enunciado.



- Pregunta a) Con los datos presentados previamente, proponga un modelo de regresión lineal que relacione las variables presentadas (la expectativa de vida en años (variable dependiente) con el gasto en salud per cápita (variable independiente)). Evalué el modelo creado en base a los criterios que fueron revisados en clases. (10 puntos)

Respuesta: Se aplica un modelo de regresión lineal simple, ya que una sola variable independiente (gasto_salud_percapita) explica el comportamiento de la variable dependiente (expectativa_anual). Aplicado este modelo se obtienen los resultados que indican lo siguiente:

- Intercepto (c) = 7.761e+01
- pendiente (m) = 6.715e-04
- residuos “Residual standard error” (ϵ_i) = 2.3

Por lo tanto, el modelo es el siguiente:



- $y_i = (mx_i * c) + \epsilon_i \Rightarrow (\text{abajo})$
- expectativa_anual = 0.0006715 * Gasto_salud_per capita + 77.61

De este modelo podemos decir que por cada dólar adicional gastado en salud, la expectativa de vida en años aumenta en aproximadamente 0.0006715 unidades, partiendo de una expectativa de vida en años base de 77.61 cuando el gasto en salud es cero.

La evaluación del modelo se hace en base a tres parámetros:

Parámetros	Explicación / Condición / Ho	Valor Obtenido
Coeficiente de determinación (r^2)	Este coeficiente indica que proporción de la variación de la variable independiente es explicada por las variables dependientes. En este caso se busca que la condición sea mayor a 0.95.	0.3070257
p-value de la pendiente	Ho: La pendiente de la recta es cero	2e-16 ***
p-value del intercepto	Ho: El interceptó de la recta es cero.	0.00496 **

De los resultados obtenidos se determina lo siguiente:

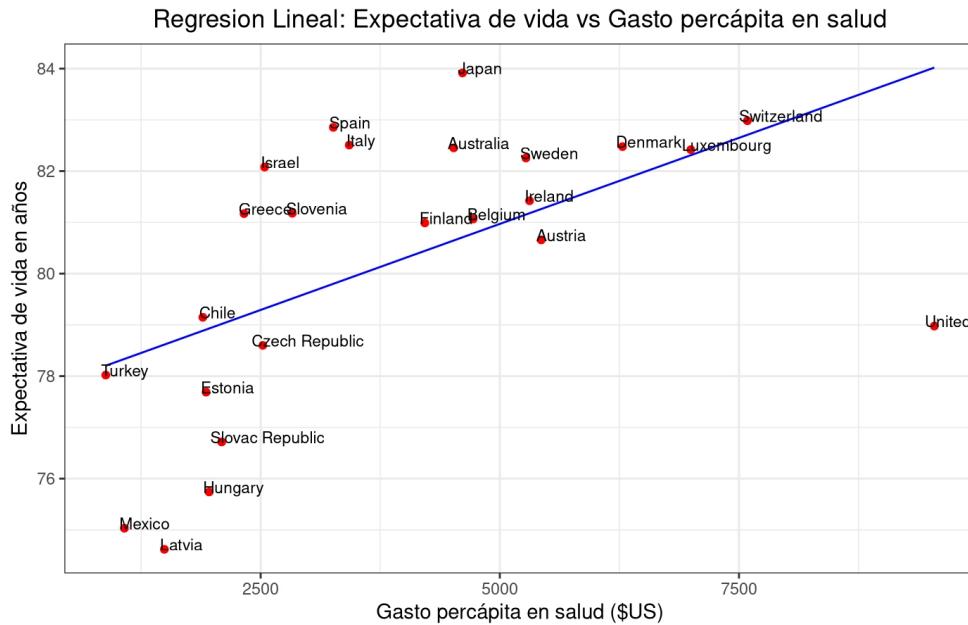
1. Como el Coeficiente de determinación es menor que el 0.95 esperado, se puede decir que el modelo responde mal al conjunto de datos evaluados.
2. El p-value de la pendiente es menor que el nivel de significancia de 0,05, por lo tanto, se rechaza la Ho y podemos decir con un 95% de confianza que la pendiente de la recta no es cero.
3. El p-value del intercepto es menor que el nivel de significancia de 0,05, por lo tanto, se rechaza la Ho y podemos decir con un 95% de confianza que él interceptó de la recta no es cero.

Finalmente, se concluye que si bien el modelo es malo ($r^2 < 0.95$) con el conjunto de datos analizados. Pero él interceptó y la pendiente están bien calculadas y son representativas. En consecuencia, el modelo es correcto y está bien calculado, pero malo para los datos analizados.

Lo anterior se puede visualizar en el gráfico donde se ven países (ej.: Estados Unidos, Lituania, Japón) que quedan muy alejados de la recta generada.



Gráfica
Regresión
Lineal



Código R:

```
1 #=====
2 # Primera pregunta (35 puntos)
3 # En este ejercicio, usted analizará un conjunto de datos que reflejan la relación que existe entre la
4 # expectativa de vida en años de las personas en algunos países del mundo y el gasto percápita en salud
5 # (Datos correspondientes al año 2015 recolectados desde https://ourworldindata.org/grapher/life-expectancy-vs-health-expenditure).
6 # Los datos se han organizado y se encuentran en el archivo ".csv" disponible junto al enunciado.
7 # Los datos se pueden graficar de la siguiente manera:
8 #=====
9
10 library(ggplot2)
11
12 datos = read.csv("Datos_pregunta2.csv",sep = ";", header = T)
13
14 grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
15   geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
16   theme_bw() + xlab("Gasto percápita en salud ($US)") + ylab("Expectativa de vida en años") +
17   ggtitle("Expectativa de vida vs Gasto percápita en salud") +
18   theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
19 plot(grafico)
20
21 #=====
22 # a) Con los datos presentados previamente, proponga un modelo de regresión lineal que relacione las
23 # variables presentadas (la expectativa de vida en años (variable dependiente) con el gasto en salud
24 # per cápita (variable independiente)). Evalúe el modelo creado en base a los criterios que fueron
25 # revisados en clases. (10 puntos)
26 #=====
27
28 # Calculando la regresión
29 regresion = lm(expectativa_anual ~ Gasto_salud_dolares, datos)
30
31 # Imprime datos de regresión
32 print(regresion)
33
34 # Imprime un resumen de la regresión
35 summary(regresion)
36
37 # Imprime el valor del r^2 del modelo
38 print(summary(regresion)$r.squared)
39
40
41 grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
42   geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
43   theme_bw() + xlab("Gasto percápita en salud ($US)") + ylab("Expectativa de vida en años") +
44   ggtitle("Regresión Lineal: Expectativa de vida vs Gasto percápita en salud") +
45   geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="blue", size=0.5, se = FALSE) +
46   theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
47
48 plot(grafico)
49
```



Consola R:

```
Console Terminal x Background Jobs x
R 4.3.3 - ~/r
> ##### 
> # Primera pregunta (35 puntos)
> # En este ejercicio, usted analizará un conjunto de datos que reflejan la relación que existe entre la
> # expectativa de vida en años de las personas en algunos países del mundo y el gasto per cápita en salud
> # (Datos correspondientes al año 2015 recolectados desde - https://ourworldindata.org/grapher/life-
> # expectancy-vs-health-expenditure).
> # Los datos se han organizado y se encuentran en el archivo ".csv" disponible junto al enunciado.
> # Los datos se pueden graficar de la siguiente manera:
> #####
>
> library(ggplot2)
>
> datos = read.csv("Datos_pregunta2.csv",sep = ";", header = T)
>
> grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
+   geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
+   theme_bw() + xlab("Gasto per cápita en salud ($US)") + ylab("Expectativa de vida en años") +
+   ggtitle("Expectativa de vida vs Gasto per cápita en salud") +
+   theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
> plot(grafico)
>
> #####
> # a) Con los datos presentados previamente, proponga un modelo de regresión lineal que relacione las
> # variables presentadas (la expectativa de vida en años (variable dependiente) con el gasto en salud
> # per cápita (variable independiente)). Evalúe el modelo creado en base a los criterios que fueron
> # revisados en clases. (10 puntos)
> #####
>
> # Calculando la regresión
> regresion = lm(expectativa_anual ~ Gasto_salud_dolares, datos)
>
> # Imprime datos de regresión
> print(regresion)

Call:
lm(formula = expectativa_anual ~ Gasto_salud_dolares, data = datos)

Coefficients:
(Intercept) Gasto_salud_dolares
7.761e+01      6.715e-04

>
> # Imprime un resumen de la regresión
> summary(regresion)

Call:
lm(formula = expectativa_anual ~ Gasto_salud_dolares, data = datos)

Residuals:
Min     1Q     Median     3Q    Max
-5.0440 -0.8353  0.2689  1.7005  3.2080

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 7.761e+01 9.545e-01 81.314 < 2e-16 ***
Gasto_salud_dolares 6.715e-04 2.151e-04   3.122 0.00496 **
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.3 on 22 degrees of freedom
Multiple R-squared: 0.307, Adjusted R-squared: 0.2755
F-statistic: 9.747 on 1 and 22 DF, p-value: 0.004962

>
> # Imprime el valor del r^2 del modelo
> print(summary(regresion)$r.squared)
[1] 0.3070257
>
> grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
+   geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
+   theme_bw() + xlab("Gasto per cápita en salud ($US)") + ylab("Expectativa de vida en años") +
+   ggtitle("Regresión Lineal: Expectativa de vida vs Gasto per cápita en salud") +
+   geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="blue", size=0.5, se = FALSE) +
+   theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
>
> plot(grafico)
Warning message:
The following aesthetics were dropped during statistical transformation: label.
i This can happen when ggplot fails to infer the correct grouping structure in the data.
i Did you forget to specify a `group` aesthetic or to convert a numerical variable into a factor?
>
```

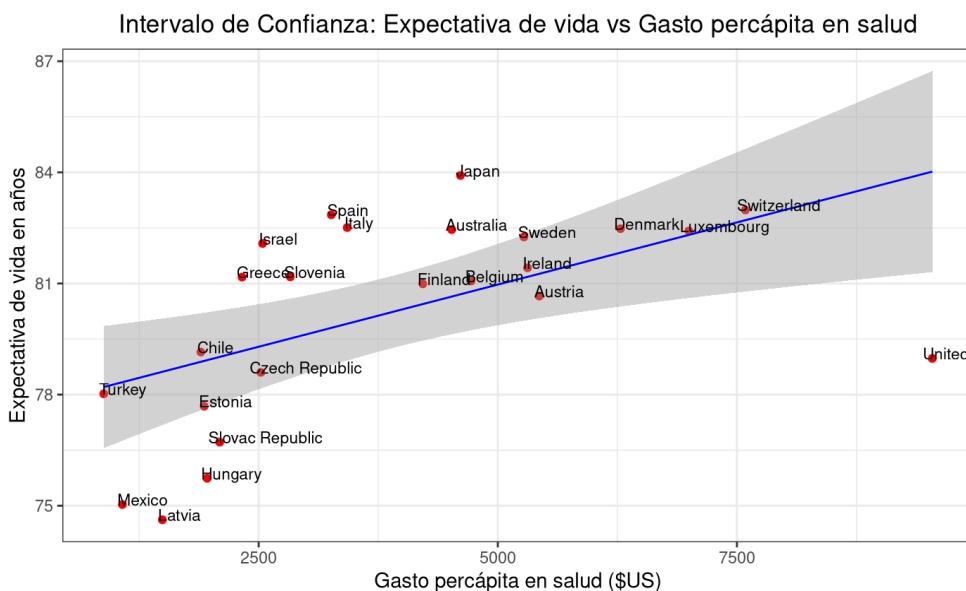
- Pregunta b) Con los datos obtenidos previamente, grafique el intervalo de confianza asociado a los datos y el modelo de regresión, analice los resultados a lo largo del ejercicio y explique si el modelo de regresión se ajusta a los datos. Según lo visto en clases, ¿Qué otro tipo de regresión lineal se adaptaría de mejor manera a los datos? ¿Por qué? (15 puntos)



Respuesta b.1: Al graficar el intervalo de confianza (Zona gris) podemos visualizar que algunos países no están bien representados, como son Japón, Australia, EE. UU., Lituania, entre otros. Pero el resto de países que están en la zona gris, como son Chile, Austria, Dinamarca, Suiza, entre otros, están bien representados con un 95% de confianza.

Por ende, y como se mencionó en la pregunta anterior a), el modelo no es bueno para el conjunto de datos evaluados.

Gráfico
Intervalo de
Confianza



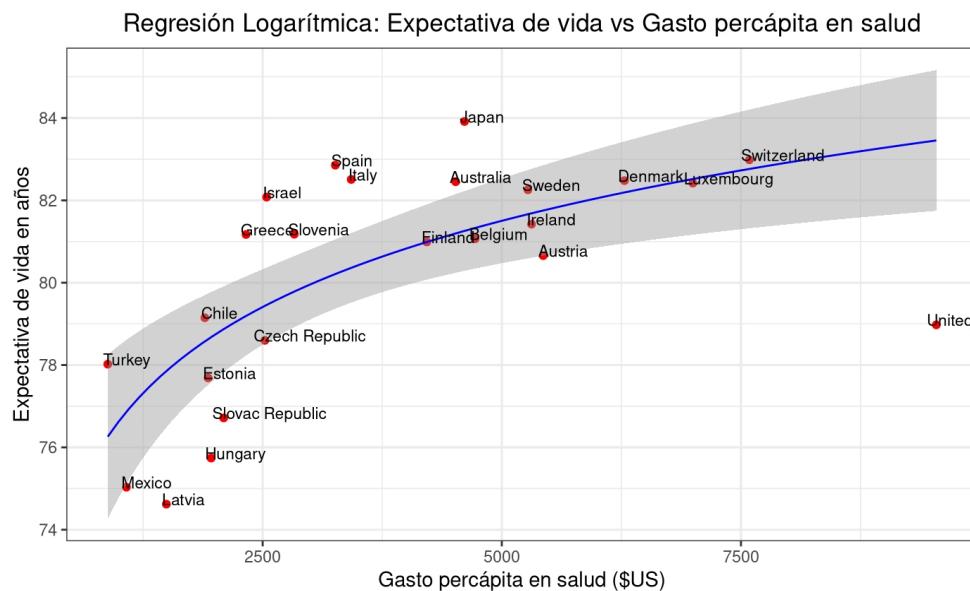
Respuesta b.2: Según lo visto en clases y basándose en la forma en que están distribuidos los datos, un modelo de regresión logística, podría ajustarse mejor a los datos.

Al aplicar este modelo de regresión logística se obtuvo un coeficiente de determinación (r^2) de 0.4822518, el cual al ser comparado con el r^2 de la regresión lineal: 0.3070257 se ve una mejora de un 57% con respecto al modelo de regresión lineal, no obstante, la condición es que sea mayor a 0.95 para ser considerado un buen modelo. Por ende, el modelo de regresión logística es una mejora con respecto al modelo de regresión lineal, pero no alcanza a cumplir con el criterio para ser considerado como modelo válido para estos datos.

Todo lo anterior se puede visualizar en el gráfico de los intervalos de confianza de la regresión logística, donde hay países que quedan más cerca de la zona gris, comparado con el modelo de regresión lineal, como es México, pero aún quedan países bastante alejados, como lo es EE. UU.

Cabe mencionar, que para poder obtener el modelo y gráfica, me base en los códigos de esta página ⇒ https://rpubs.com/Joaquin_AR/229736

Gráfica
Regresión
Logística
con
Intervalos
de
Confianza



Código R:

```

47
48   plot(grafico)
49
50  # =====
51  # b) Con los datos obtenidos previamente, grafique el intervalo de confianza asociado a los datos y el
52  # modelo de regresión, analice los resultados a lo largo del ejercicio y explique si el modelo de
53  # regresión se ajusta a los datos. Según lo visto en clases, ¿Qué otro tipo de regresión lineal se
54  # adaptaría de mejor manera a los datos? Por qué?. (15 puntos)
55  # =====
56
57  # Intervalo de confianza
58  confianza = confint(regresion, level=0.95)
59  print(confianza)
60
61  grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
62  geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
63  theme_bw() + xlab("Gasto percápita en salud ($US)") + ylab("Expectativa de vida en años") +
64  ggtitle("Intervalo de Confianza: Expectativa de vida vs Gasto percápita en salud") +
65  geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="blue", size=0.5) +
66  theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
67
68  plot(grafico)
69
70
71  # Ajustar el modelo de regresión logarítmica
72  r_log = lm(expectativa_anual ~ log(Gasto_salud_dolares), data = datos)
73
74  # Imprime el valor del r^2 del modelo
75  print(summary(r_log)$r.squared)
76
77  # Resumen del modelo
78  summary(r_log)
79
80  grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
81  geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
82  theme_bw() + xlab("Gasto percápita en salud ($US)") + ylab("Expectativa de vida en años") +
83  ggtitle("Regresión Logarítmica: Expectativa de vida vs Gasto percápita en salud") +
84  geom_smooth(method = "lm", formula = y ~ log(x), level=0.95, color="blue", size=0.5) +
85  theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
86
87  plot(grafico)
88

```



Consola R:

```
Console Terminal x Background jobs x
R 4.3.3 - ~/r
> =====
> # b) Con los datos obtenidos previamente, grafique el intervalo de confianza asociado a los datos y el
> # modelo de regresión, analice los resultados a lo largo del ejercicio y explique si el modelo de
> # regresión se ajusta a los datos. Según lo visto en clases, ¿Qué otro tipo de regresión lineal se
> # adaptaría de mejor manera a los datos? ¿Por qué? (15 puntos)
> =====
>
> # Intervalo de confianza
> confianza = confint(regresion, level=0.95)
> print(confianza)
      2.5 %    97.5 %
(Intercept) 7.563347e+01 79.592414568
Gasto_salud_dolares 2.254475e-04 0.001117566
>
> grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
+   geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
+   theme_bw() + xlab("Gasto per cápita en salud ($US)") + ylab("Expectativa de vida en años") +
+   ggtitle("Intervalo de Confianza: Expectativa de vida vs Gasto per cápita en salud") +
+   geom_smooth(method = "lm", formula = y ~ x, level=0.95, color="blue", size=0.5) +
+   theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
>
> plot(grafico)
Warning message:
The following aesthetics were dropped during statistical transformation: label.
i This can happen when ggplot fails to infer the correct grouping structure in the data.
i Did you forget to specify a 'group' aesthetic or to convert a numerical variable into a factor?
>
> # Ajustar el modelo de regresión logarítmica
> r_log = lm(expectativa_anual ~ log(Gasto_salud_dolares), data = datos)
>
> # Imprime el valor del r^2 del modelo
> print(summary(r_log)$r.squared)
[1] 0.4822518
>
> # Resumen del modelo
> summary(r_log)

Call:
lm(formula = expectativa_anual ~ log(Gasto_salud_dolares), data = datos)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.4788 -0.9790  0.1109  1.4856  2.6587 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 55.7772     5.4121 10.306 6.96e-10 ***
log(Gasto_salud_dolares) 3.0205     0.6672  4.527 0.000167 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.988 on 22 degrees of freedom
Multiple R-squared:  0.4823, Adjusted R-squared:  0.4587 
F-statistic: 20.49 on 1 and 22 DF, p-value: 0.0001666

>
> grafico=ggplot(datos,aes(Gasto_salud_dolares ,expectativa_anual,,label=Pais)) +
+   geom_point(aes(Gasto_salud_dolares,expectativa_anual),datos,color="red",alpha=1) +
+   theme_bw() + xlab("Gasto per cápita en salud ($US)") + ylab("Expectativa de vida en años") +
+   ggtitle("Regresión Logarítmica: Expectativa de vida vs Gasto per cápita en salud") +
+   geom_smooth(method = "lm", formula = y ~ log(x), level=0.95, color="blue", size=0.5) +
+   theme(plot.title = element_text(hjust = 0.5)) + geom_text(hjust=0.1, vjust=0.1, size=3)
>
> plot(grafico)
Warning message:
The following aesthetics were dropped during statistical transformation: label.
i This can happen when ggplot fails to infer the correct grouping structure in the data.
i Did you forget to specify a 'group' aesthetic or to convert a numerical variable into a factor?
>
```

ANEXO

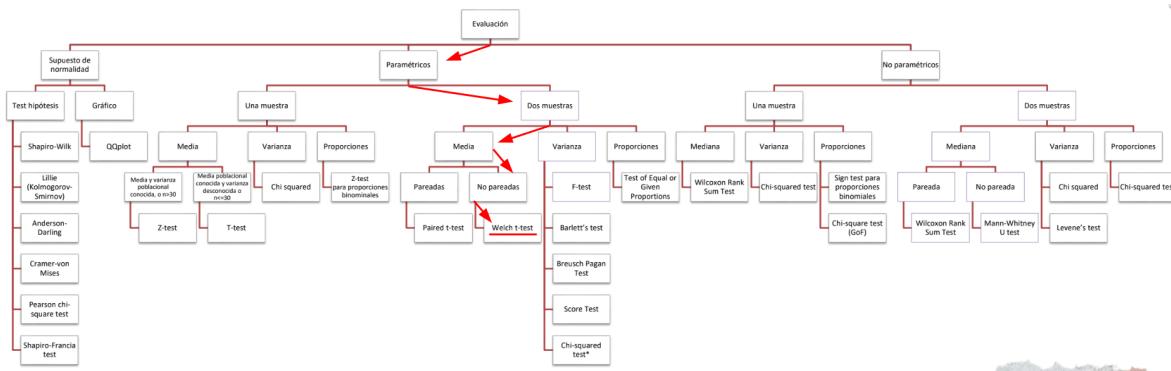


Figura 1. Ejemplo de la toma de decisión para elegir el test adecuado para responder a la Pregunta 1.