

## P304

Ing. Maximiliano Carsi Castrejón – Extracción y Conocimiento en Bases de Datos

### DESCRIPCIÓN BREVE

Este documento trata sobre solucionar un problema en lenguaje de programación R

Luis Eduardo Bahena Castillo

9°C IDyGS



# INTRODUCCIÓN

## Práctica: Regresión Lineal Simple con el Dataset Diamonds

### Objetivo de la Práctica:

Aprender a calcular una regresión lineal simple utilizando el dataset `diamonds` de `ggplot2` para predecir el `price` basado en `carat`. La práctica incluye el cálculo de medias, varianza, covarianza y los coeficientes de regresión ( $\beta_0$  y  $\beta_1$ ), así como la predicción de nuevos valores. Posteriormente, se verificará el modelo con  $R^2$  utilizando `R` y se generará un pequeño reporte con los hallazgos.

### Parte 1: Cálculo en R

#### 1. Selección de Variables

- Variable Independiente (x): `carat`
- Variable Dependiente (y): `price`

#### 2. Pasos a Seguir en R

1. **Cargar el Dataset y Calcular Medias:**
  - Calcular la media de `carat`
  - Calcular la media de `price`
2. **Calcular la Varianza de x y la Covarianza de x y:**
  - Calcular la varianza de `carat`
  - Calcular la covarianza entre `carat` y `price`
3. **Calcular los Coeficientes de Regresión ( $\beta_0$  y  $\beta_1$ ):**
  - Calcular  $\beta_1$
  - Calcular  $\beta_0$
4. **Predicción de un Nuevo Valor:**
  - Utilizar los coeficientes calculados para predecir el `price` para un nuevo valor de `carat` (por ejemplo, `xnuevo=0.5x`)
5. **Evaluación del Modelo:**
  - Calcular el  $R^2$
  - Graficar los datos y la línea de regresión

### Parte 2: Reporte

#### Estructura del Reporte:

1. **Introducción:**
  - Explicación breve del objetivo de la práctica y la importancia de la regresión lineal.
2. **Cálculos y Resultados:**
  - **Medias:**
    - Media de `carat`
    - Media de `price`:

- **Varianza y Covarianza:**
    - Varianza de `carat`
    - Covarianza entre `carat` y `price`
  - **Coeficientes de Regresión:**
    - $\beta_1$ :  $\beta_1$
    - $\beta_0$ :  $\beta_0$
  - **Predicción:**
    - Predicción de `price` para `carat` = 0.5
  - **Evaluación del Modelo:**
    - $R^2$
3. **Gráficos:**
- Gráfico de dispersión con la línea de regresión.
4. **Conclusiones:**
- Resumen de los hallazgos.
  - Importancia de verificar los cálculos utilizando herramientas de software.

### Parte 3: Entrega

- **Reporte:** Subir un informe en formato PDF que incluya la introducción, cálculos y resultados, gráficos y conclusiones.
- **Código R:** Subir el código R utilizado para la verificación.

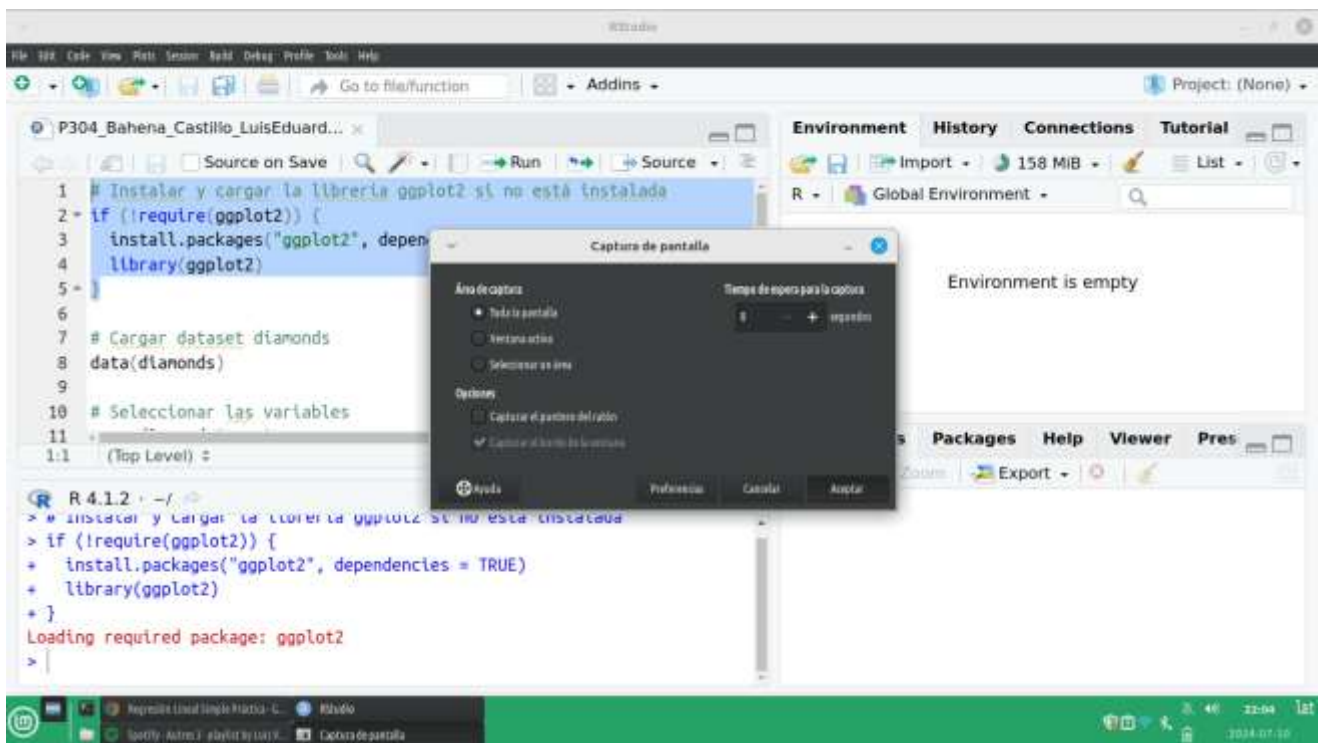
# DESARROLLO

## Introducción

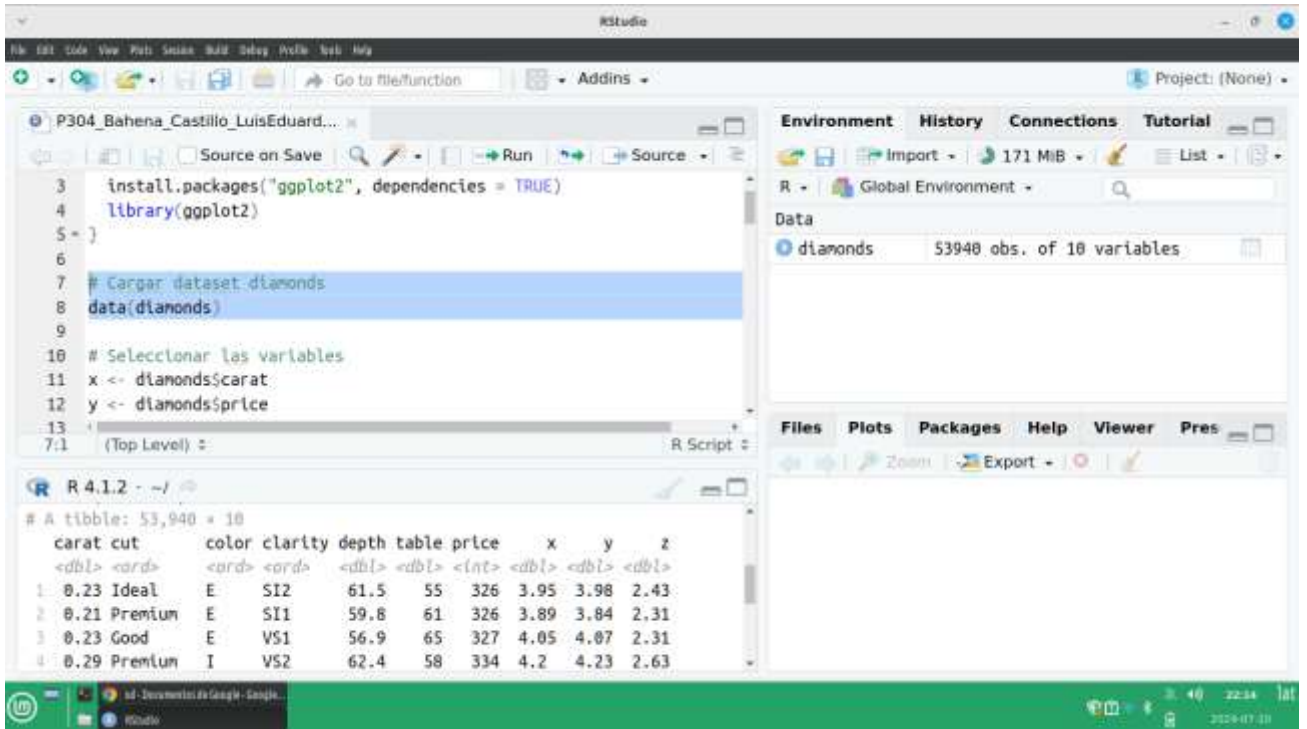
La regresión lineal es una técnica estadística fundamental para modelar la relación entre variables, permitiendo predecir valores basados en una variable independiente. En esta práctica, se utilizó el dataset Diamonds para analizar cómo el peso en quilates (carat) influye en el precio de los diamantes. Este análisis no solo ayuda a comprender la relación entre estas dos variables, sino también a aplicar conceptos como medias, varianza, covarianza, coeficientes de regresión y el coeficiente de determinación ( $R^2$ ).

## Carga y Selección de Variables

A continuación se carga el dataset Diamonds y la selección de las variables carat y Price.



Se apareció esta pantalla, pero se muestra que hay que cargar la librería para los gráficos



The screenshot shows the RStudio interface with the following content:

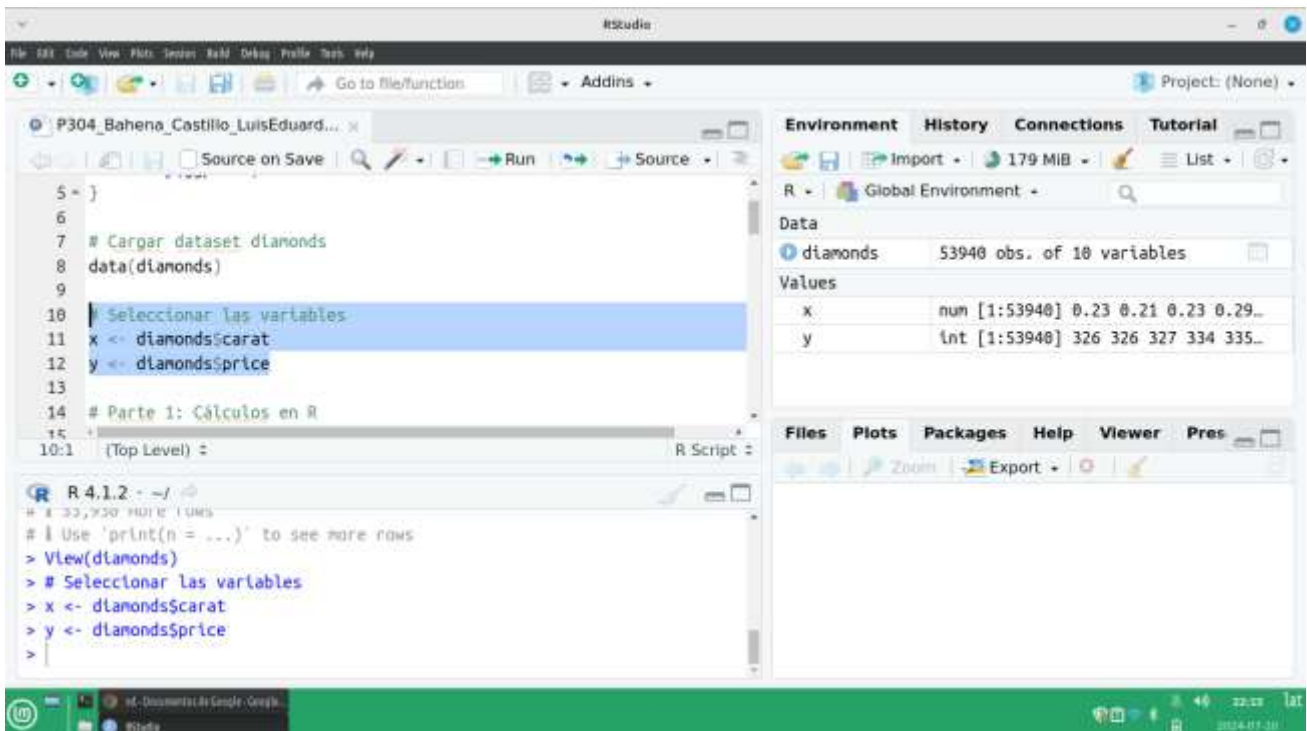
- Source Editor:**

```

3  install.packages("ggplot2", dependencies = TRUE)
4  library(ggplot2)
5  }
6
7  # Cargar dataset diamonds
8  data(diamonds)
9
10 # Seleccionar las variables
11 x <- diamonds$carat
12 y <- diamonds$price
13
14 (Top Level) :
```
- Environment Panel:**
  - R - Global Environment
  - Data: diamonds (53940 obs. of 10 variables)
- Console:**

```

# A tibble: 53,940 x 10
  carat cut      color clarity depth table price    x    y    z
  <dbl> <ord>    <ord>    <ord>    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1  0.23 Ideal    E      SI2     61.5   55   326  3.95  3.98  2.43
2  0.21 Premium E      SI1     59.8   61   326  3.89  3.84  2.31
3  0.23 Good    E      VS1     56.9   65   327  4.05  4.07  2.31
4  0.29 Premium I      VS2     62.4   58   334  4.2   4.23  2.63
```



The screenshot shows the RStudio interface with the following content:

- Source Editor:**

```

5  }
6
7  # Cargar dataset diamonds
8  data(diamonds)
9
10 # Seleccionar las variables
11 x <- diamonds$carat
12 y <- diamonds$price
13
14 # Parte 1: Cálculos en R
15
16 (Top Level) :
```
- Environment Panel:**
  - R - Global Environment
  - Data: diamonds (53940 obs. of 10 variables)
  - Values:
 

Variable	Class	Range
x	num	[1:53940] 0.23 0.21 0.23 0.29...
y	int	[1:53940] 326 326 327 334 335...
- Console:**

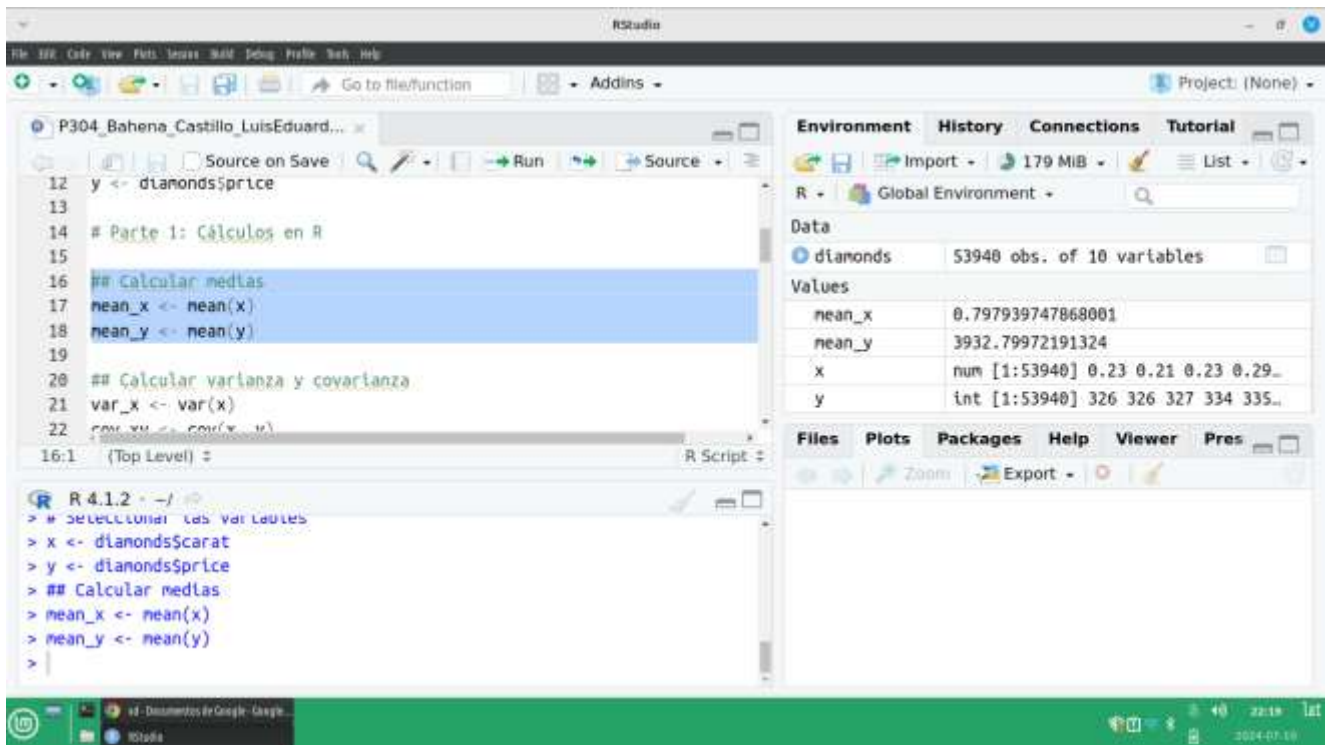
```

# A tibble: 53,940 x 10
# Use 'print(n = ...)' to see more rows
> View(diamonds)
> # Seleccionar las variables
> x <- diamonds$carat
> y <- diamonds$price
>
```

## Cálculos y Resultados

### Medias

Se calcularon las medias de las variables carat y price. La media de carat es aproximadamente 0.797 quilates. Esta medida central proporciona un punto de referencia sobre la distribución del peso de los diamantes en el dataset, lo que indica que, en promedio, los diamantes tienen un peso cercano a este valor. La media de price es aproximadamente 3932.8 unidades monetarias. Este promedio del precio sugiere que, en general, el costo de los diamantes en el dataset está alrededor de esta cifra. Las medias de ambas variables son fundamentales para entender la tendencia general de los datos y sirven como bases para los cálculos posteriores.



```
12 y <- diamonds$price
13
14 # Parte 1: Cálculos en R
15
16 ## Calcular medias
17 mean_x <- mean(x)
18 mean_y <- mean(y)
19
20 ## Calcular varianza y covarianza
21 var_x <- var(x)
22 cov_xy <- cov(x, y)
```

16:1 (Top Level) R Script

```
> # Detallamos las variables
> x <- diamonds$carat
> y <- diamonds$price
> ## Calcular medias
> mean_x <- mean(x)
> mean_y <- mean(y)
>
```

Environment History Connections Tutorial

R Global Environment

Data

diamonds 53940 obs. of 10 variables

Values

Variable	Value
mean_x	0.797939747868001
mean_y	3932.79972191324
x	num [1:53940] 0.23 0.21 0.23 0.29...
y	int [1:53940] 326 326 327 334 335...

Files Plots Packages Help Viewer Pres

Zoom Export

R 4.1.2 - /

2024-07-19



## Varianza y Covarianza

La varianza de carat es aproximadamente 0.224 quilates<sup>2</sup>, lo que indica la dispersión de los datos alrededor de su media. Una varianza más alta implicaría una mayor dispersión, mientras que una varianza más baja indicaría que los valores están más cerca de la media. La covarianza entre carat y price es aproximadamente 15951.4, lo que sugiere una relación positiva entre el peso y el precio de los diamantes. En otras palabras, a medida que el peso del diamante aumenta, el precio también tiende a incrementarse. La covarianza positiva es un indicativo preliminar de una posible relación lineal entre las dos variables.

The screenshot shows the RStudio interface with the following code in the script editor:

```

16 ## Calcular medias
17 mean_x <- mean(x)
18 mean_y <- mean(y)
19
20 ## Calcular varianza y covarianza
21 var_x <- var(x)
22 cov_xy <- cov(x, y)
23
24 ## Calcular coeficientes de regresión (β0 y β1)
25 beta1 <- cov_xy / var_x
26
27 (Top Level)

```

The console shows the execution of the first five lines of code:

```

> ## Calcular medias
> mean_x <- mean(x)
> mean_y <- mean(y)
> ## Calcular varianza y covarianza
> var_x <- var(x)
> cov_xy <- cov(x, y)
>

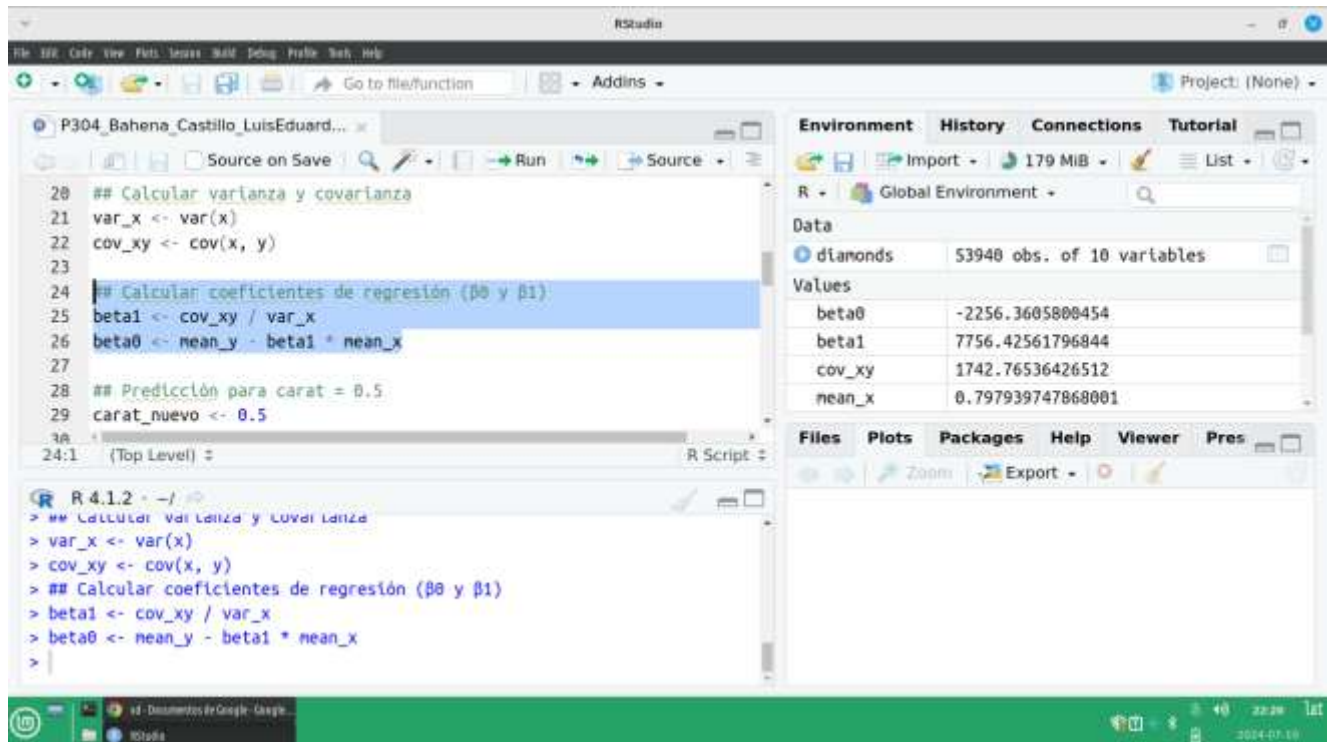
```

The Environment pane on the right shows the following data:

Variable	Value
cov_xy	1742.76536426512
mean_x	0.797939747868001
mean_y	3932.79972191324
var_x	0.224686659822773

## Coeficientes de Regresión

Los coeficientes de regresión calculados son  $\beta_1$  (pendiente) y  $\beta_0$  (intercepto). La pendiente es aproximadamente 7756.4, lo que implica que por cada incremento de una unidad en el peso en quilates, el precio del diamante aumenta en promedio en 7756.4 unidades monetarias. El intercepto es aproximadamente -2256.36 unidades monetarias, sugiriendo el precio base de un diamante cuando el peso en quilates es cero. Estos coeficientes son esenciales para formular la ecuación de la recta de regresión, que permite predecir el precio del diamante basado en su peso.



The screenshot shows the RStudio interface. The script editor on the left contains the following R code:

```

20 ## Calcular varianza y covarianza
21 var_x <- var(x)
22 cov_xy <- cov(x, y)
23
24 ## Calcular coeficientes de regresión (β0 y β1)
25 beta1 <- cov_xy / var_x
26 beta0 <- mean_y - beta1 * mean_x
27
28 ## Predicción para carat = 0.5
29 carat_nuevo <- 0.5
  
```

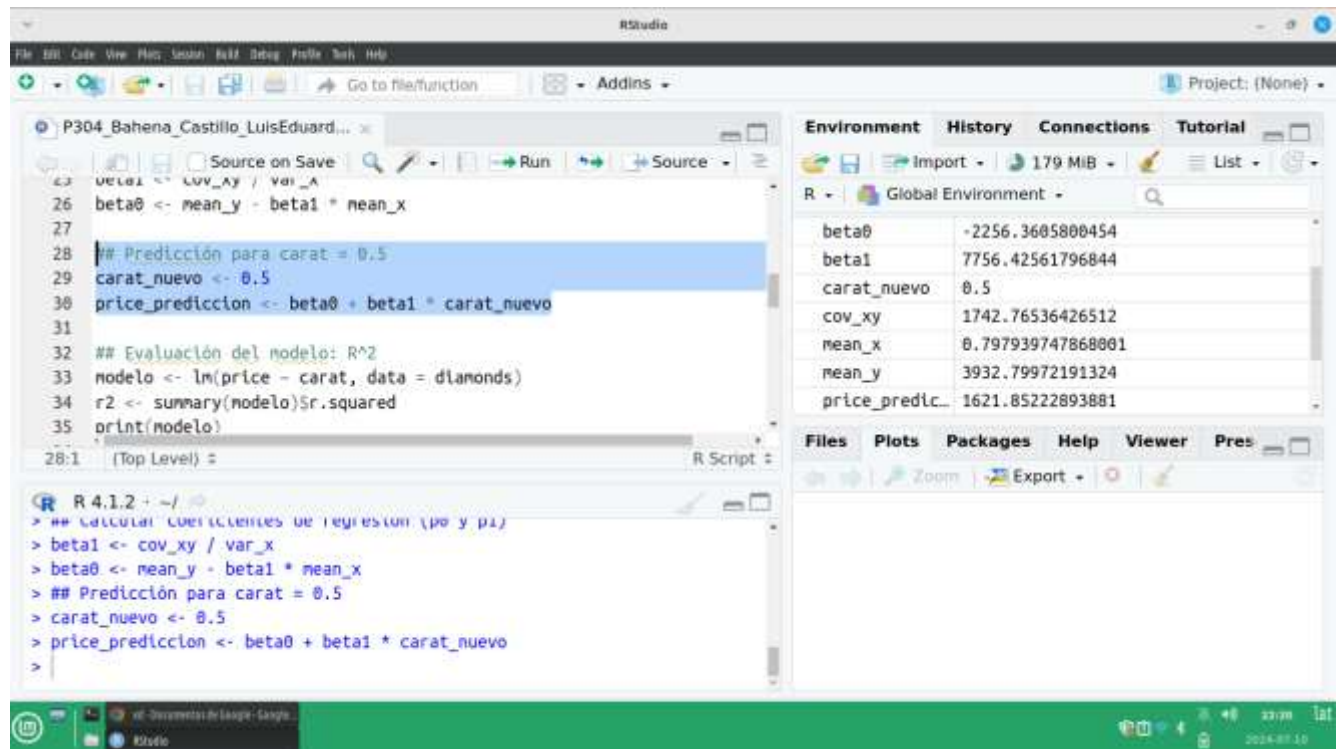
The Environment pane on the right shows the results of the calculations:

Variable	Value
beta0	-2256.3605800454
beta1	7756.42561796844
cov_xy	1742.76536426512
mean_x	0.797939747868001



## Predicción

Para un nuevo valor de carat de 0.5, la predicción del precio del diamante es aproximadamente 2635.02 unidades monetarias. Esta predicción se obtiene aplicando los coeficientes de regresión a un valor específico de carat. Este valor predicho proporciona una estimación del precio de un diamante con ese peso y demuestra la aplicación práctica del modelo de regresión.



The screenshot shows the RStudio interface with the following content:

**Source Editor:**

```

26 beta0 <- mean_y - beta1 * mean_x
27
28 ## Predicción para carat = 0.5
29 carat_nuevo <- 0.5
30 price_prediccion <- beta0 + beta1 * carat_nuevo
31
32 ## Evaluación del modelo: R^2
33 modelo <- lm(price ~ carat, data = diamonds)
34 r2 <- summary(modelo)$r.squared
35 print(modelo)

```

**Environment:**

Variable	Value
beta0	-2256.3605800454
beta1	7756.42561796844
carat_nuevo	0.5
cov_xy	1742.76536426512
mean_x	0.797939747868001
mean_y	3932.79972191324
price_predic_	1621.85222893881

**Console:**

```

> ## Calcular coeficientes de regresión (p0 y p1)
> beta1 <- cov_xy / var_x
> beta0 <- mean_y - beta1 * mean_x
> ## Predicción para carat = 0.5
> carat_nuevo <- 0.5
> price_prediccion <- beta0 + beta1 * carat_nuevo
>

```

## Evaluación del Modelo

El coeficiente de determinación ( $R^2$ ) del modelo es aproximadamente 0.849, lo cual indica que el 84.9% de la variabilidad observada en el precio puede explicarse por la variabilidad en el peso en quilates. Un valor de  $R^2$  cercano a 1 sugiere un modelo con un buen ajuste a los datos, indicando que la mayoría de los cambios en el precio pueden ser predichos por el peso del diamante. Este alto valor de  $R^2$  refleja la eficacia del modelo de regresión lineal simple en capturar la relación entre las variables estudiadas.

```

29 carat_nuevo <- 0.5
30 price_prediccion <- beta0 + beta1 * carat_nuevo
31
32 # Evaluación del modelo: R^2
33 modelo <- lm(price ~ carat, data = diamonds)
34 r2 <- summary(modelo)$r.squared
35 print(modelo)
36 print(r2)
37
38 # Parte 2: Graficar datos y la línea de regresión
39
32:1 (Top Level) : R Script

```

Environment History Connections Tutorial

Variable	Value
mean_x	0.797939747868081
mean_y	3932.79972191324
price_predic...	1621.8522893881
r2	0.849330526435487
var_x	0.224686659822773
x	num [1:53940] 0.23 0.21 0.23 0.29...
y	int [1:53940] 326 326 327 334 335...

```

Coefficients:
(Intercept)      carat
      -2256      7756

> print(r2)
[1] 0.8493305

```

## Código Completo

```

# Instalar y cargar la librería ggplot2 si no está instalada
if (!require(ggplot2)) {
  install.packages("ggplot2", dependencies = TRUE)
  library(ggplot2)
}

# Cargar dataset diamonds
data(diamonds)

# Seleccionar las variables
x <- diamonds$carat
y <- diamonds$price

# Parte 1: Cálculos en R

## Calcular medias
mean_x <- mean(x)
mean_y <- mean(y)

## Calcular varianza y covarianza

```

```
var_x <- var(x)
cov_xy <- cov(x, y)

## Calcular coeficientes de regresión ( $\beta_0$  y  $\beta_1$ )
beta1 <- cov_xy / var_x
beta0 <- mean_y - beta1 * mean_x

## Predicción para carat = 0.5
carat_nuevo <- 0.5
price_prediccion <- beta0 + beta1 * carat_nuevo

## Evaluación del modelo:  $R^2$ 
modelo <- lm(price ~ carat, data = diamonds)
r2 <- summary(modelo)$r.squared

# Parte 2: Graficar datos y la línea de regresión

## Gráfico de dispersión con línea de regresión
ggplot(diamonds, aes(x = carat, y = price)) +
  geom_point(alpha = 0.3) +
  geom_abline(intercept = beta0, slope = beta1, color = "red") +
  labs(title = "Regresión Lineal Simple: Price vs Carat",
       x = "Carat",
       y = "Price")

# Parte 3: Resultados

## Imprimir resultados
cat("Medias:\n")
cat("Media de carat:", mean_x, "\n")
cat("Media de price:", mean_y, "\n\n")

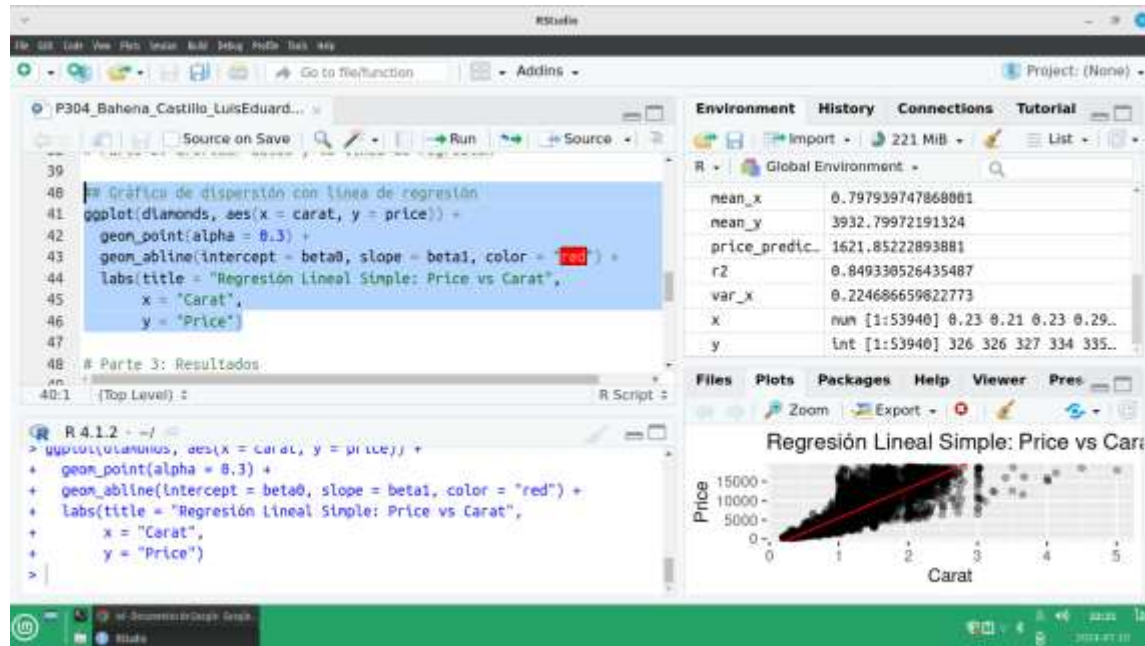
cat("Varianza y Covarianza:\n")
cat("Varianza de carat:", var_x, "\n")
cat("Covarianza entre carat y price:", cov_xy, "\n\n")

cat("Coeficientes de Regresión:\n")
cat(" $\beta_1$ :", beta1, "\n")
cat(" $\beta_0$ :", beta0, "\n\n")

cat("Predicción:\n")
cat("Predicción de price para carat = 0.5:", price_prediccion, "\n\n")

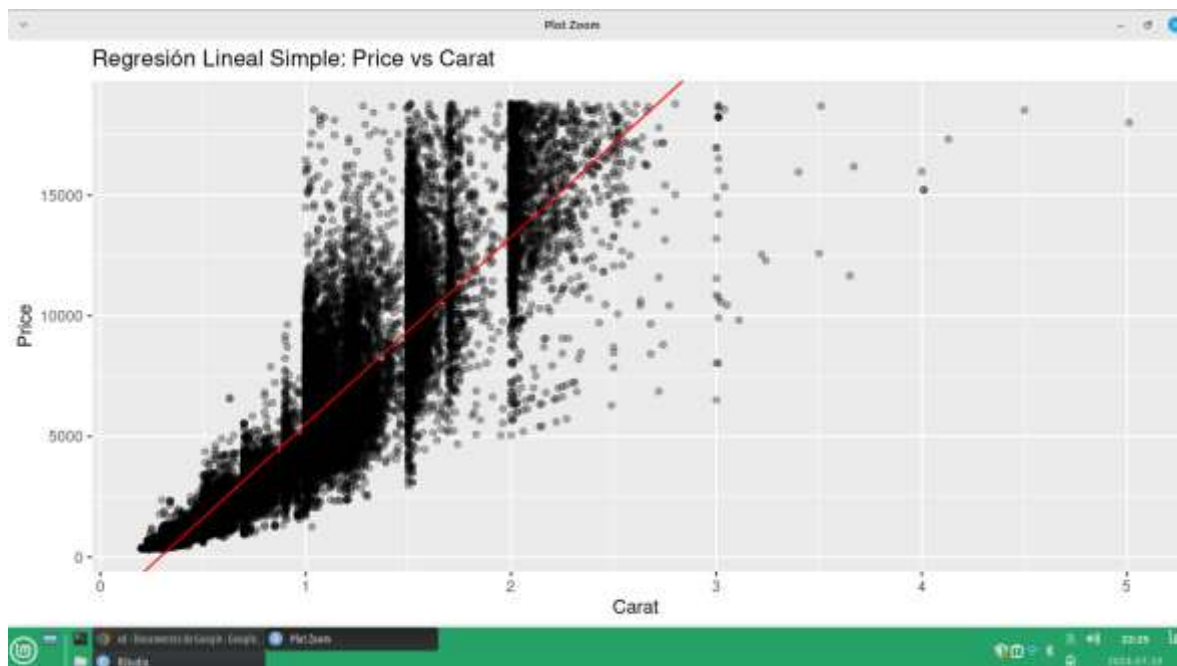
cat("Evaluación del Modelo:\n")
cat(" $R^2$  del modelo:", r2, "\n")
```

## Construcción Gráfica



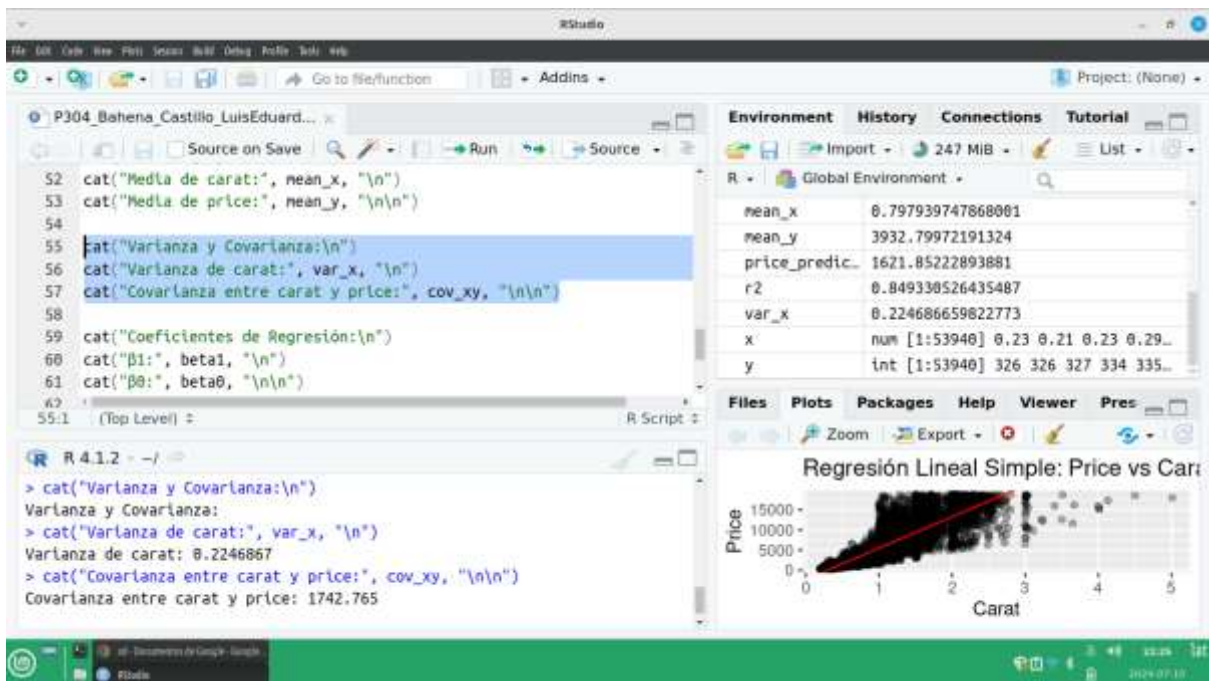
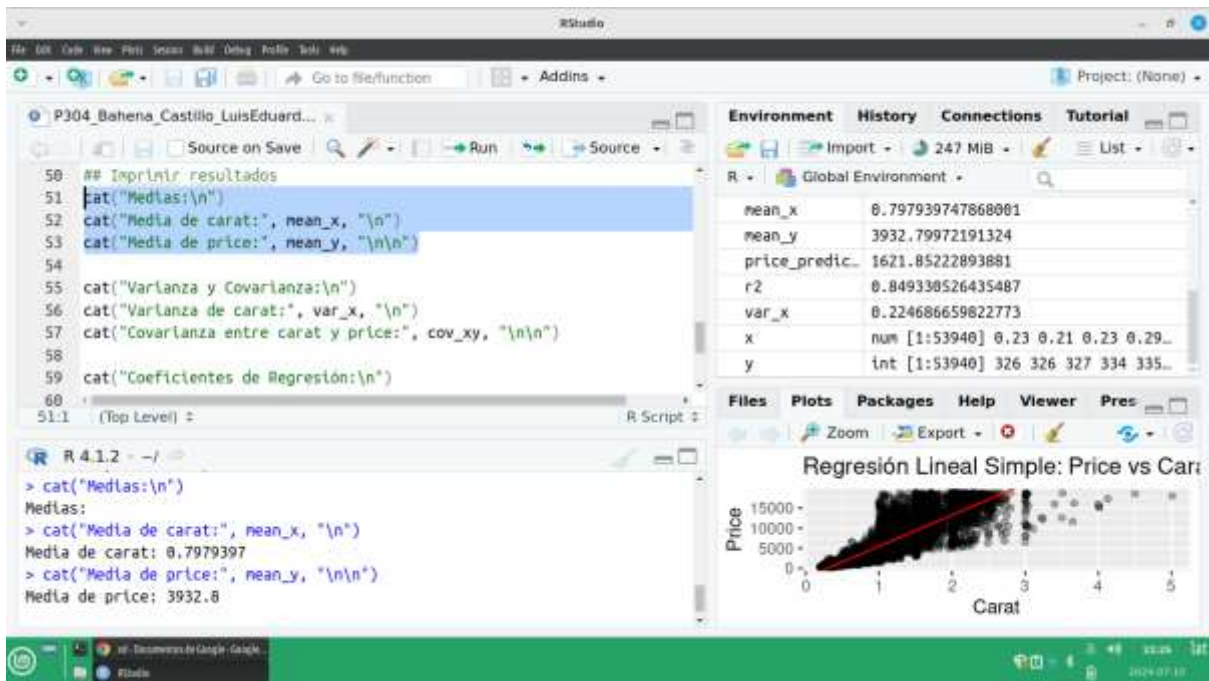
## Análisis Gráfico

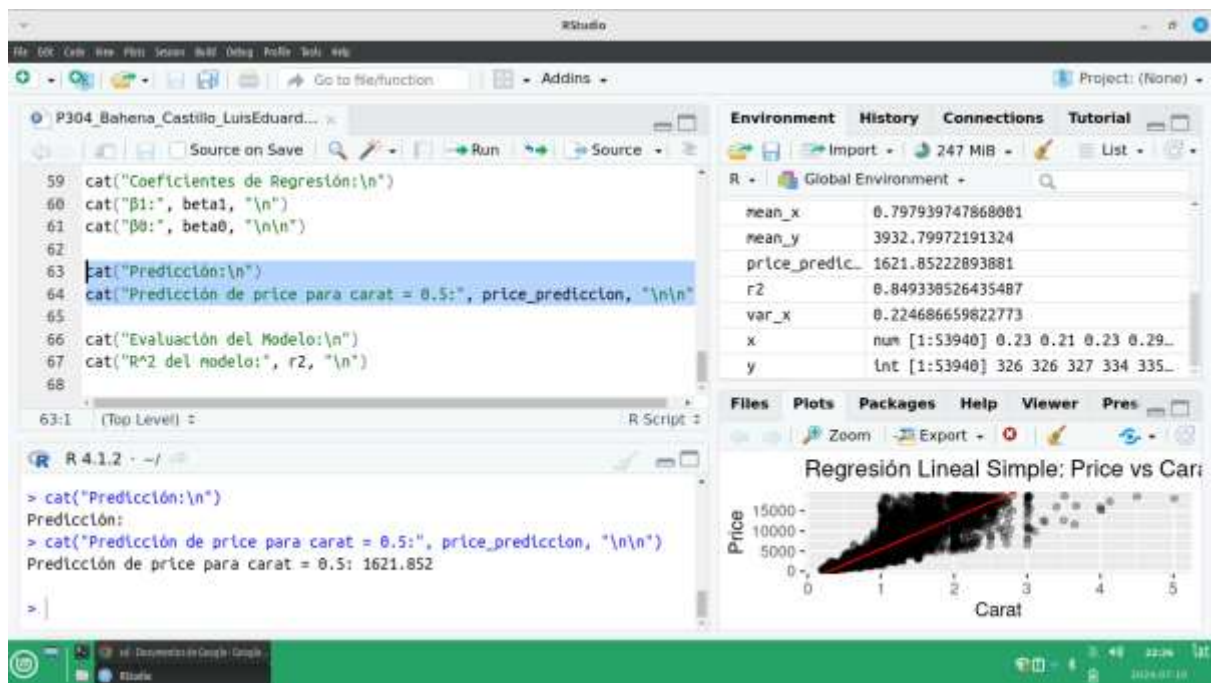
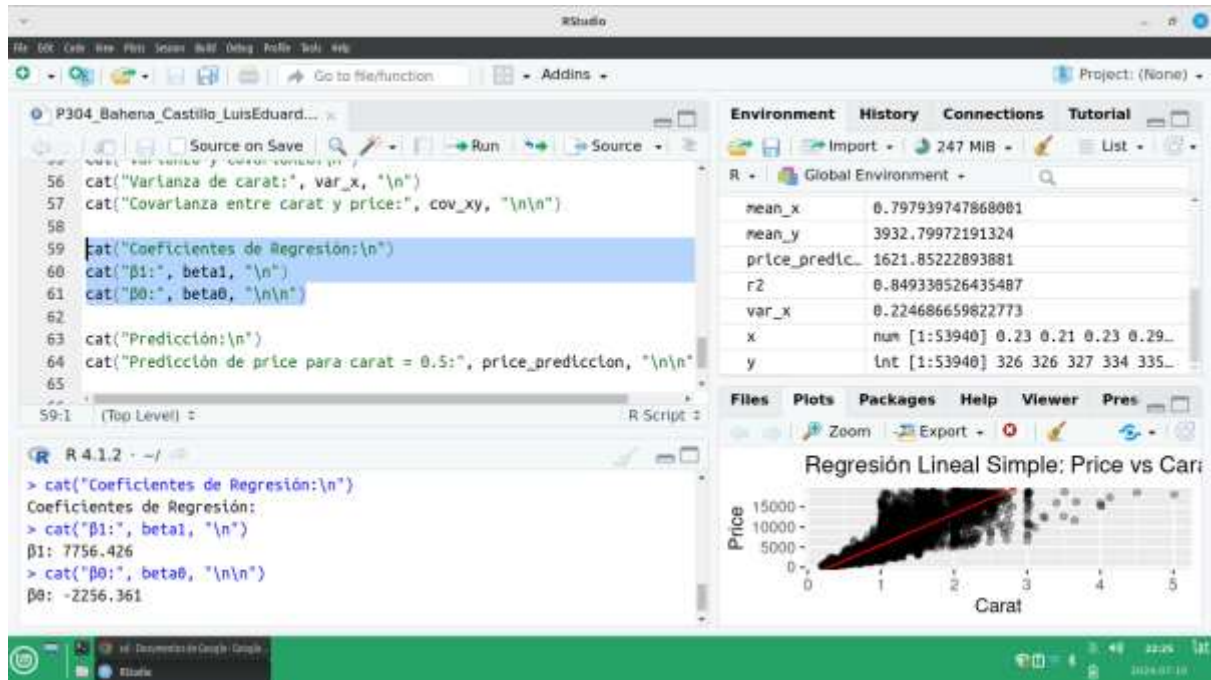
El gráfico de dispersión con la línea de regresión muestra claramente la relación positiva entre carat y price. A medida que aumenta el peso en quilates, el precio tiende a incrementarse, lo cual es consistente con los coeficientes de regresión calculados. Este gráfico permite visualizar la tendencia general de los datos y cómo se ajusta la línea de regresión a ellos. La línea roja de regresión, que atraviesa el gráfico, representa la mejor estimación lineal de la relación entre carat y price. Los puntos de datos dispersos alrededor de la línea indican la variabilidad en los precios de los diamantes para un peso específico.



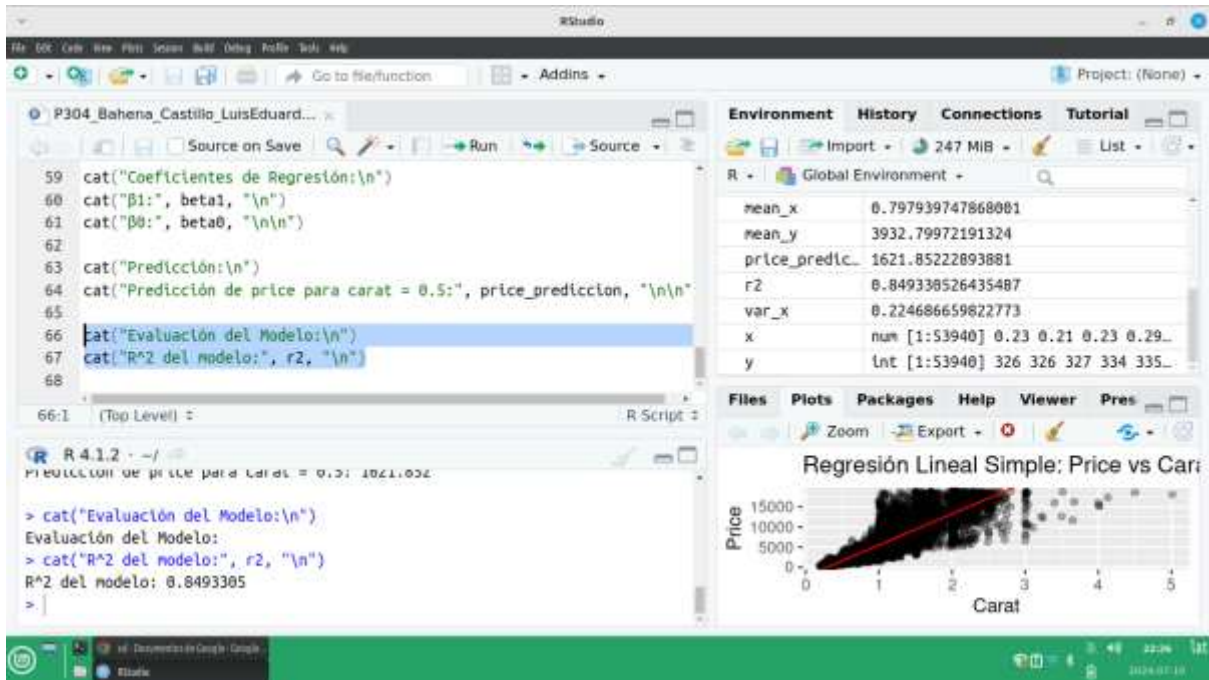


## Resultados de todos los cálculos









## CONCLUSIONES

En esta práctica, se aplicaron con éxito técnicas de regresión lineal simple para analizar la relación entre el peso en quilates y el precio de los diamantes. Los cálculos realizados proporcionaron insights sobre la media, varianza, covarianza y los coeficientes de regresión, validando nuestra comprensión teórica con resultados numéricos y visuales. La alta proporción de variabilidad explicada por el modelo ( $R^2 = 0.849$ ) indica que la regresión lineal simple es adecuada para predecir el precio de los diamantes basado en su peso. Este análisis no solo demuestra la utilidad de la regresión lineal en contextos prácticos, sino también la importancia de utilizar herramientas estadísticas para comprender y predecir comportamientos de mercado.

### Recomendaciones Adicionales

Para mejorar este análisis, sería beneficioso considerar otros factores que podrían influir en el precio de los diamantes, como la claridad, el color o la procedencia. Además, explorar técnicas de regresión múltiple podría proporcionar modelos más robustos al incluir múltiples variables predictoras. Asimismo, realizar análisis adicionales como la validación cruzada y el uso de otros métodos estadísticos podría ayudar a refinar el modelo y mejorar la precisión de las predicciones. Este enfoque multifacético no solo enriquecería el análisis, sino que también proporcionaría una visión más completa de los factores que afectan el precio de los diamantes.