***PRÁCTICA 2: ESTADÍSTICA DESCRIPTIVA BIVARIANTE. CORRELACIÓN Y REGRESIÓN LINEAL SIMPLE***

Atanasov Angelov, Daniel – [daniel.atanasov24@estudiantes.uva.es](mailto:daniel.atanasov24@estudiantes.uva.es)

Sanz Tomé, Raúl – [raul.sanz24@estudiantes.uva.es](mailto:raul.sanz24@estudiantes.uva.es)

**Resumen:**

En esta práctica de estadística con Python, hemos trabajado con datos bivariante para analizar la relación entre dos variables mediante la correlación y la regresión lineal simple. Hemos calculado la recta de regresión, interpretado la correlación y realizado predicciones. Gracias a los gráficos y al análisis de los resultados obtenidos, hemos entendido que no siempre hay una relación clara entre las variables, pero cuando la hay, podemos usar una variante para estimar la otra con una gran precisión.

**1. Introducción**

En esta práctica hemos trabajado con un conjunto de datos estadísticos de dos variantes, con los objetivos de relacionarlas y analizar la regresión lineal simple entre ellas. Estos datos son los mismos que empleamos en la Práctica 1: el dataset mpg.

**2. Metodología**

En primer lugar, hemos trabajado con los datos del dataset mpg para aprender los comandos de Python necesarios para la regresión lineal simple y la correlación de dos variables.

Después, hemos realizado el gráfico de dispersión con potencia y aceleración, las rectas de regresión adecuadas para cada caso y, para finalizar, analizamos los resultados de cada uno de estos modelos de representación.

También hemos simplificado el modelo del apartado b y hemos repetido los procedimientos anteriores para volver a analizar los resultados y compararlos con los iniciales.

Para finalizar, hemos realizado un estudio similar para buscar la relación entre aceleración y peso y entre desplazamiento y potencia.

**3. Resultados**

Tras comprender el uso de Python para estos procesos estadísticos, analizamos el gráfico de dispersión con potencia y aceleración. Como el coeficiente de correlación es negativo, las variables poseen una relación inversa y debido a su valor podemos predecir aproximadamente la potencia de un coche en función de su aceleración.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.

A continuación, tras realizar la recta de regresión de potencia sobre aceleración, la cual nos permite estimar los valores de la potencia a partir de los de la aceleración, podemos prever que un coche cuya aceleración es de 10 segundos tiene una potencia de 158 CV. Realizando la recta de regresión de aceleración sobre potencia, podemos estimar que la aceleración de un coche es de 20,21 segundos si su potencia es de 10 CV.

Aplicando el modelo de regresión lineal de aceleración sobre potencia y dado el punto de potencia = 130 CV y aceleración = 17 segundos, observamos que, si la potencia es de 300 CV, el coche tendría una aceleración de 5,88 segundos.

Tras distinguir entre los coches de cuatro y los de ocho cilindros, observamos que los coches de cuatro cilindros poseen una potencia cercana a los 78 CV, mientras que los de ocho cilindros están en el entorno de los 158 CV.

También observamos que existen las siguientes diferencias en la aceleración de ambos tipos de cilindros: los coches de 4 cilindros poseen una aceleración más lenta que los de 8 cilindros y estos tienen una mayor variabilidad y más valores atípicos.

Gráfico, Gráfico de cajas y bigotes

El contenido generado por IA puede ser incorrecto.

Analizando los residuos de los modelos anteriores, observamos que, en los dos modelos que relacionan la potencia y la aceleración, los residuos están distribuidos sin seguir ningún tipo de distribución, por lo que el modelo se puede emplear para realizar predicciones con una gran precisión. En cuanto al modelo de potencia – cilindros, los datos están agrupados de esa forma debido a que dolo existen dos valores para los cilindros, mientras que el hecho de que estén tan dispersos verticalmente indica que existen muchas excepciones, es decir, que la potencia varía mucho dentro de cada grupo, por lo tanto, estamos ante un modelo débil.

Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.Gráfico, Gráfico de dispersión

El contenido generado por IA puede ser incorrecto.Gráfico, Gráfico de líneas

El contenido generado por IA puede ser incorrecto.

Después de eliminar los puntos extraños del modelo del apartado b, hemos guardado los datos en un fichero aparte para repetir los pasos anteriores con estos nuevos datos, excepto el primer apartado, el gráfico de dispersión con potencia y aceleración. Entonces, ahora la potencia de un coche cuya aceleración es de 10 segundos será de 152 CV, mientras que, si la potencia es de 10 CV, la aceleración es de 19,8 segundos. En cuanto al coche de 130 CV y aceleración de 17 segundos, si la potencia es de 300 CV, la aceleración en este nuevo modelo es de 6,3 segundos.

Volvemos a distinguir entre coches de 4 cilindros y coches de 8 cilindros, pero esta vez los resultados que obtenemos son que, aproximadamente, los vehículos de 4 cilindros tienen una potencia de 79 CV y los de 8 cilindros tienen una potencia de 152 CV. Mientras que las gráficas de los residuos apenas varían con respecto a las originales.

Para finalizar, realizando un estudio similar, pero ahora con el modelo completo, observamos que sí es posible hallar una posible relación entre la aceleración y el peso porque los residuos están desordenados en un patrón similar a un abanico y muchos de ellos están agrupados en el centro. Sin embargo, entre desplazamiento y potencia no es posible hallar dicha relación porque el patrón observado es irregular.

**4. Conclusiones**

Después de analizar todos estos datos, llegamos a las conclusiones de que entre la potencia y la aceleración existe una relación inversa, es decir, al aumentar la potencia, la aceleración disminuye (lo cual es bueno porque es una aceleración mayor). Además, tanto el modelo de regresión de potencia sobre aceleración como el de aceleración sobre potencia, son muy fiables para predecir los datos correspondientes a cada uno. También nos fijamos en que la variable de los cilindros no es lo suficientemente buena para establecer los valores de la potencia.

Para finalizar, después de eliminar los residuos de los modelos, observamos que estos modelos mejoran ligeramente, a pesar de que los valores varían muy levemente. Igualmente, los últimos modelos estudiados demuestran que no siempre existe relación entre las variables y que, si existe, las variables pueden estar más o menos relacionadas.