Módulo 4. Analítica en R

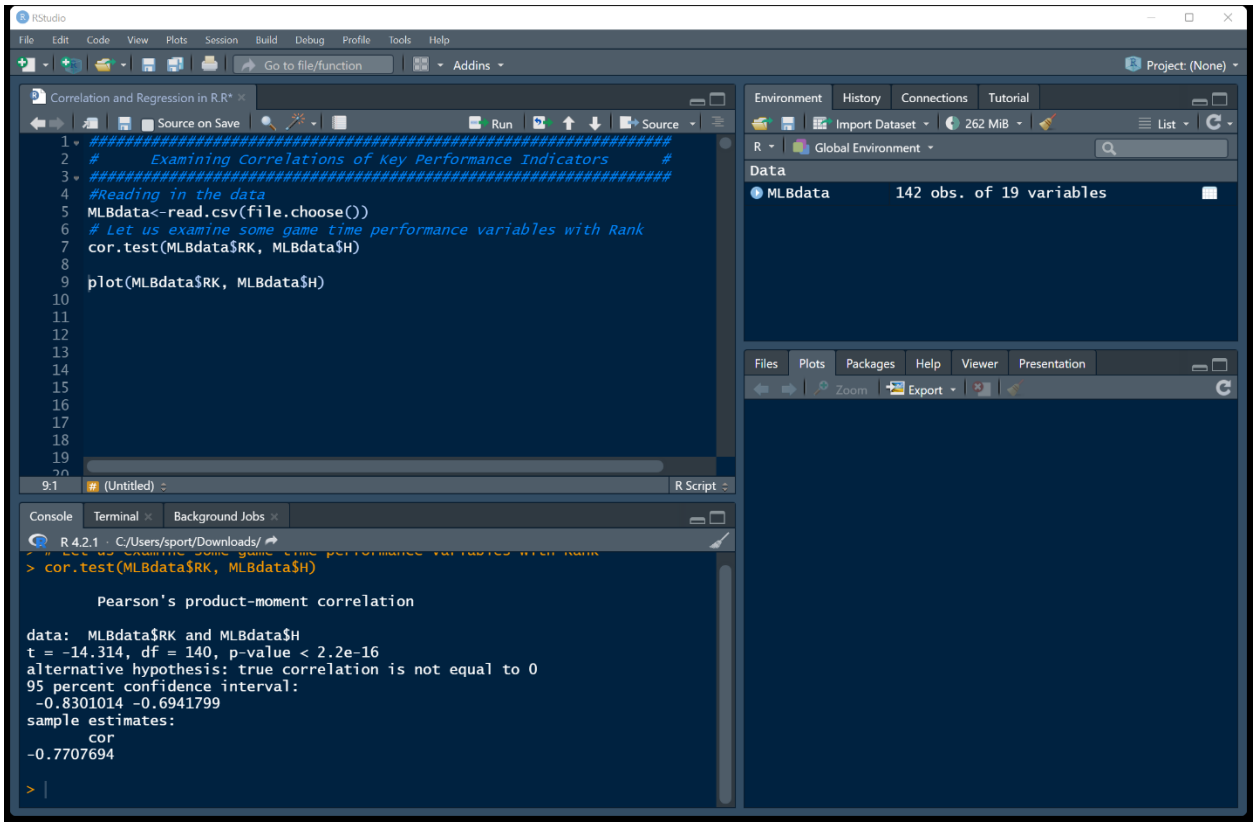
Regresión lineal simple

Este módulo está diseñado para ayudarnos a examinar relaciones y cómo ciertos factores influyen en otro factor; por ejemplo, cómo los indicadores clave de rendimiento (KPI) influyen en el rendimiento. Cubriremos el modelo estadístico más popular y el análisis más comúnmente implementado (llamado regresión), en particular, el modelo de regresión de mínimos cuadrados ordinarios. Aprender sobre la regresión nos preparará para aprender cómo entrenar y probar datos, lo que nos llevará a implementar con éxito modelos de aprendizaje automático supervisado. Aunque el aprendizaje automático no se enseña en este módulo, lo que aprendamos aquí nos preparará para ello.

Primero, comencemos examinando las asociaciones entre dos variables numéricas. Esto se llama análisis de correlación, más conocido como coeficiente de correlación de Pearson.

En R, vamos a importar el conjunto de datos de la MLB y escribir el código para la función de correlación como se muestra a continuación en la imagen.

Imagen 1. Análisis de correlación en R que examina la asociación entre la clasificación y los hits



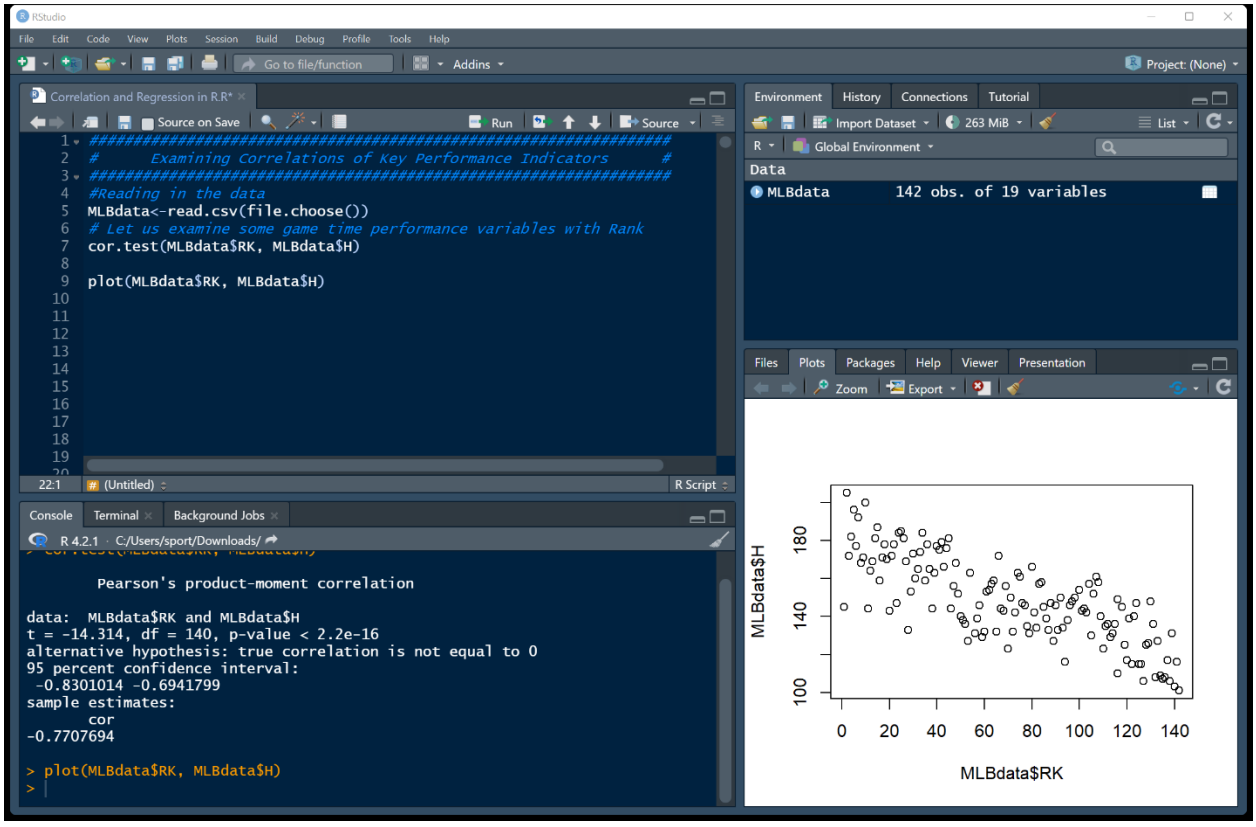
Fuente: captura de pantalla de RStudio [Software], 2011.

La imagen de arriba muestra una correlación de -0,77 en la consola, lo que indica una correlación negativa. Esto implica que, a medida que aumentan los hits, la clasificación disminuye; por lo tanto, hay una mejor clasificación (ya que una clasificación más baja, un número más cercano a 1, es mejor).

Sin embargo, para comprobar que esta correlación sea precisa, también debemos verificar la suposición de linealidad. Si hay linealidad, entonces el valor de la correlación es correcto; de lo contrario, deberíamos reconsiderar otro tipo de análisis que pueda tener en cuenta asociaciones no lineales.

Basándonos en la siguiente imagen que muestra un comando de gráfico, podemos asumir de manera segura la linealidad y que el coeficiente de correlación de -0,77 es una representación precisa de la relación entre la clasificación y los hits.

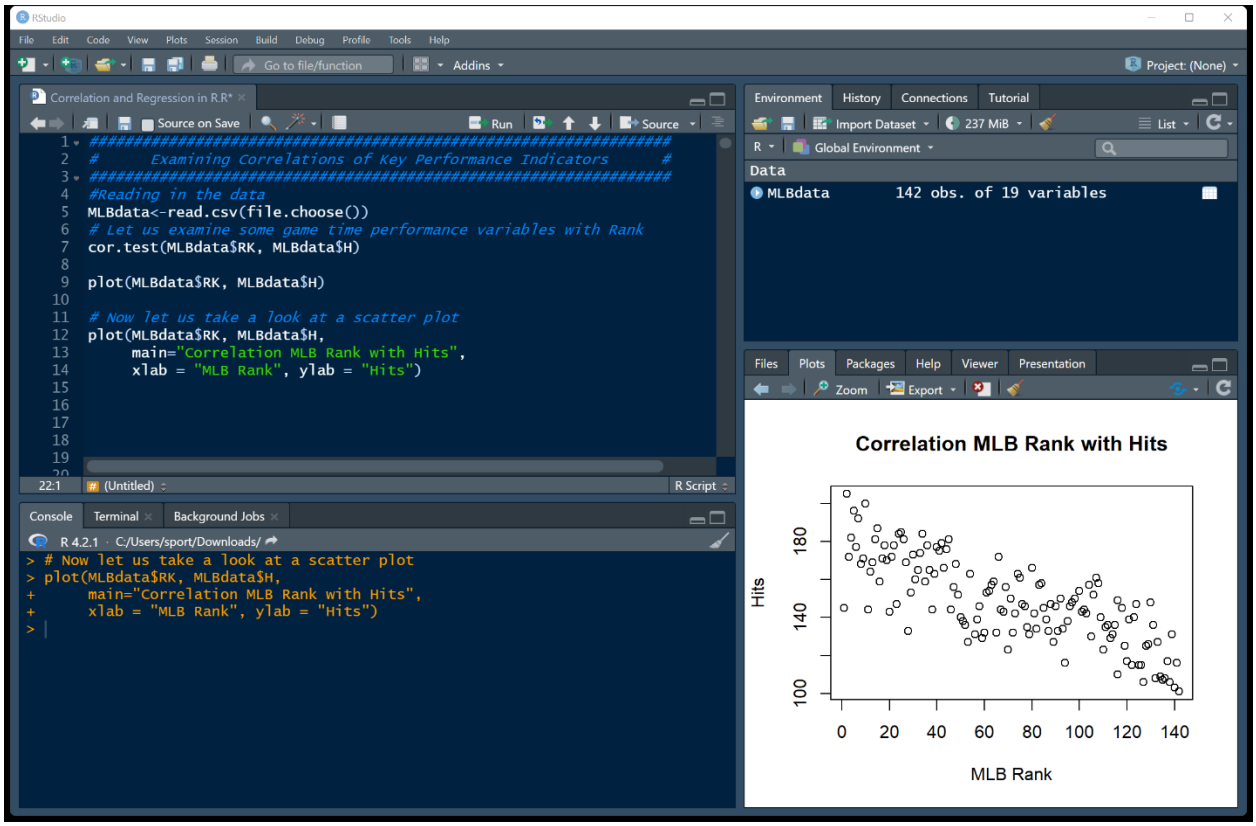
Imagen 2. Gráfico básico de la correlación entre clasificación y hits



Fuente: captura de pantalla de RStudio [Software], 2011.

Aunque podemos mejorar este gráfico con comandos regulares para los títulos de los ejes, como se muestra a continuación, recomendamos que se implemente el paquete "Grammar of Graphics" (ggplot) ya que hay personalizaciones que solo ese paquete puede generar. Sin embargo, si necesitas un gráfico rápido, como suele ser cuando se trabaja en el deporte profesional, un gráfico se puede generar rápidamente con el comando plot como se muestra en la imagen siguiente.

Imagen 3. Gráfico básico en R con título y etiquetas de ejes



Fuente: captura de pantalla de RStudio [Software], 2011.

Dado que hemos comenzado a hacer gráficos, haremos una breve desviación, ya que hemos generado un gráfico básico usando plot(). Esta es una función que viene con el paquete base de R, pero la generación de gráficos en R puede ser una forma de arte si se instalan paquetes como ggplot. Ggplot es un paquete cuyo nombre es un acrónimo de "grammar of graphics".

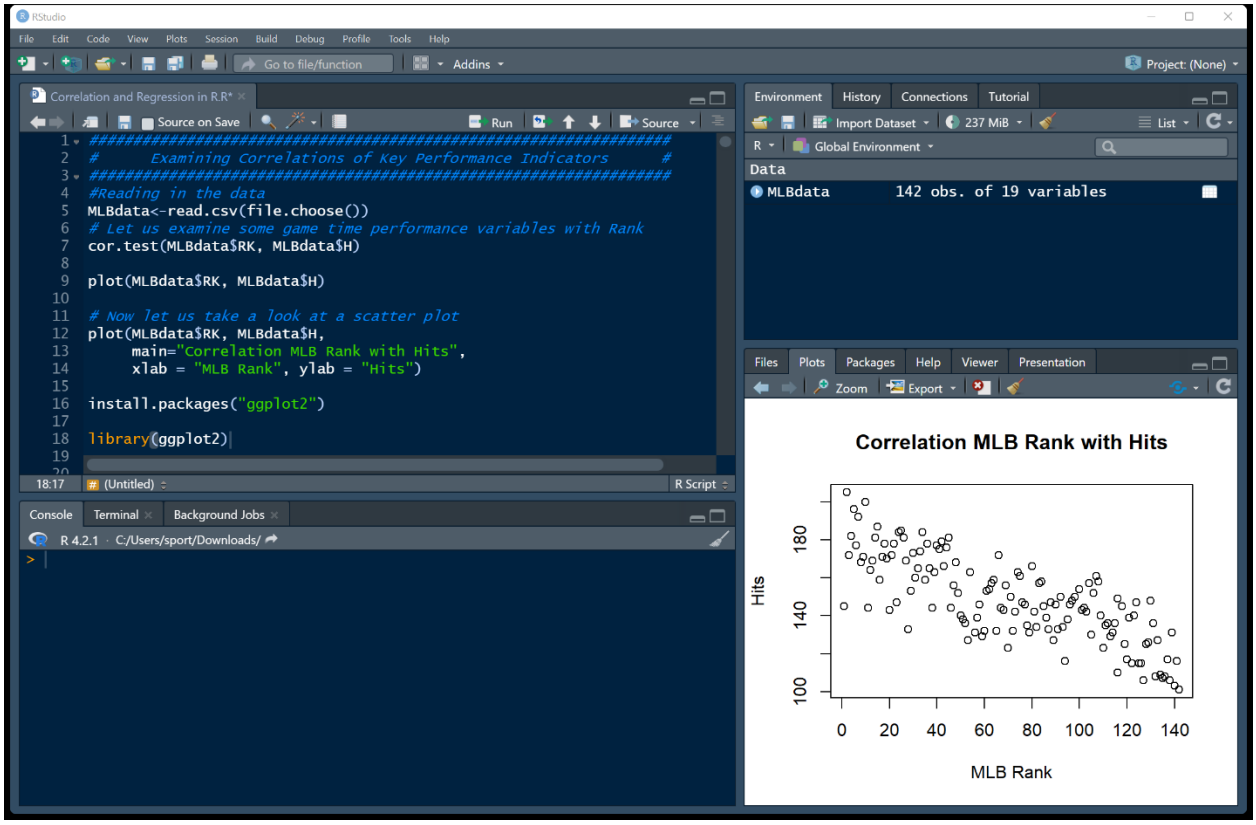
Para instalar Grammar of Graphics, utilizamos la siguiente línea de código:

install.packages("ggplot2")

library(ggplot2)

Hay que Asegurarse de que en el comando install.packages, la función ggplot2 esté dentro de comillas y dentro de la función library no esté, como se muestra en la imagen siguiente:

Imagen 4. Instalación y llamado de la librería ggplot2, "Grammar of Graphics"



Fuente: captura de pantalla de RStudio [Software], 2011.

Es importante identificar que ggplot funciona en base a capas. En la primera línea se incluye el conjunto de datos, seguido de la función aesthetic, que incluye la variable que deseamos colocar en el eje x y luego la variable que ocupará el eje y, seguida por geom_ que representa la figura geométrica, ya sea un gráfico de dispersión, una columna, un diagrama de cajas o un gráfico de líneas.

Las visualizaciones de datos son una de las principales formas de comunicar información. En este módulo, vamos a repasar en detalle qué tipo de visualizaciones son más efectivas según el tipo de datos, los análisis y la audiencia a la que deseamos transmitir la información.

En este algoritmo, la línea que mejor se adapta se identifica al reducir la suma de errores cuadrados, comúnmente conocidos como residuos.

Ecuación de regresión para la población

Ecuación de regresión estimada para muestras

La regresión lineal múltiple es una extensión de la regresión lineal simple donde se incluyen múltiples predictores en el modelo para evaluar su influencia en la variable dependiente.

Conceptos de validación cruzada

K fold

Método de retención (holdout)

Introducir otros valores a

Los conjuntos de entrenamiento y prueba se utilizan comúnmente en la evaluación de algoritmos de aprendizaje automático con el fin de obtener una evaluación del rendimiento del algoritmo dentro del mismo conjunto de datos.

La regresión logística es un tipo de regresión que se implementa cuando el resultado es binario.

Los gráficos regulares de base R se pueden implementar utilizando el siguiente comando: plot().

➢ **Paquete ggplot**

Las variables numéricas pueden ser más adecuadas para ser representadas en algunas de las siguientes opciones:

- Gráficos de dispersión
- Gráficos de barras (columnas y barras)
- Diagramas de caja

Cuando se discute la dispersión y distribución de los datos, se recomienda familiarizarse con los cuartiles, los cuantiles, los deciles y los percentiles, así como con la representación visual más común de estas estadísticas resumidas, que es conocida como el diagrama de caja.

Los cuartiles, que representan el 25% de los datos, se dividen en segmentos, delimitando así el 25% inferior de los datos, normalmente indicado por un "bigote" en el diagrama de caja o el "diagrama de caja y bigotes", que se denomina primer cuartil. Luego, el segundo y tercer cuartiles se conocen comúnmente como la caja en la visualización del diagrama de caja, que representa la porción intermedia de los datos donde se encuentra la mayoría de los datos. Un punto clave a tener en cuenta al leer un diagrama de caja o un diagrama de caja y bigotes es que el segundo y tercer cuartil están divididos por una línea negra delgada que muchos confunden con la media, pero que en realidad representa la mediana, que es el punto de corte en el percentil cincuenta. Finalmente, el percentil

veinticinco superior es el bigote superior, que generalmente se ve más a la derecha y representa los 25 puntos de datos superiores en la distribución de los datos.

Cuantiles

Deciles

Percentiles

- Distribuciones
  - Normal, uniforme, estándar normal, Poisson
  - PDF, FDA

Una variable aleatoria es cualquier cosa que no sabemos y que nos gustaría saber. Cuando intentamos asignar una probabilidad a la variable aleatoria, podemos generar una función de distribución de probabilidad que consta de los diferentes valores que la variable aleatoria puede tomar junto con la probabilidad correspondiente. Esto se llama función de distribución de probabilidad (PDF por sus siglas en inglés).

La función de distribución de probabilidad se utiliza comúnmente cuando trabajamos con variables aleatorias discretas.

Es importante entender que cuando trabajamos con variables aleatorias continuas, como el tiempo en un juego o partido, no hay una probabilidad exacta para ningún valor particular, lo que significa que la probabilidad exacta de cualquier valor dado es en realidad 0. Cuando trabajamos con variables aleatorias continuas, estamos examinando el área desde 0 hasta un valor específico o un rango de valores. Por lo tanto, se implementa la función de distribución acumulativa (FDA).

Si estás interesado en el teorema de Bayes y el análisis bayesiano, esto proporcionará los conceptos básicos para obtener probabilidades para variables aleatorias discretas y probabilidades para áreas de variables aleatorias continuas.

El diagrama de caja es una excelente visualización cuando deseas mostrar la distribución de una forma clara a tu audiencia. Hace un gran trabajo al mostrar automáticamente el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo, así como cualquier valor atípico que pueda estar fuera de los valores mínimo y máximo. En resumen, el diagrama de caja generalmente muestra el conocido resumen de cinco números que se denomina estadísticas descriptivas básicas (más sobre los diagramas de caja en el módulo 4).

Curiosamente, en R y RStudio, cuando implementas la función summary(), R arroja un resumen de seis números en lugar de un resumen de cinco números, que incluye las mismas estadísticas que el diagrama de caja pero también incluye la media.

Otra función que se ejecuta comúnmente en RStudio para obtener estadísticas descriptivas es la función describe() del paquete psych, que proporciona una multitud de descriptivos exploratorios de los datos.

Concepto general de estadísticas
Estadística específica
Diferencia entre estadística y análisis

Es raro encontrar una representación tan buena del resumen estadístico en un gráfico visual. Por lo general, el diagrama de caja se representa con una caja y unos bigotes, donde la caja representa el rango intercuartílico (RIC). El RIC se puede calcular restando Q1 de Q3,
por lo tanto: RIC = Q3 - Q1

Imagen de un diagrama de caja con el resumen de cinco a seis números. ▪
- Diagramas de caja
- Cuartiles
- Resumen
- Descripto por
- Visualización de datos

Otro tipo de visualización de datos que se utiliza comúnmente es el gráfico de dispersión. ¿Qué es exactamente un gráfico de dispersión? Es una visualización en la que se trazan puntos de las variables x e y. Volviendo a lo básico, piensa en la estructura básica x-y como se muestra a continuación:
Imagen de un gráfico x-y
Es la asignación de coordenadas x y en un plano 2D.
¿Cuándo debemos usar un gráfico de dispersión?
La función principal de un gráfico de dispersión es mostrar la relación entre dos variables. Si ambas variables de interés son numéricas, puede ser interesante examinar la correlación entre ellas.

¿Qué ocurre cuando queremos mostrar datos a lo largo del tiempo y las tendencias?

- Gráfico de líneas (es la opción más común).

El gráfico de líneas es la visualización más utilizada para examinar una variable en un período de tiempo determinado. La función para implementar es geom_line() cuando

trabajamos con el paquete ggplot2 en R y RStudio, ya que es un paquete poderoso de visualización de datos.

En este caso particular, el conjunto de datos tiene una variable que es algún tipo de componente de tiempo, ya sea minutos jugados, duración del partido, días de entrenamiento, meses en la temporada, etc. La variable de tiempo suele asignarse a la variable x, y la variable y en el argumento ggplot en R tomará la variable real de interés. Por ejemplo, si queremos hacer un seguimiento del número de asistencias a lo largo del tiempo, entonces el número de asistencias iría en la variable y y el tiempo en x.

¿Cómo solucionar problemas cuando nuestros datos son inherentemente categóricos y queremos generar un gráfico de líneas?

En este escenario, es posible que tengas ciertos puntos de datos individuales para cada día de rendimiento; sin embargo, para visualizar esto como una línea, literalmente tendremos que conectar los puntos. Si ejecutáramos el ggplot regular con la función geom_line(), se generaría un error que indicaría que hay un solo punto de datos para cada tiempo. Por lo tanto, debes conectarlos utilizando el argumento group = dentro de ggplot.

Observa el ejemplo de la imagen y el código R a continuación: sintaxis básica de R, es mejor retener el conocimiento de R intentando resolver un problema.

Conceptos básicos de R, funciones y tipos de datos
Vectores y ordenamiento
Indexación, manipulación de datos y gráficos
Conceptos básicos de programación, ifelse y bucles for para comandos

Referencias

Allaire, J. J. (2011). *RStudio* [Software]. Posit. https://posit.co/downloads/.