Regresión lineal múltiple y polinomial

Raúl Navarro

Contenido

- Modelo de regresión lineal simple vs. modelo de regresión lineal múltiple
- Supuestos de la regresión lineal múltiple
- Estimación de los coeficientes de regresión
- Interpretación de los coeficientes de regresión
- Evaluación de la calidad del ajuste del modelo
- Selección de variables independientes
- Regresión lineal múltiple con variables categóricas

Modelo de regresión lineal simple vs. modelo de regresión lineal múltiple

En estadística, la regresión lineal es una técnica que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. La regresión lineal puede ser simple o múltiple, dependiendo de si se utiliza una sola variable independiente o varias para explicar la variabilidad en la variable dependiente.

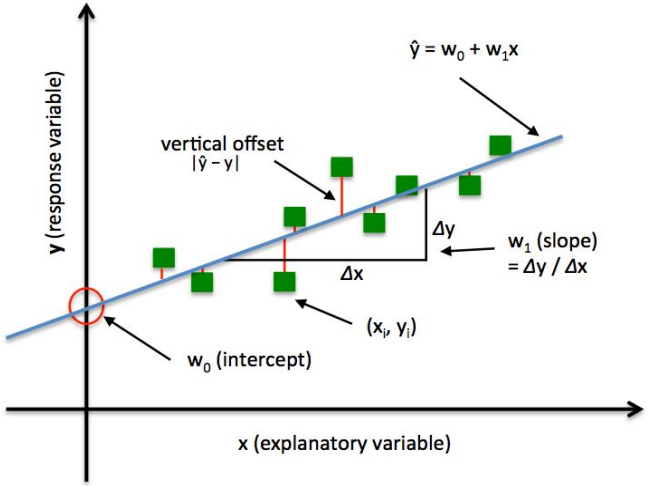
Modelo de regresión lineal simple vs. modelo de regresión lineal múltiple

La regresión lineal simple es aquella en la que se utiliza una única variable independiente para explicar la variabilidad en la variable dependiente. El modelo de regresión lineal simple se puede representar matemáticamente como:

$Y = \beta 0 + \beta 1X + \varepsilon$

Donde:

- Y es la variable dependiente (también llamada variable de respuesta)
- X es la variable independiente (también llamada variable explicativa)
- β0 es la ordenada al origen (el valor de Y cuando X es igual a cero)
- β1 es la pendiente de la recta de regresión (representa el cambio en Y por unidad de cambio en X)
- ε es el error aleatorio (representa la variabilidad no explicada por la variable independiente)



Modelo de regresión lineal simple vs. modelo de regresión lineal múltiple

En la regresión lineal múltiple, se utilizan dos o más variables independientes para explicar la variabilidad en la variable dependiente. El modelo de regresión lineal múltiple se puede representar matemáticamente como:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp + \varepsilon$$

Donde:

- Y es la variable dependiente
- X1, X2, ..., Xp son las variables independientes
- β0 es la ordenada al origen
- β1, β2, ..., βp son las pendientes de las rectas de regresión de cada variable independiente
- ε es el error aleatorio

La principal diferencia entre el modelo de regresión lineal simple y el modelo de regresión lineal múltiple es que en este último se incluyen varias variables independientes para explicar la variabilidad en la variable dependiente. Por lo tanto, en el modelo de regresión lineal múltiple, cada variable independiente se considera en relación a las demás variables independientes incluidas en el modelo, lo que permite capturar las interacciones entre las variables y su efecto conjunto en la variable dependiente.

Breve acercamiento a modelos de regresión lineal múltiple

Supuestos de la regresión lineal múltiple

La regresión lineal múltiple es una técnica estadística que se utiliza para analizar la relación entre una variable dependiente y varias variables independientes. Sin embargo, para que los resultados de la regresión lineal múltiple sean confiables, es necesario que se cumplan ciertos supuestos. A continuación, se detallan algunos de los supuestos más importantes de la regresión lineal múltiple:

1. **Linealidad**: el modelo de regresión lineal múltiple se basa en la suposición de que la relación entre las variables independientes y la variable dependiente es lineal. Esto significa que el efecto de una variable independiente sobre la variable dependiente es constante, independientemente del valor de las demás variables independientes.
2. **Independencia**: los errores en la regresión deben ser independientes entre sí, lo que significa que los errores de una observación no deben estar correlacionados con los errores de otras observaciones. Si hay dependencia entre los errores, esto puede sesgar los coeficientes de regresión y los intervalos de confianza.
3. **Homocedasticidad**: los errores en la regresión deben tener una varianza constante en todos los niveles de las variables independientes. Si la varianza de los errores no es constante, se dice que hay heterocedasticidad. La heterocedasticidad puede hacer que los intervalos de confianza y los valores p sean incorrectos, lo que puede conducir a conclusiones erróneas.

Supuestos de la regresión lineal múltiple

1. **Normalidad**: los errores en la regresión deben seguir una distribución normal. Si los errores no siguen una distribución normal, esto puede sesgar los coeficientes de regresión y los intervalos de confianza. Además, la distribución normal de los errores es un requisito para realizar pruebas estadísticas significativas.
2. **Multicolinealidad**: cuando dos o más variables independientes están altamente correlacionadas entre sí, se dice que existe multicolinealidad. La multicolinealidad puede dificultar la interpretación de los coeficientes de regresión y hacer que los intervalos de confianza sean amplios. Además, puede conducir a la selección incorrecta de variables para el modelo.

Es importante tener en cuenta que la violación de uno o más supuestos de la regresión lineal múltiple puede afectar la validez de los resultados obtenidos. Por lo tanto, es importante realizar pruebas de diagnóstico para verificar si se cumplen los supuestos antes de interpretar los resultados de la regresión. Si los supuestos no se cumplen, pueden ser necesarios ajustes al modelo para mejorar su precisión y confiabilidad.

Estimación de los coeficientes de regresión

Para estimar los coeficientes de regresión, se utiliza el método de **mínimos cuadrados**, que consiste en encontrar los valores de los coeficientes que minimizan la suma de los cuadrados de las diferencias entre los valores observados de la variable dependiente y los valores predichos por el modelo. En otras palabras, se busca el ajuste de línea que mejor se ajuste a los datos.

Estimación de los coeficientes de regresión

La ecuación general del modelo de regresión lineal múltiple es:

$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta kXk + \varepsilon$

Donde:

- Y es la variable dependiente.
- β0 es el término constante o intercepto.
- β1, β2, ..., βk son los coeficientes de regresión que indican el efecto de las variables independientes X1, X2, ..., Xk sobre la variable dependiente.
- X1, X2, ..., Xk son las variables independientes.
- ε es el término de error.

Los coeficientes de regresión se estiman mediante el uso de datos de muestra y se denotan como b0, b1, b2, ..., bk. El coeficiente de regresión b0 se estima como la media de los valores observados de la variable dependiente cuando todas las variables independientes son iguales a cero. Los demás coeficientes de regresión se estiman como las pendientes de las líneas de regresión de cada variable independiente.

Estimación de los coeficientes de regresión

Los valores estimados de los coeficientes de regresión se utilizan para construir la ecuación de regresión múltiple y para **predecir los valores de la variable dependiente a partir de los valores de las variables independientes**.

Es importante tener en cuenta que los coeficientes de regresión son sensibles a la presencia de valores atípicos y a la violación de los supuestos de la regresión lineal múltiple. Por lo tanto, es importante realizar pruebas de diagnóstico para verificar si se cumplen los supuestos antes de interpretar los resultados de la regresión. Si los supuestos no se cumplen, pueden ser necesarios ajustes al modelo para mejorar su precisión y confiabilidad.

Interpretación de los coeficientes de regresión

La interpretación de los coeficientes de regresión es esencial para comprender cómo las variables independientes influyen en la variable dependiente en un modelo de regresión lineal múltiple. Los coeficientes de regresión muestran cuánto cambia la variable dependiente por unidad de cambio en cada variable independiente, manteniendo todas las demás variables independientes constantes.

Interpretación de los coeficientes de regresión

Un coeficiente de regresión positivo indica que un aumento en el valor de la variable independiente está asociado con un aumento en el valor de la variable dependiente, mientras que un coeficiente negativo indica que un aumento en el valor de la variable independiente está asociado con una disminución en el valor de la variable dependiente.

La interpretación de los coeficientes de regresión también puede proporcionar información sobre la magnitud de la influencia de cada variable independiente en la variable dependiente. Es importante tener en cuenta que el valor de cada coeficiente de regresión depende de la escala de medición de la variable independiente correspondiente. Por ejemplo, si una variable independiente se mide en unidades muy pequeñas, como milímetros, entonces el valor del coeficiente de regresión asociado será mucho más pequeño que si la misma variable se mide en unidades más grandes, como metros.

Evaluación de la calidad del ajuste del modelo

La evaluación de la calidad del ajuste del modelo de regresión múltiple es un paso importante en el análisis de regresión, ya que permite determinar la adecuación del modelo para representar los datos y hacer predicciones precisas. Existen diferentes medidas que se utilizan para evaluar la calidad del ajuste del modelo, las cuales se describen a continuación:

1. **R-cuadrado** (R^2): El coeficiente de determinación, también conocido como R-cuadrado, es una medida de la proporción de la varianza total de la variable dependiente que es explicada por el modelo de regresión múltiple. R^2 varía entre 0 y 1, y cuanto más cercano a 1 sea su valor, mayor será la capacidad del modelo para explicar la variabilidad en la variable dependiente.
2. **Error estándar de la estimación (SEE)**: El error estándar de la estimación es una medida de la variabilidad de los residuos, que es la diferencia entre los valores observados de la variable dependiente y los valores predichos por el modelo de regresión múltiple. Un valor bajo de SEE indica que el modelo puede predecir con mayor precisión los valores de la variable dependiente.

Evaluación de la calidad del ajuste del modelo

1. **Análisis de varianza (ANOVA)**: El análisis de varianza es una técnica estadística que se utiliza para determinar si el modelo de regresión múltiple es significativamente mejor que el modelo nulo, que no incluye ninguna variable independiente. El ANOVA calcula la suma de cuadrados totales (SCT), que es la suma de las diferencias entre los valores observados de la variable dependiente y su media, la suma de cuadrados del modelo (SCM), que es la suma de las diferencias entre los valores predichos de la variable dependiente y su media, y la suma de cuadrados del error (SCE), que es la suma de las diferencias entre los valores observados y predichos de la variable dependiente. Si la relación entre las variables independientes y la variable dependiente es significativa, entonces el SCM es significativamente mayor que el SCE.
2. **Pruebas de hipótesis de los coeficientes de regresión**: Las pruebas de hipótesis se utilizan para determinar si cada coeficiente de regresión es significativamente diferente de cero. Un coeficiente de regresión significativo indica que la variable independiente correspondiente tiene una influencia significativa sobre la variable dependiente.

Evaluación de la calidad del ajuste del modelo

1. **Gráficos de residuos**: Los gráficos de residuos se utilizan para evaluar la distribución de los residuos y verificar si se cumplen los supuestos de la regresión lineal múltiple, como la normalidad y la homogeneidad de la varianza. Un modelo de regresión múltiple adecuado debe tener residuos que estén distribuidos normalmente y tengan una varianza constante en todas las combinaciones de los valores de las variables independientes.

Selección de variables independientes

La selección de variables independientes es un paso crítico en la regresión lineal múltiple, ya que permite identificar las variables que tienen una mayor influencia en la variable dependiente y mejorar la precisión del modelo. A continuación, se describen algunos métodos comunes de selección de variables independientes:

1. **Selección por criterios estadísticos**: este método utiliza criterios estadísticos, como la significancia de los coeficientes de regresión o la bondad de ajuste del modelo, para seleccionar las variables independientes más importantes. El proceso comienza con la inclusión de todas las variables independientes en el modelo y, a continuación, se eliminan las variables que no son estadísticamente significativas.
2. **Selección hacia adelante**: este método comienza con un modelo nulo que no incluye ninguna variable independiente y, a continuación, agrega secuencialmente las variables independientes más importantes en términos de su capacidad para mejorar la calidad del ajuste del modelo.

Selección de variables independientes

1. **Selección hacia atrás**: este método comienza con un modelo que incluye todas las variables independientes y, a continuación, elimina secuencialmente las variables que son menos importantes en términos de su capacidad para mejorar la calidad del ajuste del modelo.
2. **Selección por métodos de regularización**: los métodos de regularización, como la regresión Ridge y la regresión Lasso, se utilizan para reducir el sobreajuste del modelo al penalizar los coeficientes de regresión grandes. Estos métodos se basan en la minimización de una función objetivo que incluye un término de regularización y pueden eliminar automáticamente las variables que no contribuyen significativamente a la predicción.

Es importante tener en cuenta que la selección de variables independientes debe realizarse con precaución, ya que la inclusión o exclusión de variables puede tener un impacto significativo en la calidad del ajuste del modelo y en las conclusiones obtenidas a partir del análisis. Es recomendable realizar una validación cruzada y otras técnicas de validación del modelo para evaluar la capacidad del modelo para generalizar a nuevos datos. Además, es importante considerar el contexto y la teoría detrás de las variables independientes, así como la relación entre ellas y la variable dependiente, antes de realizar la selección.

Regresión lineal múltiple con variables categóricas

En la regresión lineal múltiple, las variables categóricas o variables *dummy* se utilizan para incluir factores cualitativos en el modelo. Estas variables se crean a partir de variables categóricas que pueden tomar valores discretos y se utilizan para representar una característica específica de la categoría.

Por ejemplo, en un modelo que intenta predecir el precio de la vivienda, una variable categórica podría ser el tipo de casa (casa unifamiliar, apartamento, dúplex, etc.). Para incluir esta variable en el modelo, se crean variables dummy correspondientes a cada tipo de casa, donde la variable toma el valor 1 si la observación pertenece a esa categoría y 0 en caso contrario. Por lo tanto, si hay 4 tipos de casas, se crean 4 variables dummy.

Regresión lineal múltiple con variables categóricas

Una vez que se han creado las variables dummy, se incluyen en el modelo de la misma manera que las variables continuas. Cada variable dummy se incluye en el modelo como una variable independiente adicional, y su coeficiente de regresión indica el cambio en la variable dependiente para cada unidad adicional de la variable dummy. El coeficiente de la variable dummy de referencia (la variable dummy que se excluye del modelo) se interpreta como el valor de la variable dependiente para esa categoría.

Es importante tener en cuenta que la inclusión de variables dummy en el modelo puede aumentar el número de variables independientes y, por lo tanto, aumentar la complejidad del modelo. Esto puede llevar a un sobreajuste del modelo y reducir su capacidad para generalizar a nuevos datos. Además, es importante tener cuidado al interpretar los coeficientes de regresión de las variables dummy, ya que solo se comparan las categorías dentro de la variable categórica y no entre ellas.

En resumen, las variables categóricas se pueden incluir en la regresión lineal múltiple mediante la creación de variables dummy. Estas variables se utilizan para incluir factores cualitativos en el modelo y su inclusión se realiza de manera similar a las variables continuas. Es importante considerar la complejidad del modelo y tener cuidado al interpretar los coeficientes de regresión de las variables dummy.

Ejemplos de aplicaciones de la regresión lineal múltiple

La regresión lineal múltiple y la regresión polinomial son técnicas estadísticas muy útiles que se aplican en una amplia variedad de campos para analizar relaciones entre variables y predecir valores futuros. A continuación, se presentan algunos ejemplos de su aplicación en diferentes campos:

Economía:

- Predicción de la demanda de un producto en función del precio, ingreso de los consumidores, la publicidad y otros factores.
- Análisis de los factores que influyen en el crecimiento económico, como la inversión, el gasto público y la tasa de interés.
- Análisis de la relación entre la inflación y el desempleo, utilizando la regresión polinomial para modelar una posible relación curvilínea.

Psicología:

- Predicción del rendimiento académico en función de variables como la motivación, el tiempo de estudio y el estrés.
- Análisis de la relación entre la personalidad y el bienestar emocional, utilizando la regresión polinomial para modelar una posible relación no lineal.
- Predicción de la satisfacción laboral en función de factores como el salario, la autonomía en el trabajo y el equilibrio entre vida laboral y personal.

Ejemplos de aplicaciones de la regresión lineal múltiple

Ingeniería:

- Predicción de la resistencia de los materiales en función de variables como la composición química, la temperatura y la humedad.
- Análisis de la relación entre la velocidad del viento y la potencia generada por una turbina eólica.
- Predicción del consumo de combustible de un vehículo en función de variables como la velocidad, el tipo de terreno y la carga.

Ciencias sociales:

- Análisis de la relación entre la edad y el uso de la tecnología en la población, utilizando la regresión polinomial para modelar una posible relación no lineal.
- Predicción del voto en una elección en función de variables como la afiliación política, la edad y la educación.
- Análisis de la relación entre el tiempo que se dedica a las redes sociales y el bienestar emocional, utilizando la regresión polinomial para modelar una posible relación no lineal.

En resumen, la regresión lineal múltiple y la regresión polinomial son técnicas estadísticas versátiles que se aplican en una amplia variedad de campos para analizar relaciones entre variables y predecir valores futuros. Su aplicación en la economía, la psicología, la ingeniería y las ciencias sociales demuestra su relevancia y utilidad en la investigación y toma de decisiones en diferentes áreas.

Interpretación y análisis de resultados de casos prácticos.

En este apartado se presentará una guía para interpretar y analizar los resultados obtenidos mediante la regresión lineal múltiple o regresión polinomial en casos prácticos. Se utilizará un ejemplo hipotético para ilustrar el proceso.

Supongamos que se desea analizar la relación entre la cantidad de horas de estudio y el promedio de notas de un grupo de estudiantes. Se recolectaron datos de 30 estudiantes y se realizó una regresión lineal múltiple con una variable independiente (horas de estudio) y una variable dependiente (promedio de notas). A continuación, se presentan los resultados obtenidos:

- Coeficientes de regresión: la ecuación de la recta de regresión es y = 0.8x + 2.5, donde y es el promedio de notas y x es la cantidad de horas de estudio. Esto significa que, en promedio, por cada hora de estudio adicional, se espera un incremento de 0.8 en el promedio de notas, manteniendo constante el efecto de otras variables.
- Coeficiente de determinación (R^2): el valor obtenido es 0.65, lo que significa que el 65% de la variabilidad en el promedio de notas se explica por la cantidad de horas de estudio. El 35% restante puede ser explicado por otras variables que no se incluyeron en el modelo.

Interpretación y análisis de resultados de casos prácticos.

- Significancia estadística: se realizó un análisis de significancia mediante el test de hipótesis para cada coeficiente de regresión. En este caso, el coeficiente para la variable de horas de estudio tiene un valor de t de 6.28 y un valor p de 0.0001, lo que indica que es significativamente diferente de cero. Esto significa que la relación entre las horas de estudio y el promedio de notas es estadísticamente significativa.
- Análisis de residuos: se realizaron gráficos de los residuos para evaluar si se cumplen los supuestos de la regresión lineal. En este caso, el gráfico de residuos versus ajustes muestra una distribución aleatoria de los residuos, lo que indica que no hay patrones evidentes en los errores de predicción. El gráfico de residuos versus variables independientes no muestra patrones claros, lo que sugiere que no hay violaciones graves del supuesto de linealidad.

En conclusión, los resultados obtenidos en este caso indican que la cantidad de horas de estudio es un factor significativo para predecir el promedio de notas de los estudiantes, explicando el 65% de la variabilidad en el promedio de notas. Además, los gráficos de residuos indican que se cumplen los supuestos de la regresión lineal. Estos resultados pueden ser útiles para tomar decisiones en el contexto educativo, como la asignación de tareas o el diseño de programas de estudio. Sin embargo, es importante tener en cuenta que existen otras variables que pueden influir en el promedio de notas y que no se incluyeron en el modelo, por lo que se recomienda realizar análisis adicionales para evaluar la validez de las conclusiones.

Limitaciones y desventajas de estas técnicas

Aunque la regresión lineal múltiple y la regresión polinomial son técnicas ampliamente utilizadas y tienen muchas ventajas, también presentan algunas limitaciones y desventajas que se deben tener en cuenta al aplicarlas:

1. Supuestos: La regresión lineal múltiple y la regresión polinomial se basan en varios supuestos, como la linealidad de la relación entre las variables, la normalidad de los errores, la homocedasticidad de los errores y la independencia de los errores. Si estos supuestos no se cumplen, los resultados pueden ser inexactos o engañosos.
2. Sobreajuste: Si se incluyen demasiadas variables independientes en el modelo, el modelo puede sobreajustarse a los datos de entrenamiento y no generalizar bien a nuevos datos. Esto puede resultar en un modelo que se ajusta demasiado a los datos de entrenamiento y no se ajusta bien a los datos de prueba.
3. Multicolinealidad: Si las variables independientes están altamente correlacionadas entre sí, puede haber problemas de multicolinealidad en el modelo. Esto puede hacer que sea difícil determinar la importancia relativa de cada variable independiente en el modelo.
4. Datos atípicos: La presencia de valores atípicos o datos extremos puede tener un gran impacto en los resultados del modelo y distorsionar los resultados.

Limitaciones y desventajas de estas técnicas

1. Requerimientos de datos: La regresión lineal múltiple y la regresión polinomial requieren grandes conjuntos de datos para proporcionar resultados precisos y confiables. Si el conjunto de datos es pequeño, puede haber problemas con la precisión y la confiabilidad de los resultados.
2. Asunciones lineales: La regresión lineal múltiple y la regresión polinomial se basan en la asunción de que la relación entre las variables es lineal. Si la relación no es lineal, los resultados pueden ser engañosos.
3. Inferencia limitada: Aunque la regresión lineal múltiple y la regresión polinomial se pueden utilizar para hacer predicciones, su capacidad para hacer inferencias causales es limitada. A menudo, hay muchos factores que influyen en un resultado, y puede ser difícil aislar el efecto de una sola variable independiente.

En resumen, la regresión lineal múltiple y la regresión polinomial son técnicas poderosas y versátiles, pero deben ser utilizadas con cuidado y con una comprensión completa de sus limitaciones y desventajas.

Actividad 6. Cuestionario sobre Regresión lineal simple y múltiple

Instrucciones

Después de leer detenidamente el contenido de la presentación así como los recursos recomendados por el docente, acceder a la actividad `Cuestionario sobre Regresión lineal simple y múltiple`en la plataforma y contestar de forma individual las preguntas listadas.