Tarea 2: Regresión Lineal Dólar

David Caleb Chaparro Orozco
Universidad de Envigado

Facultad de Ingeniería
Programa Ingeniería de Sistemas
Medellín, Marzo de 2023

Codigo GitHub: (Orozco, 2023)

Observaciones:

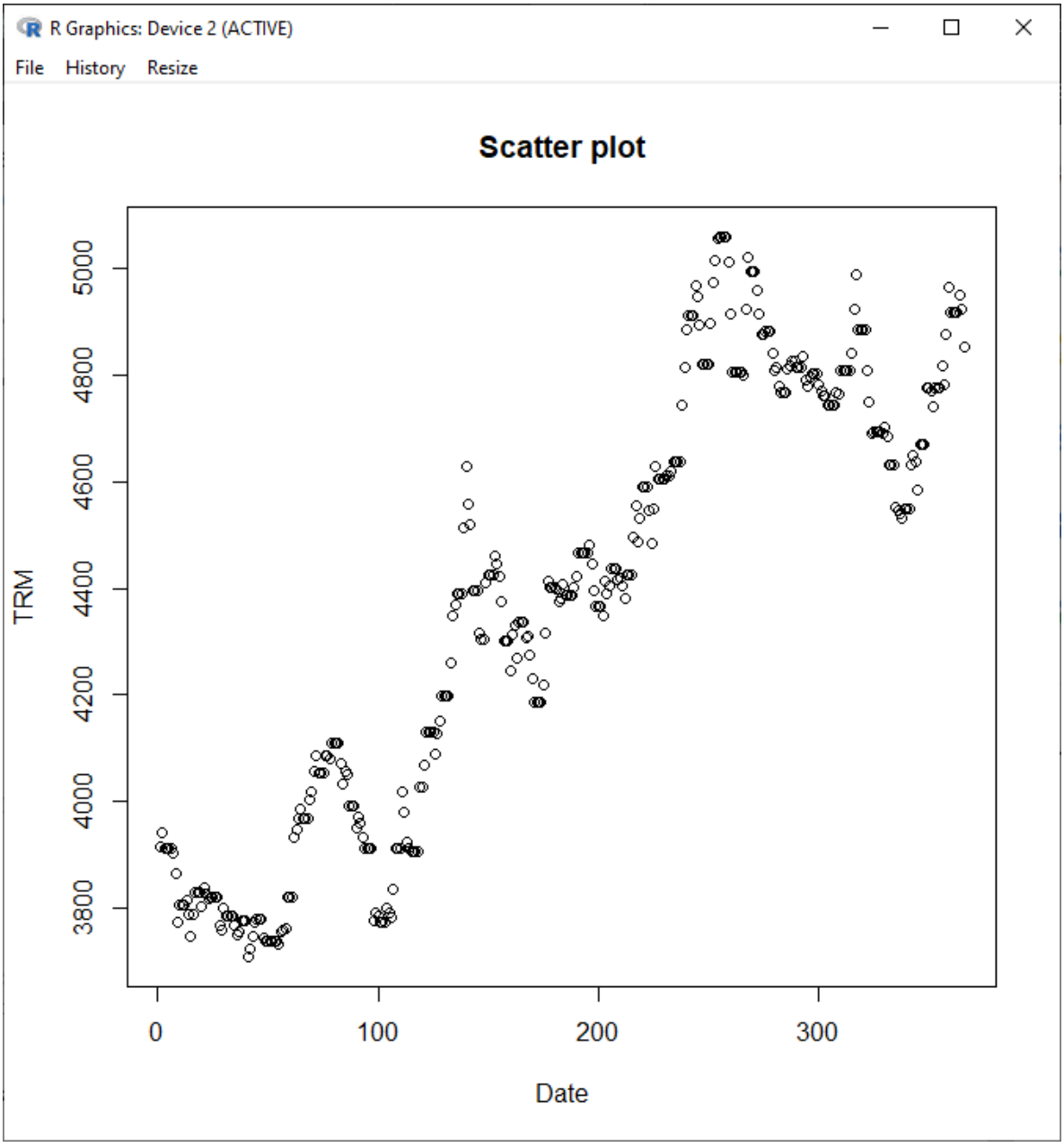
1. Para poder visualizar el código se puede realizar a través de Teams en el Equipo “Analisis de Algoritmos” - “Archivos: Entrega seguimiento 2-13 de marzo de 2023” - “Carpeta: DavidCalebChaparroOrozco” o por el contrario a través de la referencia al final del documento que te arrojará al GitHub con sus respectivos comentarios.

2. Este documento está realizado de acuerdo a los temas vistos en Análisis de Algoritmos con el docente: Diego Fernando Rangel Arciniegas - Docente IUE

Codigo GitHub: (Orozco, 2023)

Ejercicio:

1. **Grafique un diagrama de dispersión de los datos. Interpretar.**

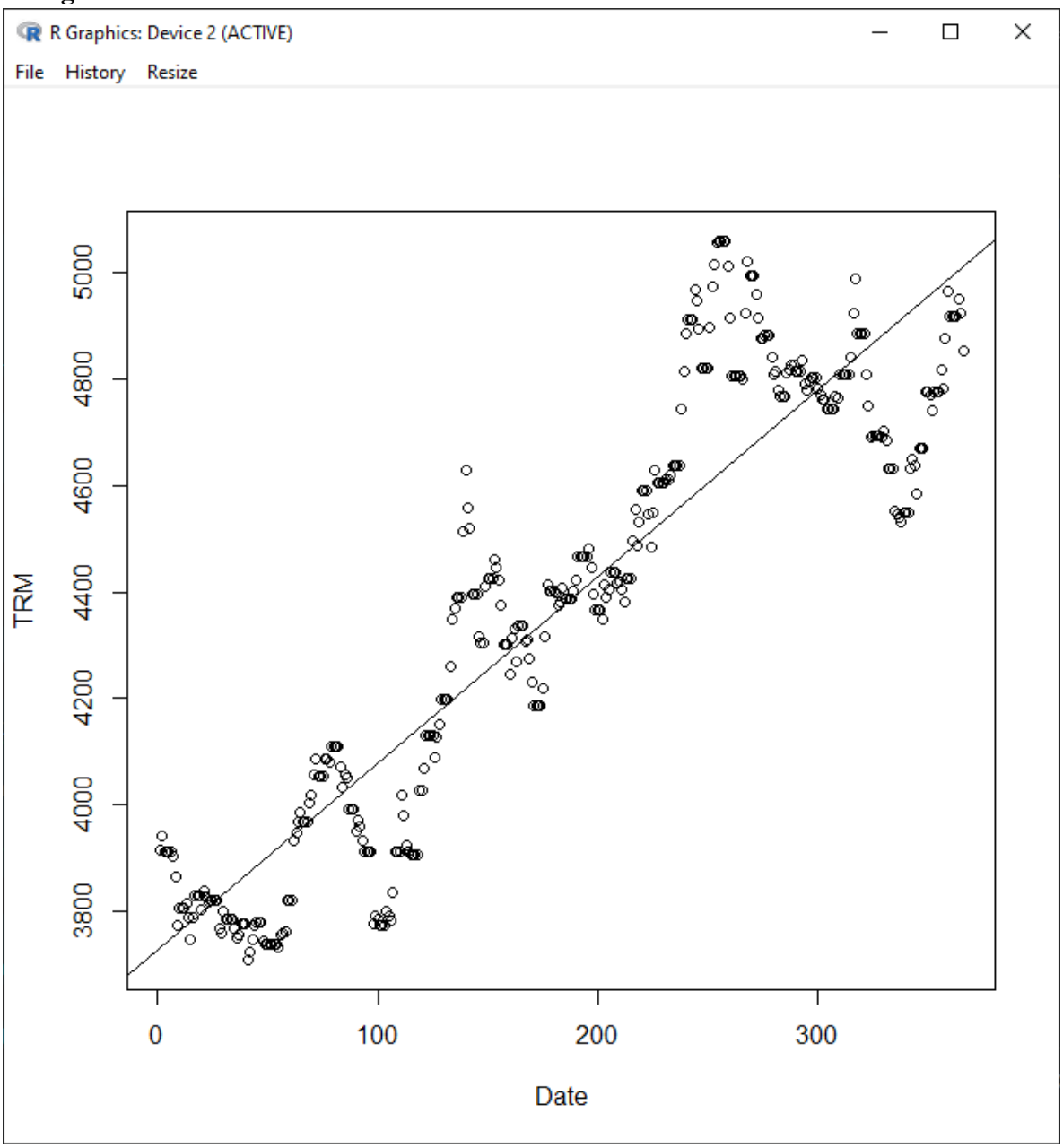


El diagrama de dispersión muestra la relación entre las variables "Date" y "TRM". Podemos observar que a medida que transcurren los días, el precio del dólar tiende a variar, lo que indica que se trata de un activo muy volátil. Además, los puntos en el gráfico están relativamente cercanos a una línea recta, lo que sugiere una relación lineal fuerte entre las dos variables. Sin embargo, es importante tener en cuenta que la presencia de una relación lineal no implica necesariamente causalidad, por lo que no podemos inferir que los cambios en la fecha sean la causa de las variaciones en

Codigo GitHub: (Orozco, 2023)

el precio del dólar.

2. **Calcule la ecuación de la recta de regresión, interprete la pendiente de la recta de regresión.**



La regresión lineal busca encontrar un modelo matemático que relacione dos variables, en este caso la fecha 'data_x' y el TRM 'data_y'. La ecuación de la recta de regresión muestra cómo 'data_y' se relaciona con 'data_x', y su fórmula es y = mx + b, donde 'm' es la pendiente y 'b' es la intersección en y. La pendiente indica cuánto cambia 'data_y' por cada unidad de cambio en 'data_x', mientras que la intersección en y indica el valor de 'data_y' cuando 'data_x' es igual a cero.

Codigo GitHub: (Orozco, 2023)

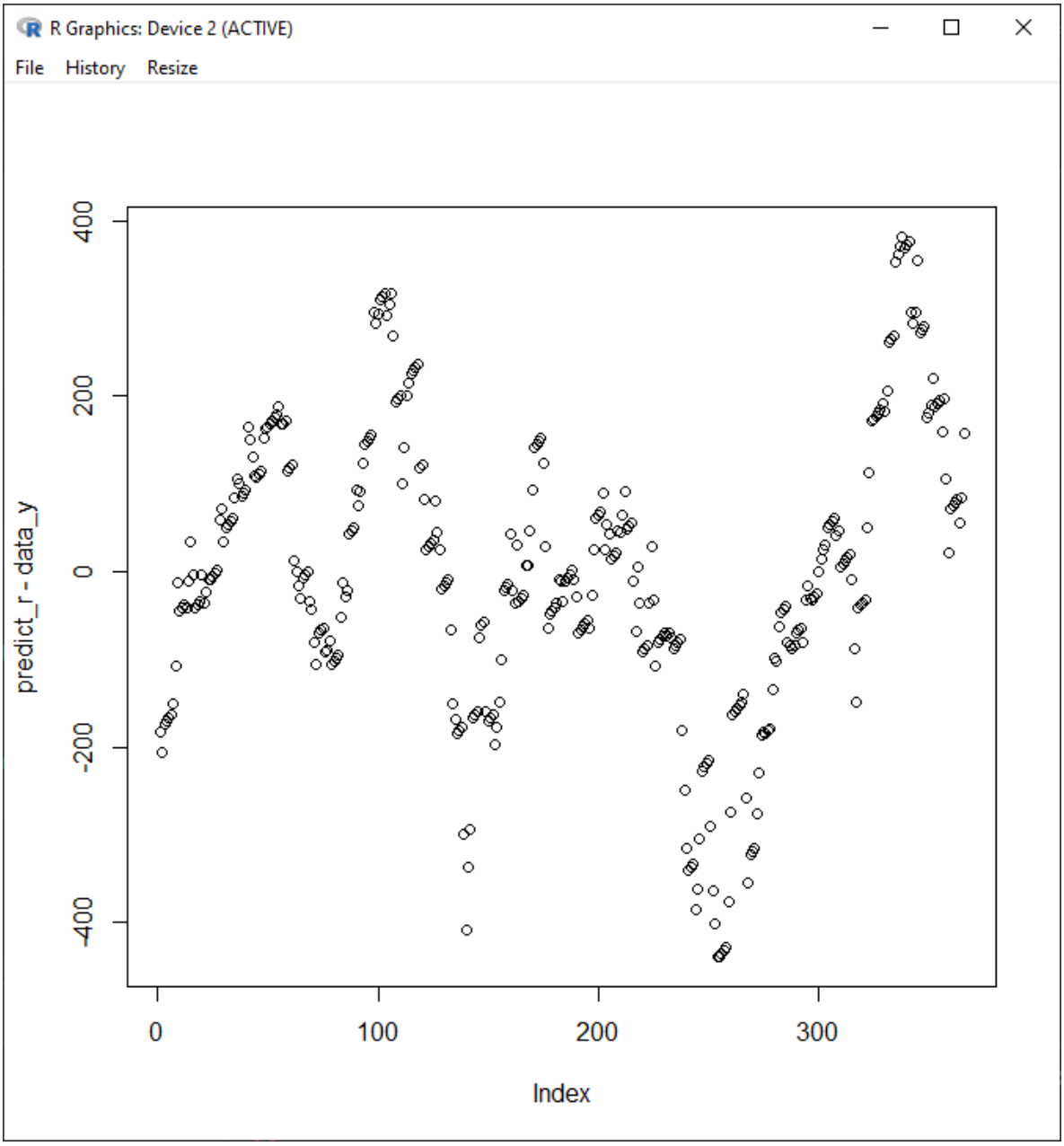
En este caso, la pendiente es positiva, lo que significa que hay una relación directa entre 'data_x' y 'data_y': a medida que 'data_x' aumenta, también aumenta 'data_y'. Además, la pendiente nos permite cuantificar esa relación, ya que indica cuánto aumenta o disminuye 'data_y' por cada unidad de cambio en 'data_x'. Por ejemplo, si la pendiente fuera 2, entonces por cada unidad de cambio en 'data_x', 'data_y' aumentaría en 2 unidades.

Es importante destacar que el objetivo de la regresión lineal es minimizar la diferencia entre los valores reales de 'data_y' y sus estimaciones respectivas, es decir, el error. El error se mide como la diferencia entre el valor real de 'data_y' y su valor estimado en cada punto. Por lo tanto, se espera que el error sea lo más pequeño posible para que la recta de regresión sea un buen modelo para los datos.

En resumen, la ecuación de la recta de regresión muestra cómo 'data_y' se relaciona con 'data_x', y la pendiente de la recta indica la dirección y magnitud de esa relación. La minimización del error es esencial para que la recta de regresión sea un buen modelo para los datos, ya que indica qué tan bien los valores estimados de 'data_y' se ajustan a los valores reales.

Codigo GitHub: (Orozco, 2023)

3. **Estime el precio del dólar utilizando la recta de regresión (para cada valor de x) y halle los residuales.**



Podemos utilizar la ecuación de la recta para predecir el valor de la TRM en una fecha específica en el futuro, dado que la relación entre la fecha y la TRM sigue siendo lineal y constante. Los residuos representan la diferencia entre el valor real y el valor estimado por la recta de regresión, por lo que podemos utilizarlos para evaluar la precisión de nuestra predicción. Si los residuos son pequeños, esto indica

Codigo GitHub: (Orozco, 2023)

que nuestra predicción es precisa y que la recta de regresión es un buen modelo para predecir los valores futuros de la TRM. Por otro lado, si los residuos son grandes, esto indica que nuestra predicción puede no ser precisa y que puede ser necesario buscar un modelo de regresión más adecuado.

Codigo GitHub: (Orozco, 2023)

4. **Calcule s^2 y compárelo con el valor hallado en la tabla ANOVA.**

```
> varianza <- sum(s_varianza)/364
> anova(regression)
Analysis of Variance Table

Response: data_y
           Df   Sum Sq  Mean Sq F value    Pr(>F)
data_x      1 50383409 50383409  1864.1 < 2.2e-16 ***
Residuals 364  9838270    27028
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para calcular la varianza del modelo, podemos usar la siguiente fórmula vista en clase:

S^2 (La Varianza del Modelo)

$$S^2 = \frac{SCE}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2}$$

Primero necesitamos calcular SCE, que se puede obtener de la tabla ANOVA. En este caso, SCE = 9838270. Luego, podemos calcular s^2 como:

$$S^2 = SCE/(n-p-1) = 9838270/(366-1-1) = 27025.75$$

Comparando este valor con el valor de la tabla ANOVA, vemos que son muy similares. El valor en la tabla ANOVA es 27028, y el valor calculado es 27025.75. La pequeña diferencia se debe a redondeos en el cálculo. En general, se espera que los valores calculados y los valores de la tabla ANOVA sean similares, y en este caso, esto se cumple.

Codigo GitHub: (Orozco, 2023)

5. **Realice la tabla ANOVA y analice la significancia del modelo.**

```
> anova(regression)
Analysis of Variance Table

Response: data_y
           Df   Sum Sq  Mean Sq F value    Pr(>F)
data_x      1 50383409 50383409  1864.1 < 2.2e-16 ***
Residuals 364  9838270    27028
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De acuerdo a la tabla ANOVA, el modelo de regresión es altamente significativo, como lo indica el valor extremadamente bajo del p-valor (menor que 2.2e-16) en la fila correspondiente a la variable 'data_x'. Esto significa que la variable 'data_x' tiene una influencia significativa en la variable de respuesta 'data_y'.

La hipótesis nula para la prueba de significancia del modelo es que todos los coeficientes de regresión son iguales a cero, lo que indica que no hay relación lineal entre las variables. Como el p-valor es tan bajo, se rechaza la hipótesis nula y se concluye que hay suficiente evidencia para decir que hay una relación lineal significativa entre las variables 'data_x' y 'data_y'.

6. **Con un nivel de significancia de 0.05 pruebe la hipótesis de que la pendiente es diferente de cero.**

```
# Con un nivel de significancia de 0.05 pruebe la hipótesis que
# de la pendiente es diferente de cero. Se hace con F value: 1864.1
pf(0,1,364)
```

Con un nivel de significancia de 0.05, se realiza la prueba de hipótesis para determinar si la pendiente de la recta de regresión es diferente de cero. Usando la distribución t de Student, se calcula el valor crítico de t para 364 grados de libertad y un nivel de significancia de 0.05, lo que resulta en un valor crítico de t de aproximadamente 1.96. Luego, se calcula el valor de t para la pendiente de la recta de regresión y se compara con el valor crítico de t. Si el valor de t calculado es mayor que el valor crítico de t, entonces se rechaza la hipótesis nula de que la pendiente es igual a cero y se concluye que hay suficiente evidencia estadística para sugerir que hay una relación significativa entre las variables predictoras y la variable respuesta.

En este caso, el valor calculado de t es mayor que el valor crítico de t, y el valor de p asociado es muy cercano a cero (0), lo que indica una fuerte evidencia en contra de la hipótesis nula. Por lo tanto, se rechaza la hipótesis nula y se concluye que hay una relación significativa entre la fecha y la TRM

Codigo GitHub: (Orozco, 2023)

7. **Halle e interprete R^2 y r.**

$$R^2 = 1 - \frac{SCE}{STC}$$

```
> SCE <- 9838270
> SCE
[1] 9838270
> STC <- 50383409 + 9838270
> STC
[1] 60221679
> R2 <- 1 - (SCE/STC)
> R2
[1] 0.8366324
```

R^2: El 83.67% de la variación total en la TRM puede ser explicada por la fecha a través de la recta de regresión. Por lo tanto, la calidad del ajuste del modelo es buena, ya que la mayoría de la variación de la TRM puede ser explicada por la fecha.

```
> r <- sqrt(R2)
> r
[1] 0.9146761
```

r: Esto indica que hay una fuerte correlación positiva entre la variable de respuesta y la variable predictor, lo que significa que a medida que la fecha aumenta, también lo hace la TRM en general.

Codigo GitHub: (Orozco, 2023)

8. Mencione las ventajas y desventajas de la regresión simple

Regresión simple	
Ventajas	**Desventajas**
1. Es una técnica simple y fácil de entender. 2. Proporciona una medida de la fuerza y la dirección de la relación entre dos variables. 3. Se puede utilizar para predecir valores futuros de una variable a partir de los valores de otra variable. 4. Permite identificar valores atípicos y observaciones influyentes. 5. Se puede utilizar para evaluar la importancia de una variable independiente en la predicción de una variable dependiente.	1. La relación entre las variables debe ser lineal. Si la relación no es lineal, la regresión lineal simple no será adecuada. 2. La regresión lineal simple asume que las variables son independientes. Si hay una relación entre las variables independientes, esto puede afectar la interpretación de los coeficientes de regresión. 3. La regresión lineal simple puede ser sensible a valores atípicos y observaciones influyentes. 4. La regresión lineal simple no puede manejar variables categóricas o nominales, a menos que se utilicen técnicas de codificación adecuadas. 5. Si se utilizan para predecir valores futuros, las predicciones pueden ser imprecisas si las condiciones en las que se basa el modelo cambian significativamente en el futuro.

Codigo GitHub: (Orozco, 2023)

Referencias:

- Orozco, D. C. (15 de 03 de 2023). https://github.com/DavidCalebChaparroOrozco/AnalysisofAlgorithms/tree/main/Class05. Obtenido de https://github.com/

Codigo GitHub: (Orozco, 2023)