

P301

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: Comparación de Regresión Lineal con Diferentes Variables Predictoras en el Dataset mtcars

Objetivo de la Práctica:

Comparar el ajuste de diferentes modelos de regresión lineal simple utilizando varias columnas del dataset `mtcars` como variables predictoras (x) para predecir la variable `mpg` (y). Los estudiantes deben evaluar los modelos calculando el R^2 y el MSE, y determinar cuál modelo proporciona el mejor ajuste.

Instrucciones:

- Preparación del Entorno:**
 - Asegúrate de tener R y RStudio instalados.
 - Cargar el dataset `mtcars` que viene incluido en R.
- Cálculo de Medias y Varianzas:**
 - Para cada variable independiente (x) seleccionada, calcular las medias y varianzas necesarias.
- Cálculo de los Coeficientes de Regresión:**
 - Calcular los coeficientes de regresión para cada variable predictora.
- Predicción de Valores y Evaluación del Modelo:**
 - Calcular los valores predichos, R^2 y MSE para cada modelo.
- Comparación y Análisis:**
 - Comparar los valores de R^2 y MSE entre los diferentes modelos.
 - Crear una tabla comparativa y subirla junto con un informe de los resultados.

Variables a Utilizar:

Seleccionar al menos 3 (mejores) de las siguientes variables como predictoras (x) para predecir `mpg` (y):

- `disp`: Desplazamiento (pulgadas cúbicas)
- `hp`: Caballos de fuerza
- `drat`: Relación del eje trasero
- `qsec`: Tiempo de 1/4 de milla
- `wt`: Peso (1000 lbs)
- `vs`: Forma del motor (0 = V-shaped, 1 = straight)
- `am`: Tipo de transmisión (0 = automática, 1 = manual)
- `gear`: Número de marchas
- `carb`: Número de carburadores

Informe:

1. **Introducción:**

- Explicación breve del objetivo de la práctica y la importancia de evaluar diferentes modelos de regresión.

2. **Metodología:**

- Descripción de los pasos seguidos para calcular los coeficientes de regresión, R^2 y MSE para cada variable predictora.
- Explicación de cómo se seleccionaron las variables predictoras.

3. **Resultados:**

- Presentación de la tabla comparativa de los resultados.
- Gráficos de dispersión con las líneas de regresión ajustadas para cada modelo.

4. **Análisis:**

- Comparación de los valores de R^2 y MSE entre los diferentes modelos.
- Identificación del modelo con el mejor ajuste (mayor R^2 y menor MSE).
- Discusión sobre los posibles motivos de las diferencias en el rendimiento de los modelos.

5. **Conclusiones:**

- Resumen de los hallazgos.
- Importancia de seleccionar la variable predictora adecuada para el modelado de regresión.

Entrega:

- **Tabla Comparativa:** Subir una tabla con los resultados de R^2 y MSE para cada variable predictora.
- **Informe:** Subir un informe en formato PDF que incluya la introducción, metodología, resultados, análisis y conclusiones.

DESARROLLO

La regresión lineal es una técnica estadística fundamental que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. Este informe detalla el desarrollo y la aplicación de una función de regresión lineal en R, que se puede aplicar a cualquier conjunto de datos para predecir una variable en función de otra.

Introducción

En esta práctica, se explora la capacidad predictiva de diferentes variables del dataset mtcars sobre la eficiencia de combustible (mpg) de varios modelos de automóviles. Este análisis es crucial porque permite entender cómo características específicas como caballos de fuerza (hp), tiempo de 1/4 de milla (qsec) y peso (wt) influye en el consumo de combustible. Además, ayudará a evaluar cómo se comparan entre sí los modelos de regresión lineal simple en términos de precisión.

La eficiencia de combustible es un aspecto esencial tanto para los fabricantes de automóviles como para los consumidores, ya que impacta directamente en los costos operativos y las emisiones ambientales. Por lo tanto, identificar las variables que mejor

predicen el consumo de combustible puede proporcionar insights valiosos para mejorar el diseño de vehículos y fomentar prácticas de conducción más eficientes.

En esta práctica, se utilizará el dataset mtcars, un conjunto de datos ampliamente utilizado que contiene información detallada sobre varias características de automóviles. Las variables que he seleccionado para este análisis son hp (caballos de fuerza), qsec (tiempo en recorrer 1/4 de milla en segundos) y wt (peso del automóvil en miles de libras). Estas variables fueron elegidas porque representan factores mecánicos y de rendimiento que intuitivamente podrían tener un impacto significativo en la eficiencia de combustible.

Metodología

Carga de Datos y Preparación del Entorno

El primer paso consiste en cargar el dataset mtcars en el entorno de trabajo. Este conjunto de datos incluye información detallada sobre varios aspectos de automóviles, lo que permitirá acceder a las variables hp, qsec y wt para utilizarlas como predictores en los modelos de regresión. La carga y preparación adecuada de los datos son esenciales para asegurar la integridad del análisis y la validez de los resultados que se obtendrán.

Cálculo de Medias y Varianzas

Para cada variable seleccionada como predictora, se procede a calcular la media y la varianza. Estos cálculos proporcionan estadísticas descriptivas fundamentales que me permiten entender la distribución y dispersión de cada variable en el conjunto de datos. La media me indica el valor central de la variable, mientras que la varianza me muestra cómo se dispersan los datos alrededor de la media. Esta información es crucial para evaluar la naturaleza de las variables predictoras antes de ajustar los modelos de regresión.

Ajuste de Modelos de Regresión Lineal

Se ajustará modelos de regresión lineal simple para cada una de las variables predictoras (hp, qsec, wt) con el objetivo de predecir mpg. Este proceso implica calcular los coeficientes de regresión, que determinan la relación lineal entre cada variable predictora y la variable dependiente (mpg). Se evaluará la calidad del ajuste de cada modelo utilizando métricas estándar como el coeficiente de determinación (R^2) y el error cuadrático medio (MSE). Estas métricas permiten cuantificar la capacidad predictiva de cada modelo y su precisión.

Predicción de Valores y Evaluación del Modelo

Utilizando los modelos ajustados, se calcularán los valores predichos de mpg para cada variable predictora. Posteriormente, evaluaré el desempeño de estos modelos mediante el cálculo de R^2 y MSE. R^2 mide la proporción de la variabilidad en mpg que puede explicarse mediante la variable predictora, mientras que MSE evalúa la magnitud de los errores de predicción. Estas métricas permitirán comparar objetivamente la capacidad predictiva de cada modelo y determinar cuál ofrece el mejor ajuste a los datos observados.

Código Completo

```
# Cargar el dataset mtcars
data(mtcars)

# Función para ajustar modelo y calcular métricas
fit_lm <- function(data, x_var, y_var) {
  model <- lm(data[[y_var]] ~ data[[x_var]])
  summary(model)
  return(model)
}

# Ajuste de modelos y cálculo de métricas
models <- list()
for (var in c("hp", "qsec", "wt")) {
  models[[var]] <- fit_lm(mtcars, var, "mpg")
}

# Función para calcular R^2 y MSE
calculate_metrics <- function(model, data, x_var, y_var) {
  predictions <- predict(model, newdata = data)
  r_squared <- cor(data[[y_var]], predictions)^2
  mse <- mean((data[[y_var]] - predictions)^2)
  return(list(R2 = r_squared, MSE = mse))
}

# Calcular métricas para cada modelo
metrics <- list()
for (var in c("hp", "qsec", "wt")) {
  metrics[[var]] <- calculate_metrics(models[[var]], mtcars, var, "mpg")
}

# Crear tabla comparativa de resultados
results <- data.frame(
  Variable = c("hp", "qsec", "wt"),
  R2 = sapply(metrics, function(x) x$R2),
  MSE = sapply(metrics, function(x) x$MSE)
)

# Mostrar la tabla
print(results)
# Graficar dispersión con líneas de regresión ajustadas
par(mfrow = c(1, 3))
for (i in 1:length(models)) {
  plot(mtcars[[names(models)[i]]], mtcars$mpg,
       xlab = names(models)[i], ylab = "mpg", main = names(models)[i])
  abline(models[[i]], col = "blue")
}
```

Capturas

1. Cargar el dataset mtcars

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

```

1 # Cargar el dataset mtcars
2 data(mtcars)
3
4 # Función para ajustar modelo y calcular métricas
5 fit_lm <- function(data, x_var, y_var) {
6   model <- lm(data[[y_var]] ~ data[[x_var]])
7   summary(model)
8   return(model)
9 }
10
11
12

```

The Environment pane on the right shows the 'Data' section with 'mtcars' loaded, containing 32 observations of 11 variables. The console at the bottom shows the output of the 'data(mtcars)' command, displaying a table of car specifications.

	mpg	displacement	horsepower	weight	acceleration	mileage_per_gallon	number_of_gears	number_of_cylinders	quarter_mile_time	highway_mileage_per_gallon
Porsche 914-2	26.0	120.3	91	4.43	2.140	16.70	0	1	5	2
Lotus Europa	30.4	95.1	113	3.77	1.513	16.90	1	1	5	2
Ford Pantera L	15.8	351.0	264	4.22	3.170	14.50	0	1	5	4
Ferrari Dino	19.7	145.0	175	3.62	2.770	15.50	0	1	5	6
Maserati Bora	15.0	301.0	335	3.54	3.570	14.60	0	1	5	8
Volvo 142E	21.4	121.0	109	4.11	2.780	18.60	1	1	4	2

2. Ajustar modelos y cálculo de métricas

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

```

3
4 # Función para ajustar modelo y calcular métricas
5 fit_lm <- function(data, x_var, y_var) {
6   model <- lm(data[[y_var]] ~ data[[x_var]])
7   summary(model)
8   return(model)
9 }
10
11 # Ajuste de modelos y cálculo de métricas
12 models <- list()
13

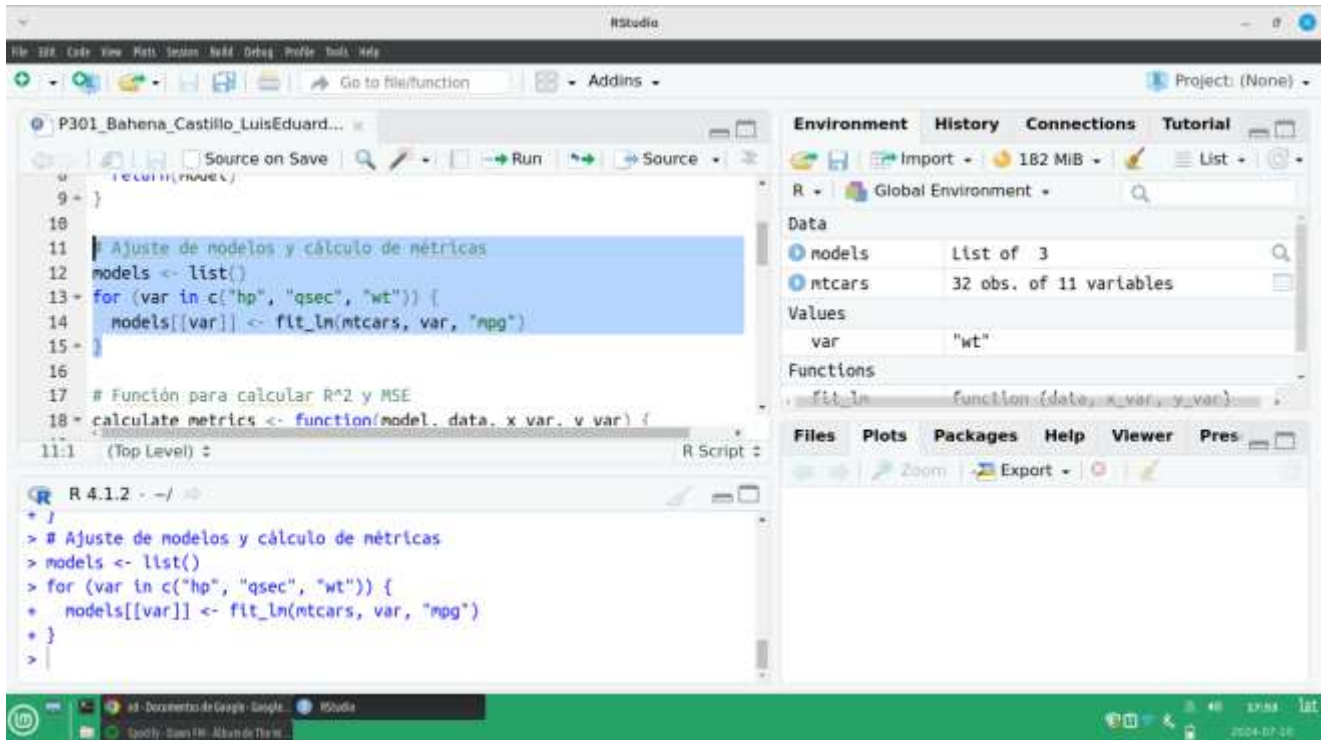
```

The Environment pane on the right shows the 'Functions' section with 'fit_lm' defined as a function (data, x_var, y_var). The console at the bottom shows the output of the 'fit_lm' function, displaying the summary of the linear model fit.

```

> # Función para ajustar modelo y calcular métricas
> fit_lm <- function(data, x_var, y_var) {
+   model <- lm(data[[y_var]] ~ data[[x_var]])
+   summary(model)
+   return(model)
+ }
>

```

```

# P301_Bahena_Castillo_LuisEduard...
9+ }
10
11 # Ajuste de modelos y cálculo de métricas
12 models <- list()
13 for (var in c("hp", "qsec", "wt")) {
14   models[[var]] <- fit_lm(mtcars, var, "mpg")
15 }
16
17 # Función para calcular R^2 y MSE
18 calculate_metrics <- function(model, data, x_var, y_var) {
19
11:1 (Top Level)
  
```

Environment: Global Environment

Data	
models	List of 3
mtcars	32 obs. of 11 variables

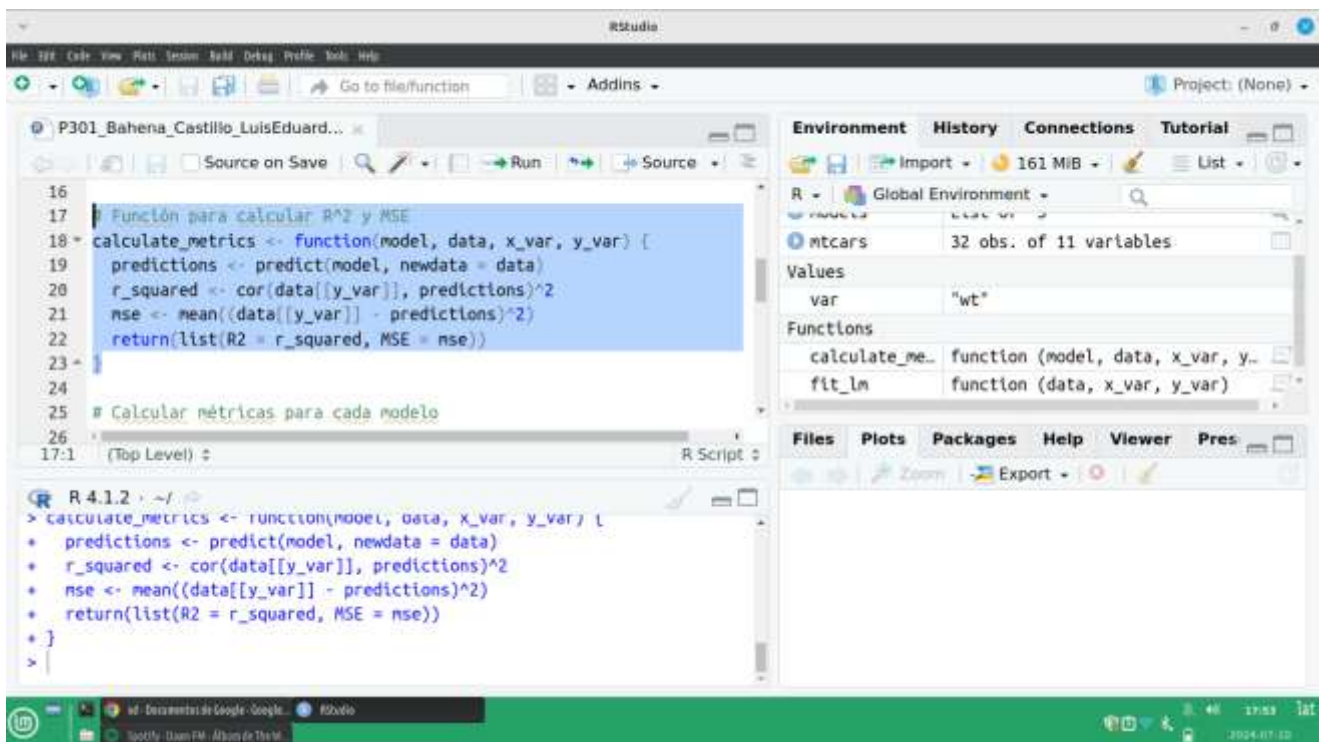
Values

var	
wt	"wt"

Functions

fit_lm	function (data, x_var, y_var)
--------	-------------------------------

3. Función para calcular R^2 y MSE



```

16
17 # Función para calcular R^2 y MSE
18 calculate_metrics <- function(model, data, x_var, y_var) {
19   predictions <- predict(model, newdata = data)
20   r_squared <- cor(data[[y_var]], predictions)^2
21   mse <- mean((data[[y_var]] - predictions)^2)
22   return(list(R2 = r_squared, MSE = mse))
23 }
24
25 # Calcular métricas para cada modelo
26
17:1 (Top Level)
  
```

Environment: Global Environment

Data	
models	List of 3
mtcars	32 obs. of 11 variables

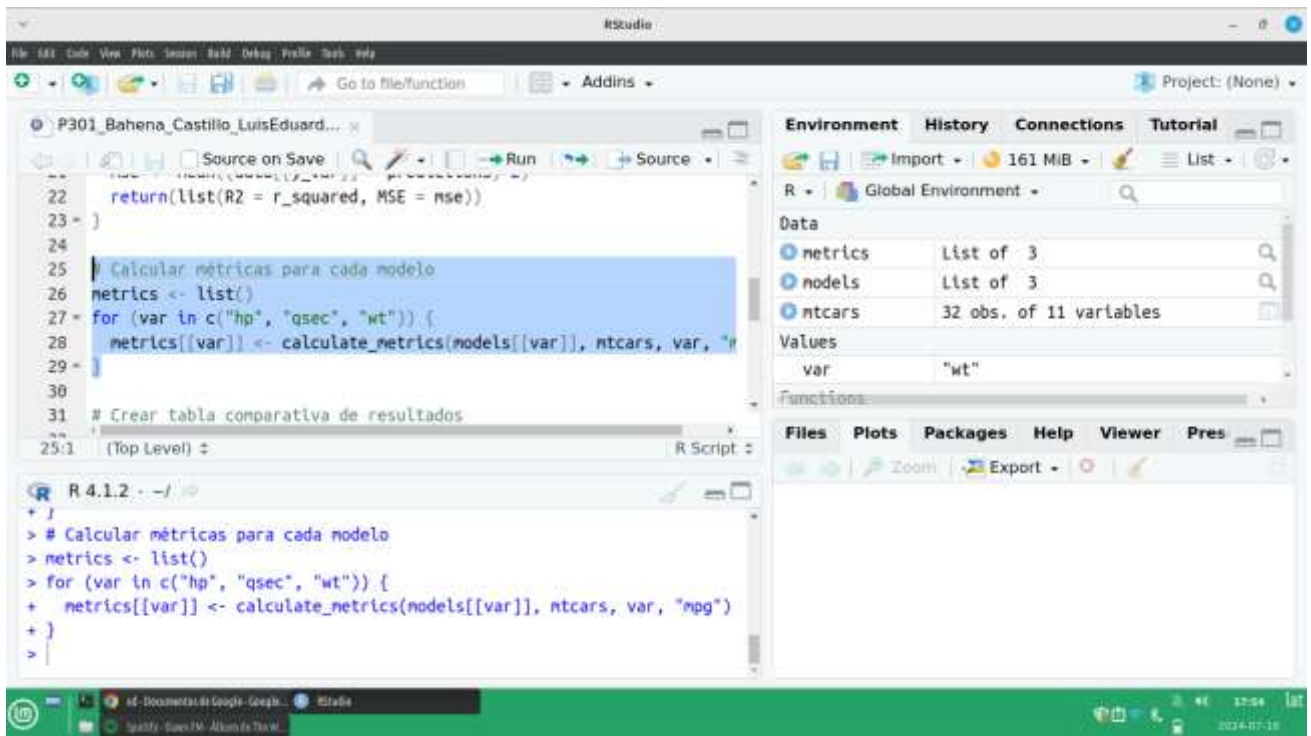
Values

var	
wt	"wt"

Functions

calculate_me...	function (model, data, x_var, y_...
fit_lm	function (data, x_var, y_var)

4. Calcular métricas para cada modelo



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

```

22   return(list(R2 = r_squared, MSE = mse))
23 }
24
25 # Calcular métricas para cada modelo
26 metrics <- list()
27 for (var in c("hp", "qsec", "wt")) {
28   metrics[[var]] <- calculate_metrics(models[[var]], mtcars, var, "mpg")
29 }
30
31 # Crear tabla comparativa de resultados
32
25:1 (Top Level)
  
```

The Environment pane on the right shows the following data objects:

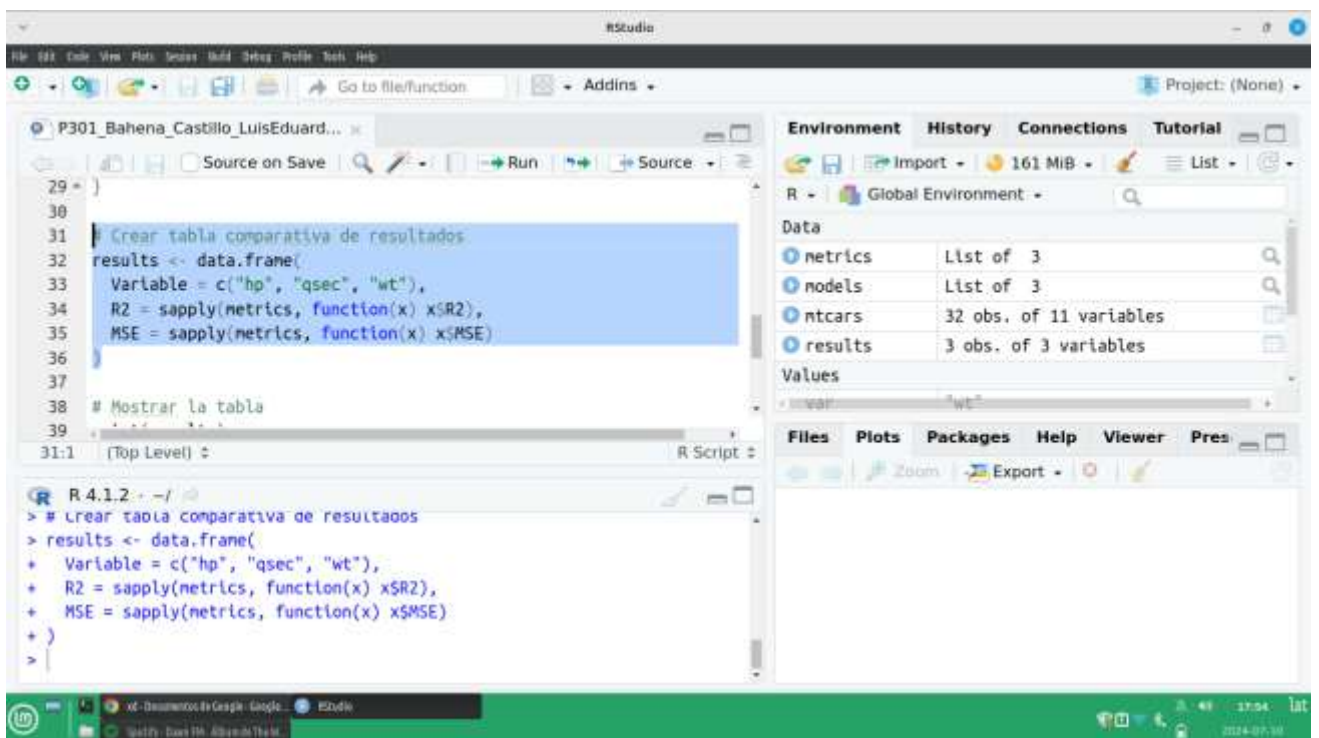
Object	Type	Size
metrics	List of 3	
models	List of 3	
mtcars	32 obs. of 11 variables	

The console shows the execution of the code:

```

> # Calcular métricas para cada modelo
> metrics <- list()
> for (var in c("hp", "qsec", "wt")) {
+   metrics[[var]] <- calculate_metrics(models[[var]], mtcars, var, "mpg")
+ }
>
  
```

5. Crear tabla comparativa de resultados



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

```

29 }
30
31 # Crear tabla comparativa de resultados
32 results <- data.frame(
33   Variable = c("hp", "qsec", "wt"),
34   R2 = sapply(metrics, function(x) x$R2),
35   MSE = sapply(metrics, function(x) x$MSE)
36 )
37
38 # Mostrar la tabla
39
31:1 (Top Level)
  
```

The Environment pane on the right shows the following data objects:

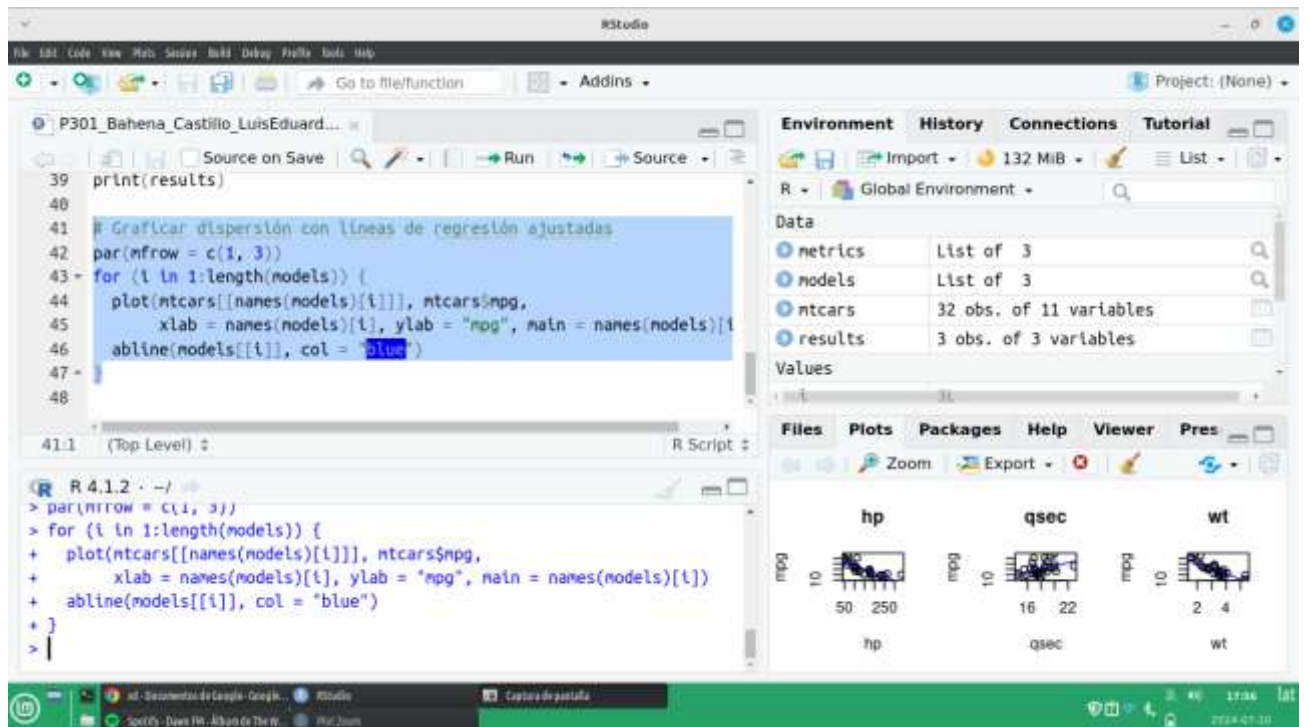
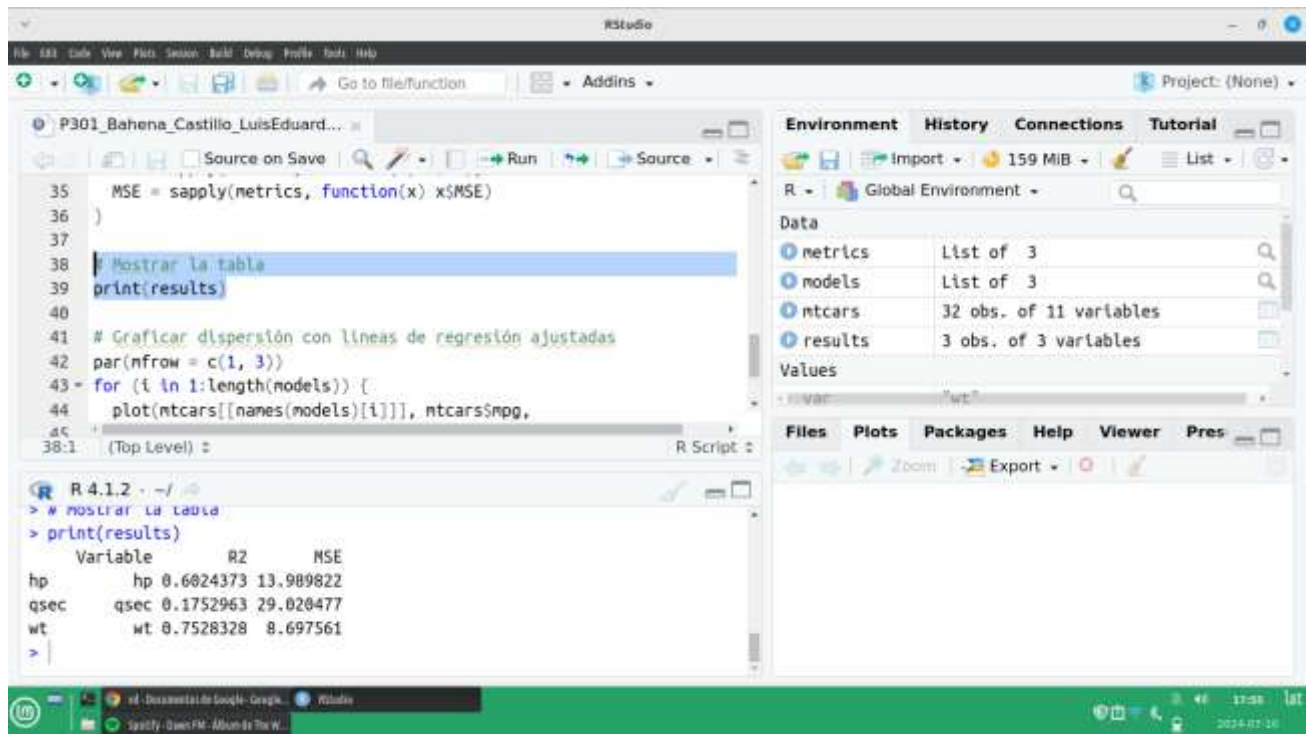
Object	Type	Size
metrics	List of 3	
models	List of 3	
mtcars	32 obs. of 11 variables	
results	3 obs. of 3 variables	

The console shows the execution of the code:

```

> # Crear tabla comparativa de resultados
> results <- data.frame(
+   Variable = c("hp", "qsec", "wt"),
+   R2 = sapply(metrics, function(x) x$R2),
+   MSE = sapply(metrics, function(x) x$MSE)
+ )
>
  
```

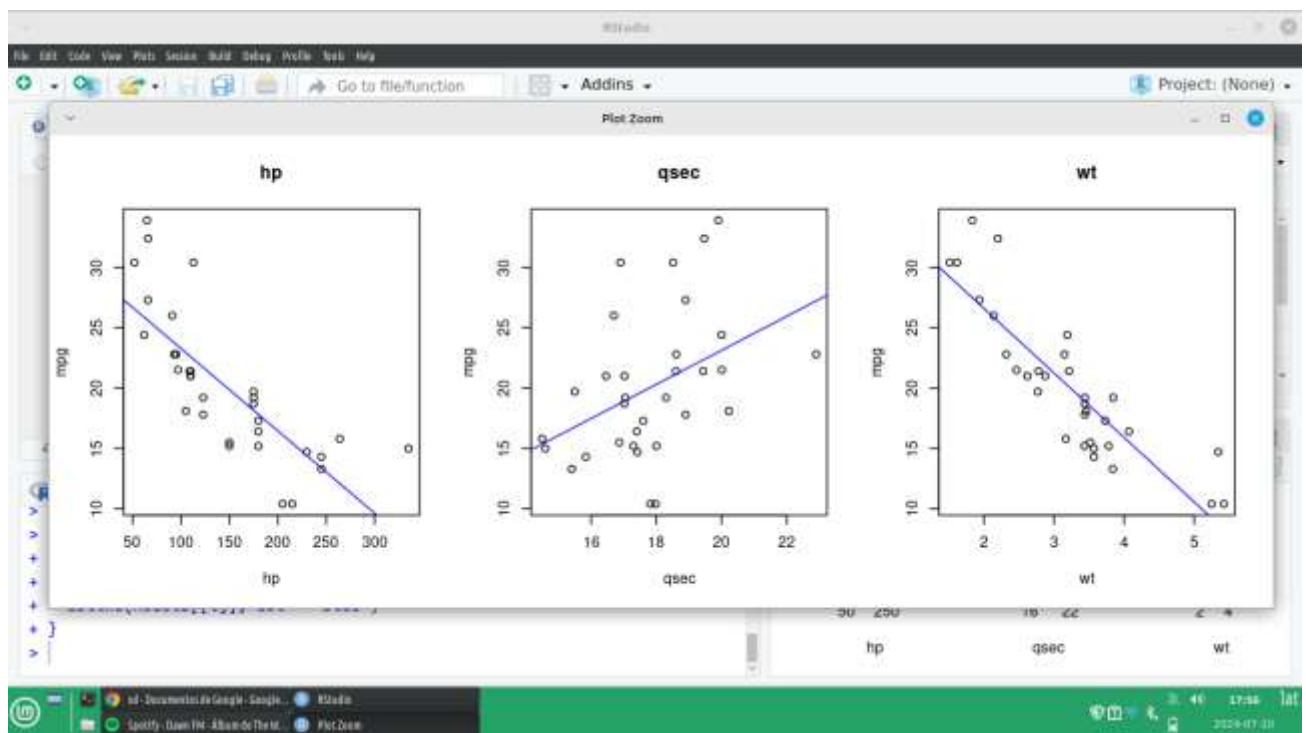

6. Mostrar la tabla comparativa y la gráfica de dsipersión



Resultados

Se presentan los resultados obtenidos de la tabla comparativa de la imagen anterior que incluye los valores de R^2 y MSE para cada uno de los modelos de regresión lineal simple ajustados. Esta tabla permitirá observar de manera clara y directa el rendimiento de cada modelo predictivo en términos de su capacidad para explicar la variabilidad en la eficiencia de combustible (mpg) y la magnitud de los errores de predicción.

Además de la tabla, se visualizan gráficamente los datos junto con las líneas de regresión ajustadas para cada variable predictora seleccionada (hp, qsec, wt). Estos gráficos de dispersión con líneas de regresión permitirán una comprensión visual de cómo cada variable influye en la eficiencia de combustible y la precisión del ajuste del modelo. Observar las líneas de regresión superpuestas a los puntos de datos ayuda a ilustrar la relación lineal entre cada predictor y la variable de interés (mpg), facilitando la interpretación y comparación de los diferentes modelos.



La representación gráfica es especialmente útil para detectar patrones, tendencias y posibles anomalías que no son evidentes únicamente a través de las estadísticas numéricas. Por ejemplo, se podrá identificar si alguno de los modelos presenta un ajuste considerablemente mejor o peor que los demás, lo cual podría sugerir la existencia de una relación más fuerte o débil entre la variable predictora correspondiente y la eficiencia de combustible.

Análisis

El análisis de los resultados se centra en comparar los valores de R^2 y MSE entre los diferentes modelos de regresión lineal simple. R^2 , o coeficiente de determinación, mide la proporción de la variabilidad en la variable dependiente (mpg) que puede explicarse mediante la variable predictora. Un R^2 más alto indica un mejor ajuste del modelo. Por

otro lado, el MSE (error cuadrático medio) evalúa la magnitud de los errores de predicción; un MSE más bajo sugiere un modelo más preciso.

Al comparar estos valores, se identifica cuál de las variables predictoras (hp, qsec, wt) proporciona el mejor ajuste para predecir la eficiencia de combustible. Este análisis ayuda a determinar qué características de los automóviles tienen un impacto más significativo en el consumo de combustible.

Además, se discute las posibles razones de las diferencias en el rendimiento de los modelos, donde las características específicas de cada variable predictora y su relación con la eficiencia de combustible se considerarán para explicar por qué algunos modelos pueden ser más efectivos que otros. Por ejemplo, se podría explorar si el peso del automóvil tiene una influencia más directa y consistente sobre el consumo de combustible en comparación con los caballos de fuerza o el tiempo de 1/4 de milla.

Evaluar las fortalezas y debilidades de cada modelo es fundamental para comprender mejor la dinámica de la predicción de mpg. Esto no solo permitirá identificar el mejor predictor, sino también ofrecer insights sobre cómo mejorar los modelos de predicción y las estrategias para optimizar la eficiencia de combustible.

CONCLUSIONES

Finalmente, se resumen los hallazgos del análisis, destacando la importancia de seleccionar cuidadosamente las variables predictoras en el modelado de regresión. Los resultados obtenidos subrayarán cuáles variables son más efectivas para predecir la eficiencia de combustible y cómo estos insights pueden ser aplicados en contextos prácticos.

Se concluye discutiendo la relevancia práctica de los resultados, sugiriendo cómo podrían ser utilizados para optimizar la eficiencia de combustible en el diseño de automóviles futuros. Los fabricantes de automóviles podrían emplear estos hallazgos para focalizarse en aspectos específicos del diseño vehicular que contribuyan a una mayor eficiencia de combustible.

Además, destaca cómo los consumidores pueden beneficiarse de esta información al tomar decisiones informadas sobre la compra de vehículos, considerando las características que afectan directamente el consumo de combustible. Este ejercicio no solo proporciona un análisis técnico detallado, sino que también subraya la utilidad práctica de la estadística y el modelado predictivo en aplicaciones del mundo real, demostrando cómo las herramientas analíticas pueden contribuir significativamente a la toma de decisiones estratégicas en la industria automotriz y más allá.