**Transformaciones para Linearizar Relaciones No Lineales**

Las transformaciones para linearizar relaciones no lineales y resolver problemas relacionados con la heterocedasticidad cambio no uniforme de la variabilidad de los errores. Entre estas transformaciones, la transformación de Box-Cox se presenta como un método útil.

La transformación Box-Cox se define para un parámetro y es expresada como:

Esta fórmula se utiliza para transformar datos con el objetivo de aproximar una relación lineal, especialmente cuando los residuos presentan heterocedasticidad.

* Cuando se aproxima a 0, la transformación equivale a aplicar el logaritmo a la variable .
* Si , la variable transformada crece más lentamente que la original.
* Si la variable transformada crece más rápidamente que la original.

La elección del valor de tiene como objetivo ajustar los datos a una distribución más adecuada. En este contexto:

* Si los residuos presentan heterocedasticidad o no normalidad, se transforma la variable .
* Si la variable explicativa muestra problemas similares, se transforma .

En general, las transformaciones son una herramienta para ajustar relaciones no lineales y aproximarlas a una forma lineal, lo que mejora la calidad del modelo de regresión. Estas se aplican modificando las variables de manera que reflejen mejor el patrón observado en los datos. Por ejemplo, si una variable crece más lentamente de lo esperado en su escala original, una transformación adecuada puede comprimir el rango de valores y lograr que la relación sea más lineal. En cambio, si una variable crece de forma muy rápida, se pueden usar transformaciones que expandan los incrementos más pequeños en comparación con los mayores, ajustando así la relación.

Además, es importante notar que las transformaciones permiten manejar situaciones donde los incrementos constantes en las variables originales generan incrementos variables en la escala transformada. Esto facilita la linealización al elegir parámetros específicos para la transformación que compensen dichas variaciones. La elección de la transformación adecuada depende, entonces, de cómo se comporta la relación en los datos: si la curva de los datos tiene una inclinación hacia arriba o hacia abajo, se seleccionará una transformación que ajuste el crecimiento o la compresión según sea necesario.

En última instancia, el objetivo es garantizar que la relación entre las variables sea lo más lineal posible, ya sea transformando una o ambas variables, dependiendo del problema específico. Esto permite que el modelo de regresión sea más preciso y fácil de interpretar, optimizando su capacidad para explicar o predecir el fenómeno en estudio.

**Estimación Máximo-Verosímil de**

La estimación del parámetro para realizar transformaciones que optimicen la linealidad y homocedasticidad de los datos se puede realizar mediante el método de máxima verosimilitud. Esto implica definir una transformación específica de la respuesta que reduzca la varianza no explicada del modelo. La transformación considerada utiliza una fórmula general donde se introduce el parámetro para modificar los datos y ajustarlos a un comportamiento más lineal. Dependiendo del valor de , se pueden lograr diferentes efectos en los datos transformados. Valores mayores a uno tienden a acelerar el crecimiento de la variable transformada respecto a la original, mientras que valores menores a uno producen un crecimiento más lento.

La función de verosimilitud ) se define como:

donde representa la varianza no explicada después de aplicar la transformación. El valor de que minimiza esta varianza es considerado óptimo, ya que asegura un mejor ajuste del modelo.

Para determinar dicho valor óptimo, se realizan iteraciones probando diferentes valores de . Cada iteración implica ajustar un modelo de regresión, calcular la varianza residual asociada, y evaluar la función de verosimilitud. El intervalo de confianza de puede estimarse utilizando límites basados en la distribución , lo que proporciona una idea de la robustez del valor óptimo obtenido.

Este enfoque no solo mejora la linealidad y homocedasticidad de los datos, sino que también resulta en un modelo más robusto frente a problemas de heterocedasticidad o no linealidad. Además, la metodología puede extenderse para manejar transformaciones simultáneas de múltiples variables, lo que la convierte en una herramienta versátil para el análisis de datos.

**Transformaciones para conseguir homocedasticidad**

El objetivo de realizar transformaciones para lograr homocedasticidad en un modelo de regresión es ajustar la relación entre la variabilidad de los residuos y la respuesta media, garantizando un modelo más estable y confiable. Este ajuste se basa en identificar cómo la varianza de los residuos se relaciona con el valor esperado de la respuesta.

Cuando la varianza crece proporcionalmente al cuadrado de la respuesta esperada, se recomienda aplicar una transformación que modifique esta relación, de manera que la varianza sea constante. Una opción común para lograr esto es utilizar transformaciones basadas en el parámetro , donde describe la relación entre la varianza y el valor esperado.

La homocedasticidad, que se refiere a la condición en que la varianza de los errores de un modelo de regresión es constante para todos los valores de la variable independiente, se puede lograr mediante transformaciones. Se establece que:

Si , se obtiene la transformación logarítmica. Para estimar la relación entre la variabilidad y la media esperada, se realiza el siguiente procedimiento:

1. Ordenar los valores de según .
2. Agrupar observaciones contiguas (4-5 por grupo).
3. Calcular la media y el rango por grupo, considerando el rango como medida de variabilidad ya que con tamaños muestrales pequeños es tan eficaz como la varianza y es algo más robusto.
4. Graficar la media contra el rango para cada grupo.

Si el gráfico del rango contra la media sigue una función , entonces se usa para transformar .

6.4.3 Consecuencias de las transformaciones

Transformar observaciones afecta la interpretación de los parámetros del modelo. Por ejemplo, si el modelo transformado es:

Las variables transformadas son:

Donde:

Esto introduce un sesgo proporcional a la varianza residual . También es fundamental considerar la nueva interpretación de los coeficientes. Por ejemplo:

representa la elasticidad, que mide el cambio porcentual en dado un cambio porcentual en .

6.5 Regresión no paramétrica

La regresión no paramétrica permite estimar directamente la forma de la función condicional sin asumir su estructura. La estimación de se realiza ponderando las observaciones cercanas al punto con pesos decrecientes:

El método del núcleo asigna pesos basados en una función de densidad, como la normal estándar:

El parámetro (ancho de ventana) controla la suavidad de la estimación. Valores pequeños de generan estimaciones más detalladas, pero menos suaves.

6.6 Predicción

En los modelos de regresión, se distinguen dos tipos de predicción:

1. **Estimación de medias condicionales :**

Su varianza es:

Donde es la medida del efecto palanca de .

1. **Predicción de nuevos valores:**

Se calcula sustituyendo en la ecuación de regresión, añadiendo la varianza residual.

**Estimación de la varianza y la media condicionada**

En el contexto de la regresión lineal, cuando se desea estimar la media condicionada de dado un valor específico de , utilizamos la ecuación:

La **varianza de la estimación de la media condicionada** se expresa como:

Donde:

* \(\sigma^2\) es la varianza residual estimada del modelo.
* \(h\_h\) es el elemento \(h\)-ésimo de la matriz de proyección \(H\), conocida como matriz de efectos palanca (\textit{leverage}):

En regresión lineal simple, se calcula como:

**Interpretación de :**

* **Efecto palanca (leverage):** mide la influencia de en su propio valor ajustado . Valores altos de indican mayor influencia.
* **Varianza de la estimación:** A medida que aumenta, la varianza de también lo hace, disminuyendo la precisión de la estimación.

El concepto del **número equivalente de observaciones ()** se define como:

La varianza de la estimación se puede expresar como:

**Implicaciones prácticas:**

* Para puntos cercanos a es pequeño, es grande, y la varianza de es pequeña.
* Para puntos alejados de : es grande, es pequeño, y la varianza de es mayor.

**Intervalos de confianza para medias condicionadas**

Para construir intervalos de confianza para la media estimada , utilizamos la distribución t de Student con grados de libertad:

Donde:

* es la desviación estándar residual.
* .

**Interpretación:**

* Los intervalos son más estrechos cerca de y se ensanchan en los extremos.

**Predicción de nuevas observaciones**

Cuando se predice un valor futuro de para un , usamos la misma estimación de la media condicionada, pero consideramos la **varianza adicional** debido a la variabilidad de los datos:

El intervalo de predicción es:

**Bandas de confianza y predicción**

Las **bandas de confianza** representan gráficamente los intervalos para todos los valores de :

* Banda de confianza para la media: Representa la precisión de a lo largo de .
* Banda de predicción: Es más amplia, ya que incorpora la variabilidad adicional.

**Riesgos de extrapolación**

**Extrapolación:** Usar el modelo para predecir valores de fuera del rango observado de .

**Riesgos:**

* La relación entre e podría no ser lineal fuera del rango observado.
* Al extrapolar, , lo que aumenta la varianza y disminuye la precisión.
* Los intervalos de predicción podrían subestimar la verdadera incertidumbre.

**Conclusiones:**

* Usar el modelo dentro del rango observado de .
* Validar extrapolaciones con datos adicionales.

**Metodología para construir modelos de regresión (Sección 6.7)**

La sección describe un enfoque estructurado para construir modelos de regresión entre dos variables:

Pasos iniciales:

* Realizar un diagrama de dispersión para verificar la relación entre las variables.
* Comprobar si un modelo de regresión lineal simple es apropiado.
* **Estimación de parámetros**:

1. Estimar los parámetros del modelo y verificar si existe una relación estadísticamente significativa.

* **Diagnóstico del modelo**:

1. Realizar pruebas diagnósticas de linealidad, normalidad, homocedasticidad e independencia.
2. Si el modelo cumple con estas condiciones, puede usarse para predicciones.

* **Transformaciones o extensiones:**

1. Si los supuestos fallan, considerar transformar las variables o incluir variables explicativas adicionales, lo que lleva a un modelo de regresión múltiple.

**Cuadro 6.1: Diagrama de flujo para construir modelos de regresión**

El diagrama de flujo ilustra el proceso de toma de decisiones para la construcción de modelos, comenzando desde la exploración de datos hasta la validación de los supuestos, incluyendo posibles transformaciones o ampliaciones del modelo.

Conceptos clave y fórmulas en regresión simple (Cuadro 6.2)

**Modelo:**

**Estimación**:

**Propiedades de los estimadores**:

Prueba de hipótesis:

* Prueba de simplificación para :

**Diagnósticos:**

Análisis de residuos para probar:

* Normalidad,
* Independencia,
* Linealidad,
* Homocedasticidad.

**Otras medidas**:

\item \textbf{**Efecto de apalancamiento** :

**Intervalo de predicción**:

**Número equivalente de observaciones**:

**Coeficiente de correlación**: