Diferencias en el tiempo de procesamiento de sistemas gestores de bases de datos

Julián Alejandro Úsuga Ortiz

Universidad Nacional de Colombia Sede Medellín

2022-11-24

Sección 1

Introducción

¿Por qué?

Existen diferentes tipos de sistemas de gestión de bases de datos, estos sistemas tienen la función de agregar, leer, actualizar y eliminar datos de una forma rápida y estable.

Sección 2

Experimento

Factor **gestor de base de datos**

Este es el factor más importante y en el cual se enfoca el diseño.

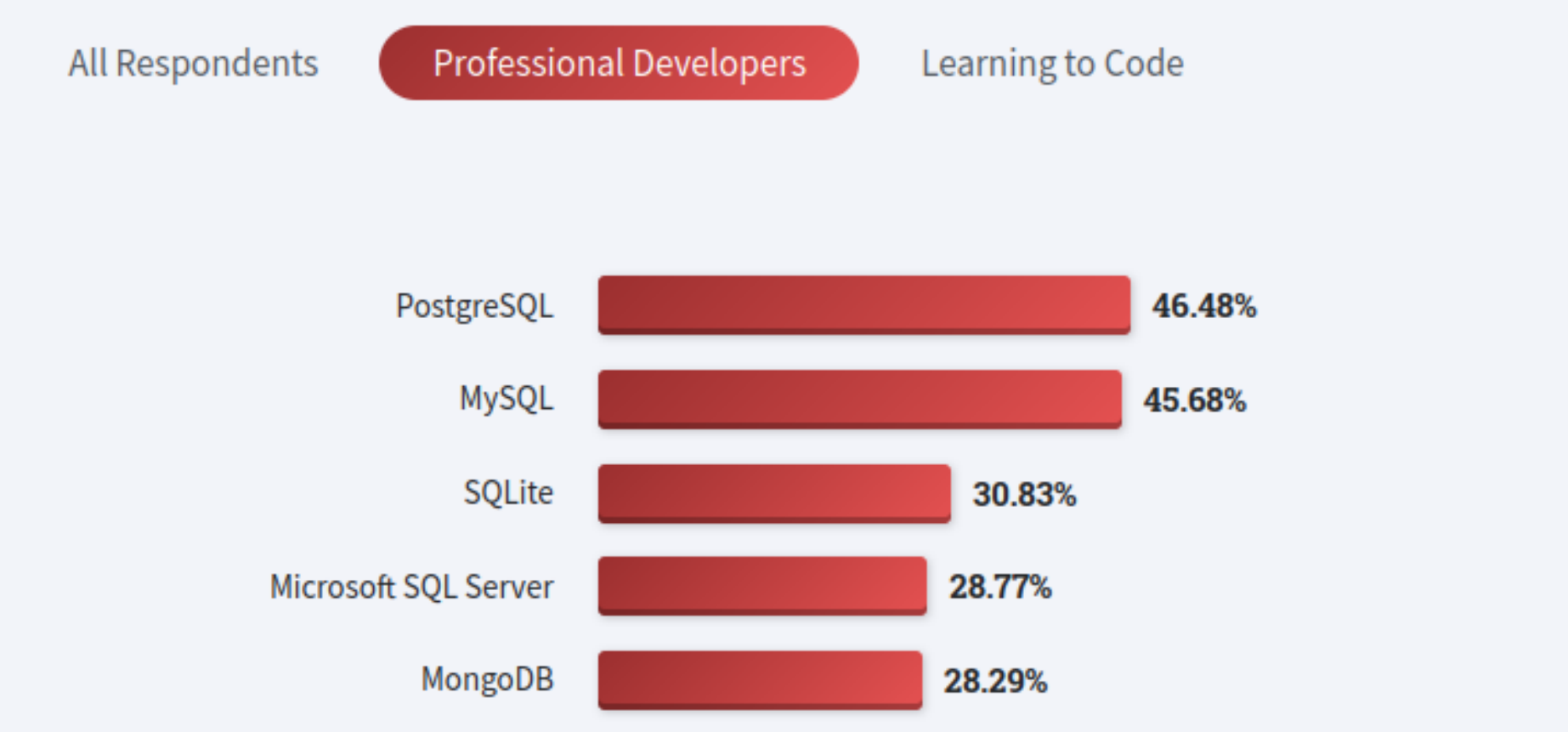


Figura 1: Resultados a la pregunta ¿En que entornos de bases de datos ha realizado un amplio trabajo de desarrollo en el último año?, StackOverflow Survey 2022

Factor **computador**

Figura 2: Los dos niveles del factor computador

Unidad experimental: Tabla de datos

Son 3 tablas que se obtuvieron de un proceso de aleatorización de una tabla más grande (1'378.033 filas). Cada tabla resultó teniendo 459.344 filas aprox.

Estos datos son de reseñas de libros y fueron descargados de Kaggle, Goodreads Books Reviews.

- Identificación del Usuario
- Identificación del Libro
- Identificación de la reseña
- **Reseña del Libro (Texto)**
- Fecha de la reseña
- Fecha de actualización de la reseña
- Fecha en la que se terminó de leerse el libro
- Fecha en la que empezó a leerse el libro
- Votos de la reseña
- Número de comentarios de la reseña

In [7]:

```
import random



nrows = len(df)

random.seed(2022)
index_list = random.sample(range(0, nrows), len(df))
index_list
```

In [9]:
```
df1 = df.iloc[index_list[0:459343]]
df2 = df.iloc[index_list[459344:918688]]
df3 = df.iloc[index_list[918689:1378033]]
```

In [10]:
```
# 0:459343 # 459343 filas
# 459344:918688 # 459344 filas
# 918689:1378033 # 459344 filas
```

In [11]:
```
df1.to_csv('df1.csv', index=False)
df2.to_csv('df2.csv', index=False)
df3.to_csv('df3.csv', index=False)
```

Bloque **consulta**

Los dos niveles del bloque son dos consultas diferentes.

```
experimentos=# SELECT COUNT(*) FROM goodreads2 WHERE review_text LIKE '%spoiler%';
 count
-------
 48443
(1 row)

Time: 869.771 ms
```

Figura 3: Consulta 1 hecha a la tabla 2 en el computador portatil con el gestor PostgreSQL

```
sqlite> SELECT COUNT(*)
FROM goodreads1 WHERE ((review_text LIKE '%entertaining%') OR (review_text LIKE '%motivating%'));
13301
Run Time: real 1.821 user 1.819504 sys 0.000000
```

Figura 4: Consulta 2 hecha a la tabla 1 en el computador portatil con el gestor SQLite

En la siguiente figura se intenta ilustrar el diseño experimental

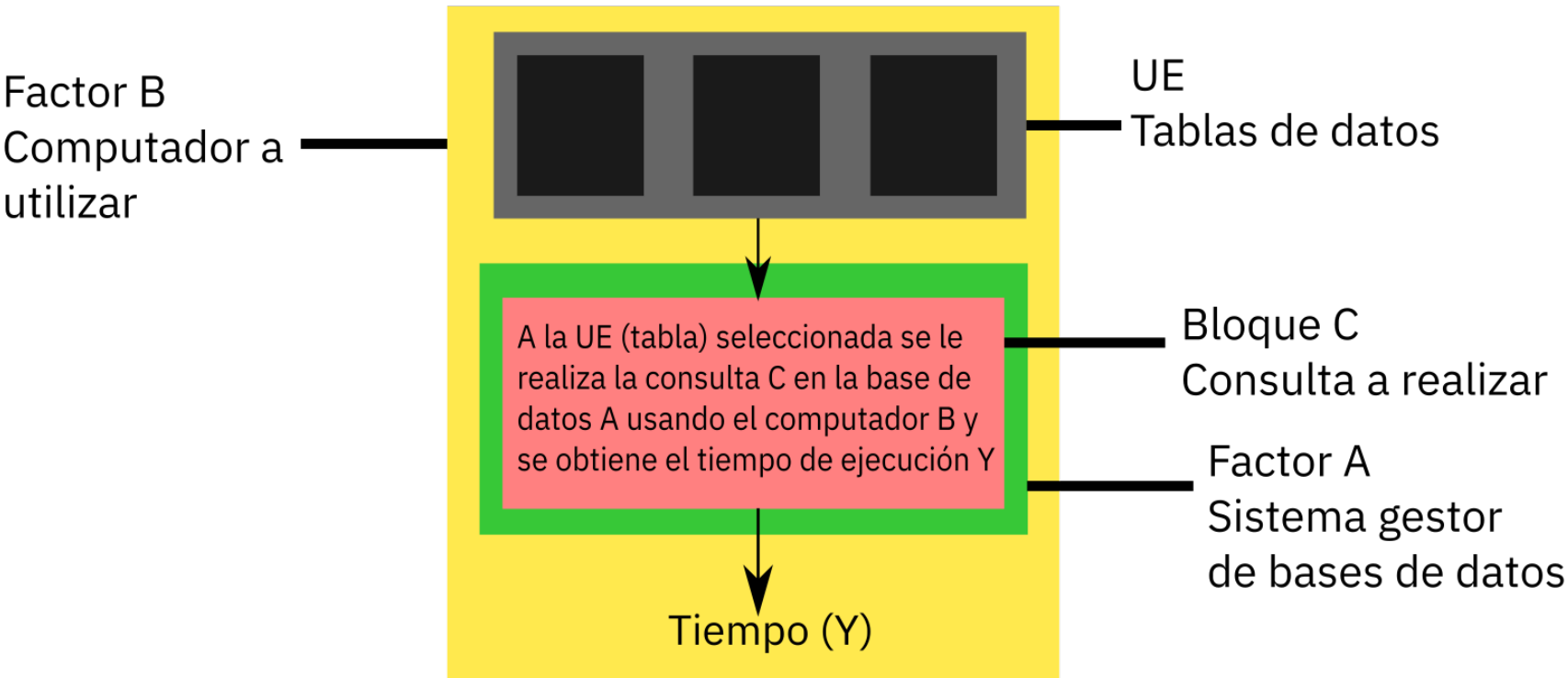


Figura 5: Diagrama del diseño

Sección 3

Como se garantizó aleatorización

Como se garantizó aleatorización

computador	gestor	consulta	tabla	orden
mesa	sqlite	1	2	1
portatil	mysql	2	2	2
portatil	sqlite	1	3	3
portatil	posgresql	1	1	4
portatil	posgresql	1	2	5

	computador	gestor	consulta	tabla	orden
32	portatil	mysql	1	3	32
33	mesa	mysql	2	3	33
34	portatil	posgresql	1	3	34
35	mesa	mysql	2	1	35
36	mesa	mysql	1	1	36

Sección 4

Modelo

Modelo

$$Y_{ijkl} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \alpha_k + \varepsilon_{ijkl}$$

Con $\varepsilon_{ijkl} \sim NID(0, \sigma^2)$. Y con las restricciones $\sum_{i=1}^{3} \tau_i = 0$, $\sum_{j=1}^{2} \beta_j = 0$, $\sum_{k=1}^{2} \alpha_k = 0$.

Donde

- La letra **i** representa cada gestor de bases de datos, en este experimento son 3 niveles (PostgreSQL, MySQL y SQLite).
- El factor A es el gestor de sistemas de base de datos usado, su efecto va a ser representado por la letra griega τ.
- La letra **j** representa cada computador usado en la realización del experimento, se usaron dos computadores con capacidades muy diferentes, un computador portátil y un computador de mesa, por lo que se consideraran estos dos niveles.

- El factor B es el computador a usar y el efecto de este será representado por la letra griega β.
- La letra **k** representa la consulta usada para obtener un resultado.
- El bloque C representa cada una de las consultas a realizar, en total se realizarán 2 consultas diferentes. De antemano se sabe que estas toman tiempos diferentes en procesar. El efecto de este bloque será representado por la letra griega α.
- El error aleatorio representado por la letra ε.
- La letra **l** representa la tabla (UE) a muestrear, decidí muestrear todas las UE para cada pareja de computador-gestor-consulta.
- Y_{ijkl} representa el tiempo en segundos que le toma al sistema gestor de bases de datos i procesar el l-esimo grupo de filas de datos (UE) usando la consulta k en el computador j.

Sección 5

Datos obtenidos

Datos obtenidos

computador	gestor	consulta	tabla	orden	tiempo
mesa	sqlite	1	2	1	0.532
portatil	mysql	2	2	2	6.470
portatil	sqlite	1	3	3	0.966
portatil	posgresql	1	1	4	0.859
portatil	posgresql	1	2	5	0.869
mesa	sqlite	1	1	6	0.547
mesa	mysql	1	2	7	3.410
mesa	sqlite	2	3	8	1.125
portatil	posgresql	2	1	9	1.655
mesa	posgresql	2	2	10	1.439

Sección 6

Resultados del análisis

Resultados del análisis

Tabla 4: Analisis de Varianza del modelo planteado

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
computador	1	0.27	0.27	0.44	0.51
gestor	2	142	71	114	1.8e-14
consulta	1	25	25	41	5.8e-07
computador:gestor	2	0.78	0.39	0.63	0.54
Residuals	29	18	0.62	NA	NA

Como se puede observar hay varios factores no significativos. Procedí a eliminar el efecto del factor de interacción entre los factores gestor de base de datos y computador, ya que este es el menos significativo. Que esta interacción no sea significativa se interpreta como que **los gestores de base de datos no tienden a comportarse mejor o peor dependiendo de que computador se usa**.

Nuevo modelo planteado

Sin el factor de interacción, el nuevo modelo planteado es el siguiente:

$$Y_{ijkl} = \mu + \tau_i + \beta_j + \alpha_k + \varepsilon_{ijkl}$$

$$\text{Con } \varepsilon_{ijkl} \sim N(0, \sigma^2).$$

Tabla 5: Analisis de Varianza sin interacción

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gestor	2	142	71	117	3.7e-15
computador	1	0.27	0.27	0.45	0.51
consulta	1	25	25	42	3.5e-07
Residuals	31	19	0.61	NA	NA

Tabla 5: Analisis de Varianza sin interacción

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gestor	2	142	71	117	3.7e-15
computador	1	0.27	0.27	0.45	0.51
consulta	1	25	25	42	3.5e-07
Residuals	31	19	0.61	NA	NA

Del análisis de varianza anterior se observa que el factor computador no es significativo, este resultado no se esperaba, ya que como se había dicho antes ambos computadores son de años distintos de fabricación y con procesadores muy diferentes, pero con este resultado se eliminará este factor.

Finalmente el modelo planteado será el siguiente

$$Y_{ijl} = \mu + \tau_i + \beta_j + \varepsilon_{ijl}$$

$$\text{Con } \varepsilon_{ijl} \sim N(0, \sigma^2).$$

Y los datos utilizados

gestor	consulta	orden	tiempo
mysql	1	7	3.410
mysql	1	14	3.690
mysql	1	18	3.690
mysql	1	26	3.450
mysql	1	32	3.690
mysql	1	36	3.440
mysql	2	2	6.470
mysql	2	20	7.900
mysql	2	22	7.380
mysql	2	28	6.760
mysql	2	33	7.140
mysql	2	35	7.300
posgresql	1	4	0.859
posgresql	1	5	0.869
posgresql	1	11	0.812

Tabla 7: Analisis de Varianza sin factor Computador

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gestor	2	142	71	119	1.5e-15
consulta	1	25	25	42	2.6e-07
Residuals	32	19	0.6	NA	NA

Tabla 7: Analisis de Varianza sin factor Computador

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gestor	2	142	71	119	1.5e-15
consulta	1	25	25	42	2.6e-07
Residuals	32	19	0.6	NA	NA

Modelo obtenido donde el factor y bloque son significativos.

Sección 7

Verificación de los supuestos

Normalidad

Tabla 8: Shapiro-Wilk normality test: `residuals(analisis)`

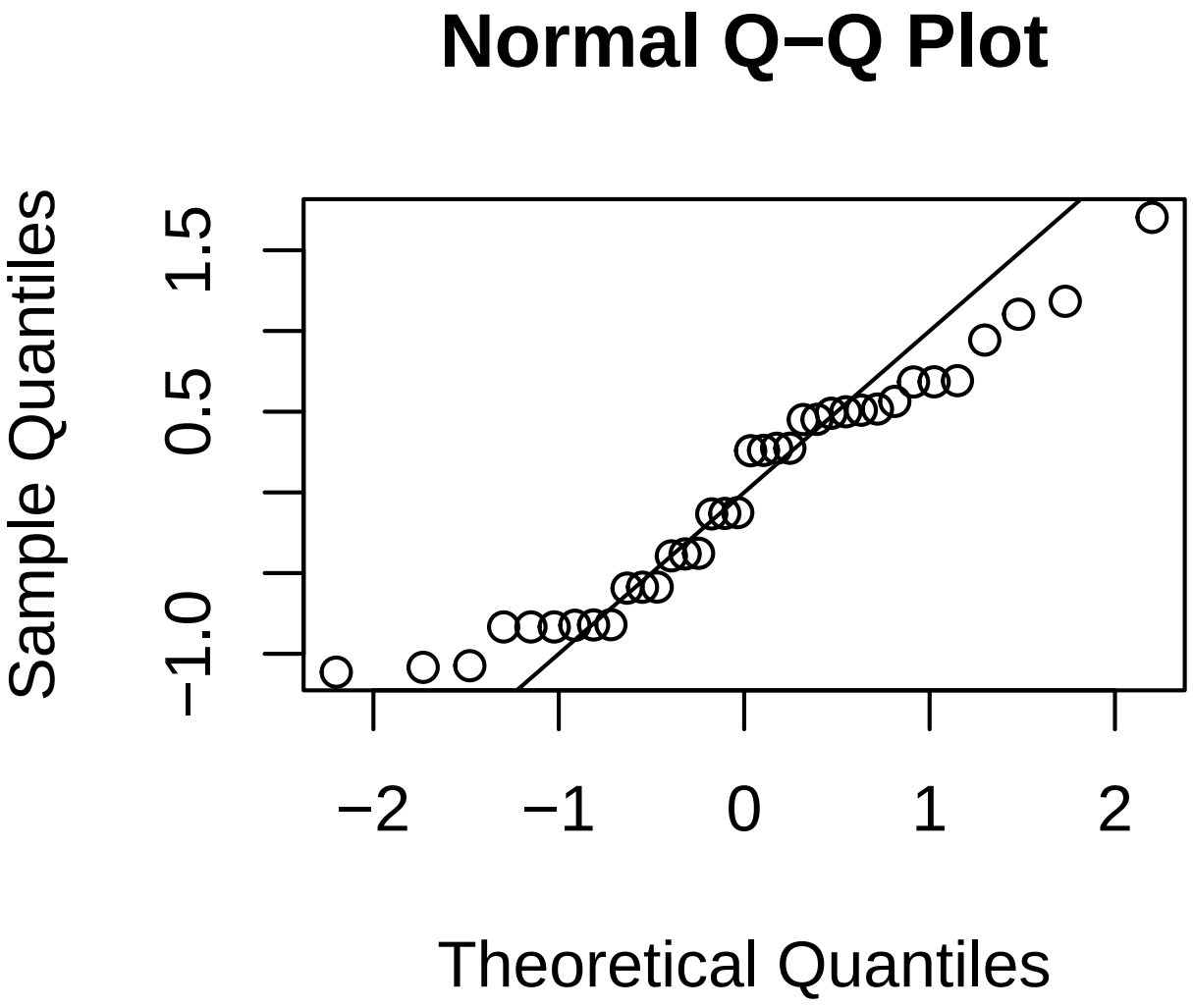
Test statistic	P value
0.9463	0.08007

Normalidad

Tabla 8: Shapiro-Wilk normality test: `residuals(analisis)`

Test statistic	P value
0.9463	0.08007

Con el test de Shapiro-Wilk se falla al rechazar la hipótesis nula (H_0: Los errores provienen de una distribución normal) y concluimos que hay suficiente evidencia para afirmar que los errores provienen de una distribución normal con una significancia del 5%.


Normal Q−Q Plot
Sample Quantiles
−1.0
0.5
1.5
−2
−1
0
1
2
Theoretical Quantiles

Independencia de los errores

Dado que contamos con el orden en que se tomaron los datos, podemos realizar el test de Durbin-Watson, en el cual la hipótesis nula afirma que los errores no tienen autocorrelación ($\rho = 0$) o son ruido blanco.

Se obtiene un valor-p mayor al nivel de significancia $\alpha = 0.05$, por lo que se falla al rechazar la hipótesis nula y se afirma que los errores tienen correlación cero con una significancia del 5%, esto es, que en el tiempo exactamente anterior las tomas no están relacionadas unas con otras.

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1392044      1.656495   0.286
##  Alternative hypothesis: rho != 0
```

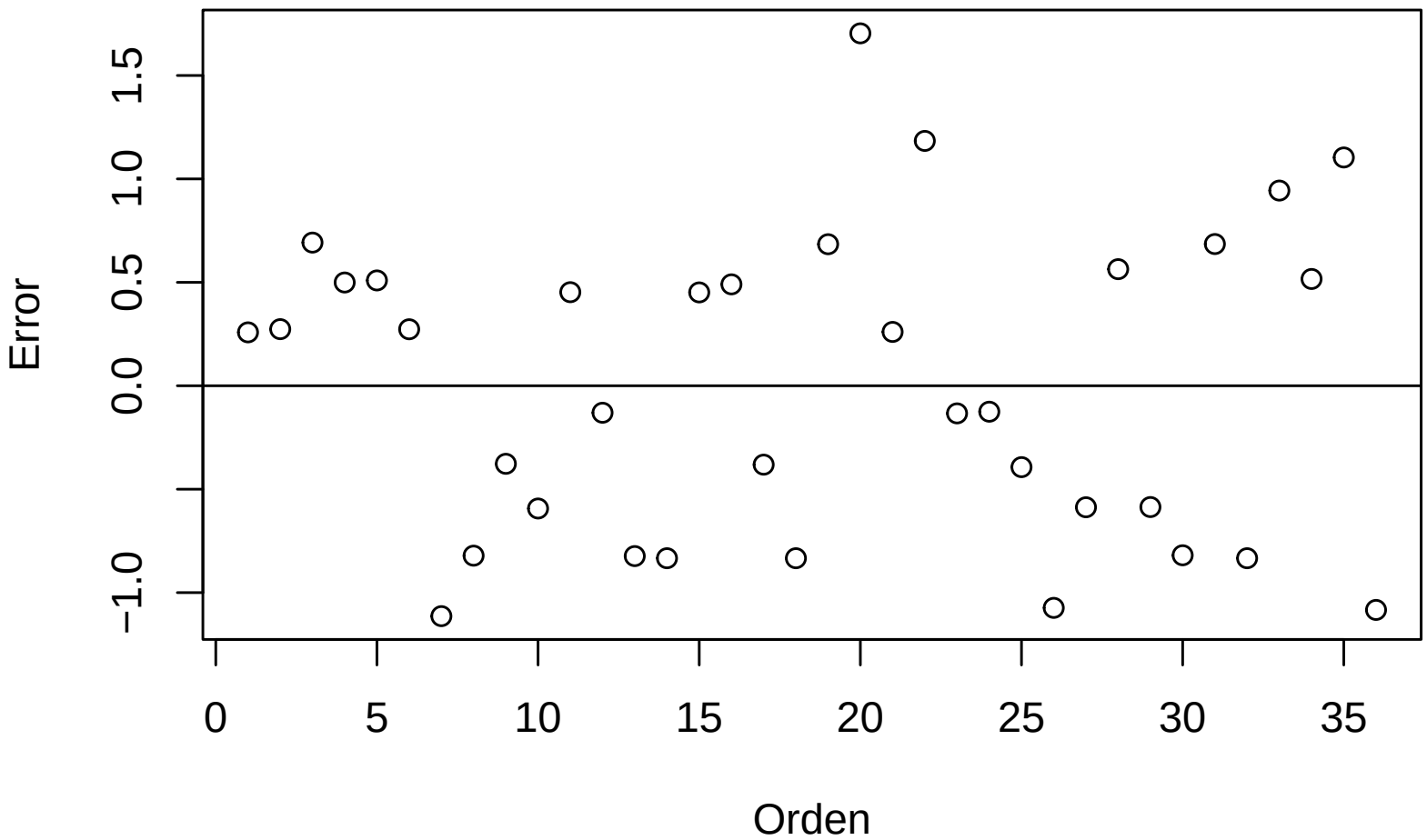


Figura 6: Orden de toma de datos vs errores

Homogeneidad de varianzas

Se realiza el test de Bartlett para probar homogeneidad de varianzas (La hipótesis nula es que las varianzas en cada grupo son las mismas).

Tabla 9: Bartlett test of homogeneity of variances: `analisis$residuals` and `datos$consulta`

Test statistic	df	P value
0.1257	1	0.7229

Tabla 10: Bartlett test of homogeneity of variances: `analisis$residuals` and `datos$gestor`

Test statistic	df	P value
6.87	2	0.03222 *

Tabla 9: Bartlett test of homogeneity of variances: `analisis$residuals` and `datos$consulta`

Test statistic	df	P value
0.1257	1	0.7229

Tabla 10: Bartlett test of homogeneity of variances: `analisis$residuals` and `datos$gestor`

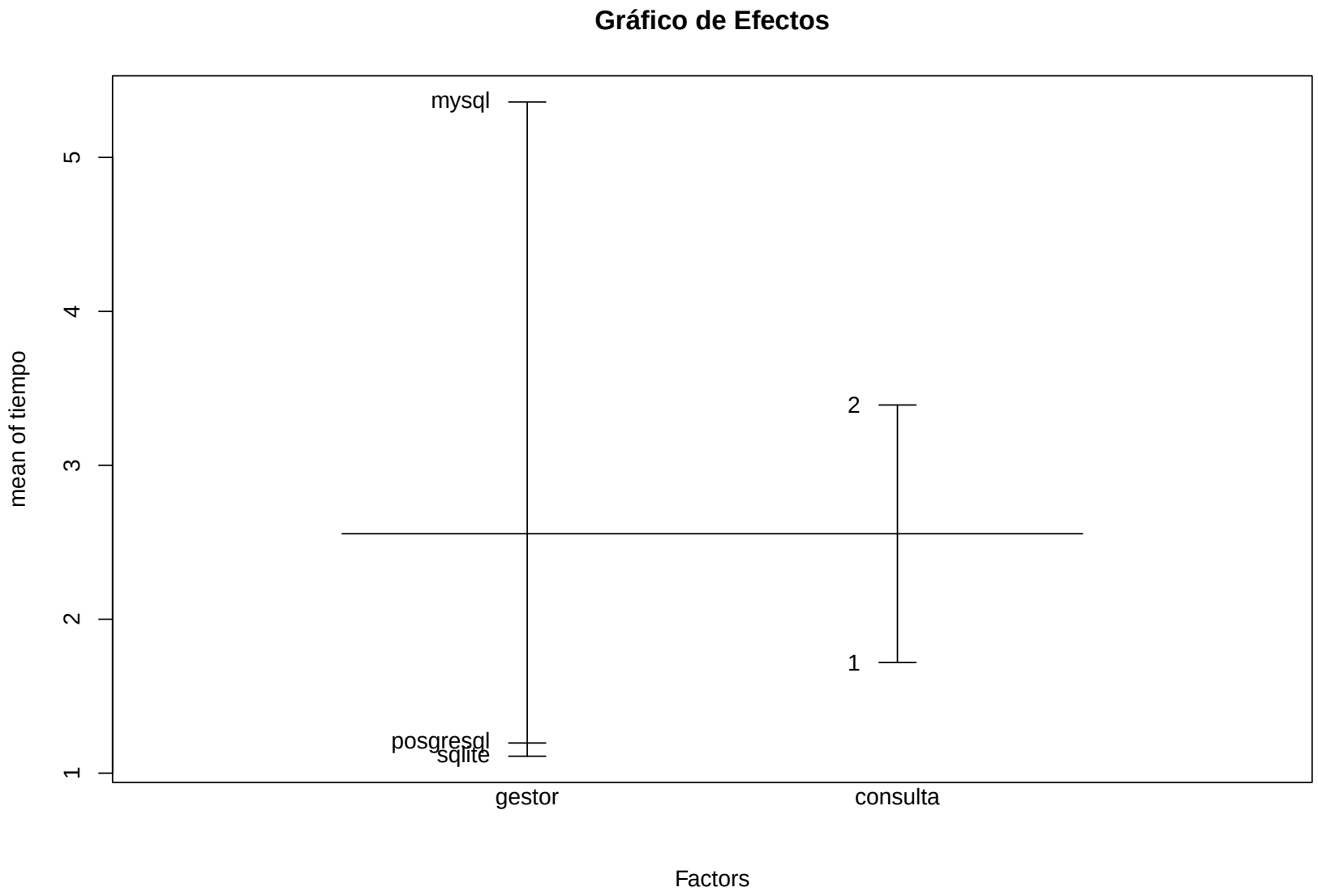
Test statistic	df	P value
6.87	2	0.03222 *

Se observa que entre los dos niveles del factor consulta existe homogeneidad de las varianzas (con una significancia del 5%) mientras que entre los tres gestores de bases de datos no existe homogeneidad de varianzas (aunque el valor del estadístico está cercano a la región de rechazo).

Sección 8

Gráficos

Gráfico de efectos



Comparaciones de medias

- **means**:

Tabla 11: Table continues below

	tiempo	std	r	LCL	UCL	Min	Max	Q25
mysql	5.36	1.911	12	4.906	5.814	3.41	7.9	3.63
posgresql	1.196	0.3737	12	0.7421	1.65	0.811	1.655	0.8568
sqlite	1.11	0.4813	12	0.6562	1.564	0.532	1.821	0.8552

	Q50	Q75
mysql	5.08	7.18
posgresql	1.157	1.494
sqlite	1.044	1.298

- **statistics**:

	MSerror	Df	Mean	CV	t.value	LSD
	0.5958	32	2.555	30.21	2.037	0.6419

- **groups**:

	tiempo	groups
mysql	5.36	a
posgresql	1.196	b
sqlite	1.11	b

La media del tratamiento i es significativamente diferente a la media del tratamiento j si

$$|\bar{y}_i - \bar{y}_j| \geq t_{\alpha/2, dfError} \sqrt{MSE \frac{2}{n}}$$

```
LSD <- qt(0.05 / 2, 32, lower.tail = FALSE) *
  sqrt(0.5957896 * (2 / 12))
cat(abs(5.36 - 1.196) >= LSD) # mysql y postgresql
```

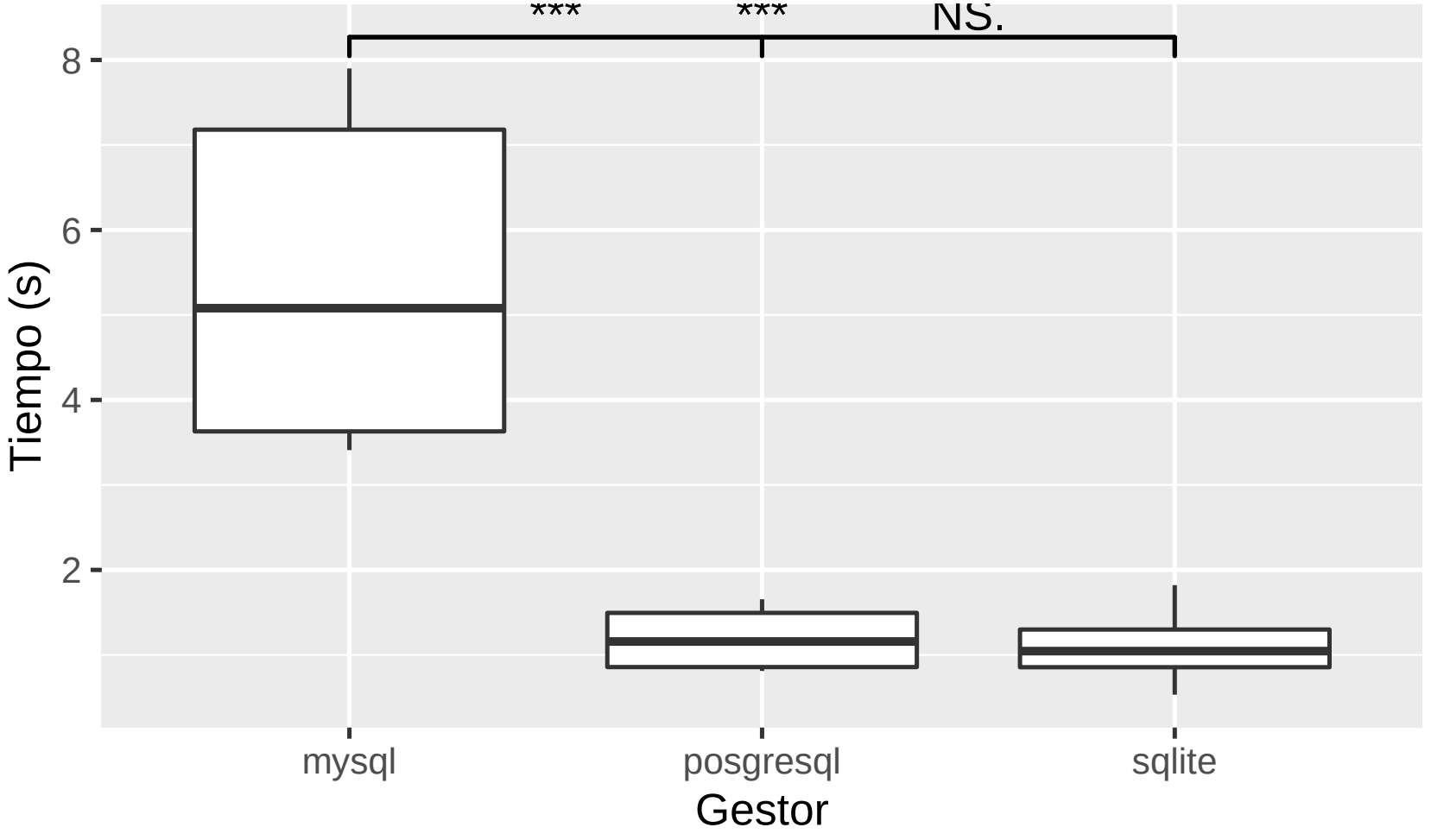
```
## TRUE
```

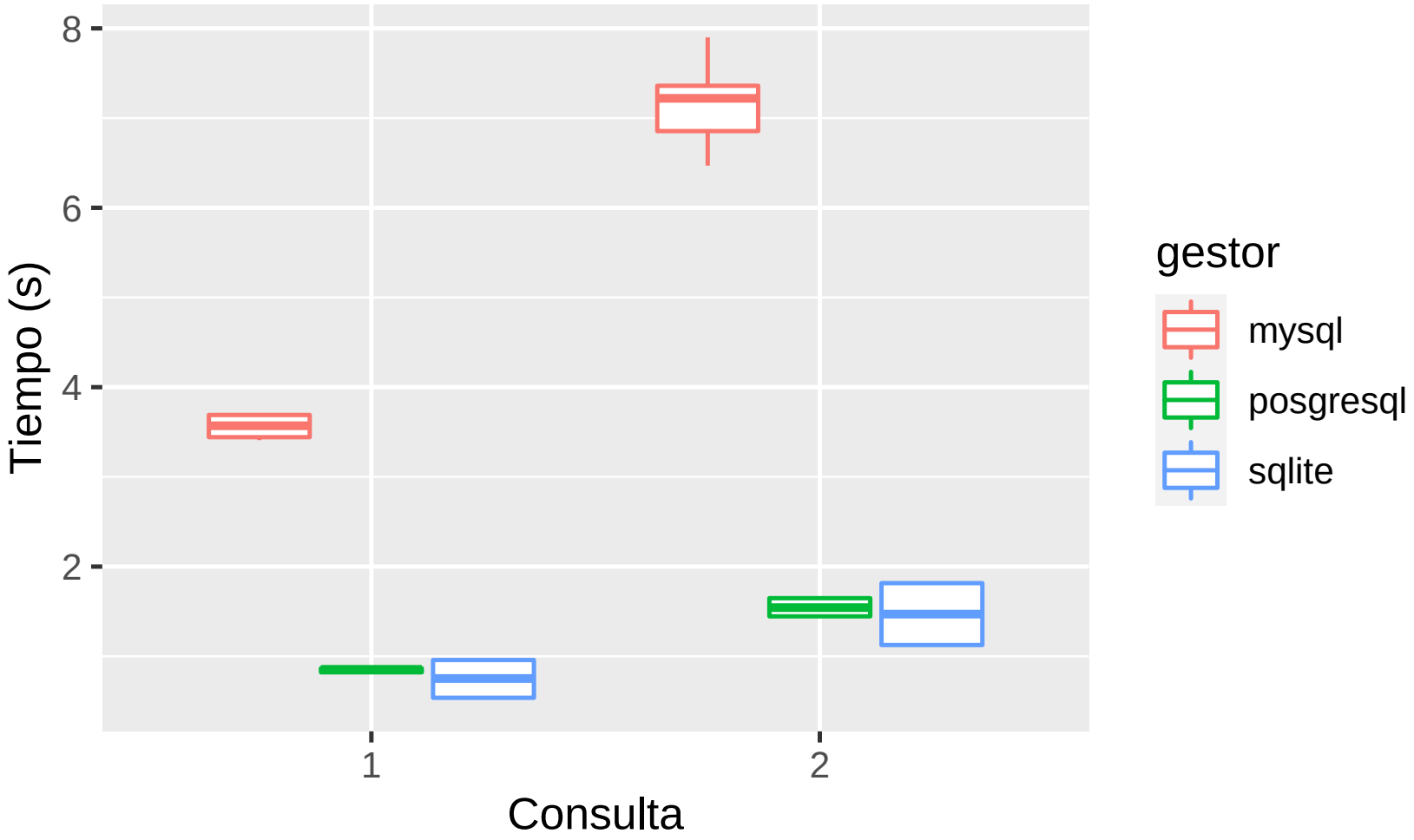
```
cat(abs(5.36 - 1.11) >= LSD) # mysql y sqlite
```

```
## TRUE
```

```
cat(abs(1.11 - 1.196) >= LSD) # sqlite y postgresql
```

```
## FALSE
```


***
***
NS.
8
6
4
2
Tiempo (s)
mysql
posgresql
sqlite
Gestor


8
6
4
2
Tiempo (s)
1
2
Consulta
gestor
mysql
posgresql
sqlite

Sección 9

Conclusiones

Conclusiones

- En este diseño se pudo observar que efectivamente entre gestores de bases de datos hay diferencias.

Conclusiones

- En este diseño se pudo observar que efectivamente entre gestores de bases de datos hay diferencias.
- El Bloque **Consulta** efectivamente era significativo y realizar una consulta buscando una palabra toma menos tiempo que una consulta buscando dos palabras.

Conclusiones

- En este diseño se pudo observar que efectivamente entre gestores de bases de datos hay diferencias.
- El Bloque **Consulta** efectivamente era significativo y realizar una consulta buscando una palabra toma menos tiempo que una consulta buscando dos palabras.
- La interacción entre computador y gestor no es significativa.

Conclusiones

- En este diseño se pudo observar que efectivamente entre gestores de bases de datos hay diferencias.
- El Bloque **Consulta** efectivamente era significativo y realizar una consulta buscando una palabra toma menos tiempo que una consulta buscando dos palabras.
- La interacción entre computador y gestor no es significativa.
- Los gestores **SQLite** y **PostgreSQL** tienen unas velocidades mayores a **MySQL**, aunque es de notar que muchas veces es más importante el caso de uso.

Conclusiones

- En este diseño se pudo observar que efectivamente entre gestores de bases de datos hay diferencias.
- El Bloque **Consulta** efectivamente era significativo y realizar una consulta buscando una palabra toma menos tiempo que una consulta buscando dos palabras.
- La interacción entre computador y gestor no es significativa.
- Los gestores **SQLite** y **PostgreSQL** tienen unas velocidades mayores a **MySQL**, aunque es de notar que muchas veces es más importante el caso de uso.
- El gestor de base de datos **PostgreSQL** ofrece muchas herramientas interesantes que los demás no, por ejemplo **Mínimos cuadrados ordinarios**, **R^2** y la **Correlación** entre dos columnas de una tabla.

Sección 10

Muchas gracias!