Transformaciones para Linearizar Relaciones No Lineales

La sección se centra en 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.

Transformación de Box-Cox

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

Esta fórmula permite transformar datos para aproximar una relación lineal, especialmente cuando los residuos muestran comportamientos heterocedásticos. Cuando es cercano a 0, la transformación es equivalente a tomar el logaritmo de .

Interpretación de la Tabla 6.4:

La tabla muestra cómo, al calcular para diferentes valores de , los resultados convergen al logaritmo de 10 a medida que tiende a 0. Esto confirma que la transformación Box-Cox logra mantener la continuidad de la función.

Efecto de la Transformación y Decisión del Parámetro

La Figura 6.17 ilustra el impacto de en la transformación. Valores de producen que la variable transformada crezca más lentamente que la original, mientras que produce un crecimiento más rápido. Esto se traduce en la elección de para ajustar los datos a una distribución más adecuada:

* Si los residuos muestran heterocedasticidad o no normalidad, se transforma .
* Si la variable explicativa presenta problemas similares}, se transforma .

**Linealización de Relaciones No Lineales**

Las Figuras 6.18 a 6.21 ilustran ejemplos prácticos de relaciones no lineales que pueden ser linealizadas mediante las transformaciones descritas. Las gráficas muestran:

* Cómo los valores de y cambian dependiendo de .
* Los residuos asociados a estas transformaciones para evaluar si la relación se aproxima a la linealidad.

Clasificación de Relaciones Linealizables

La Figura 6.22 clasifica varias formas de relaciones no lineales que pueden transformarse en relaciones lineales con distintas técnicas. Por ejemplo:

* Relaciones exponenciales del tipo pueden transformarse tomando el logaritmo natural.
* Relaciones inversas del tipo se linealizan mediante transformaciones recíprocas.

Estimación Máximo-Verosímil de

Se introduce la función de verosimilitud para encontrar el valor óptimo de que maximice la linealidad y homocedasticidad:

Donde es la varianza no explicada después de la transformación. La estimación de que minimice la varianza residual será la solución óptima.

Figura 6.23: La figura ilustra cómo determinar mediante curvas de verosimilitud, mostrando los valores límites y máximos correspondientes.

Este conjunto de herramientas y técnicas proporciona métodos robustos para transformar datos y adaptarlos a modelos de regresión lineal, resolviendo problemas de no linealidad y heterocedasticidad.

Transformaciones para conseguir homocedasticidad

La transformación adecuada para conseguir homocedasticidad en la respuesta se elige en el modelo de regresión de una manera análoga a la estudiada en el capítulo 2. Es decir:

entonces

En particular, si la varianza crece con el cuadrado de la respuesta esperada, , se obtiene la transformación logarítmica, . Para estimar la relación entre la variabilidad y la respuesta media cuando se dispone únicamente de un valor de para cada , se procede como sigue:

* Ordenar los valores de en función de valores crecientes de .
* Formar grupos de 4 o 5 observaciones contiguas.
* Calcular en cada grupo la media y el rango. Se toma el rango como medida de variabilidad, ya que con tamaños muestrales pequeños es tan eficaz como la varianza y es algo más robusto.
* Hacer un gráfico entre la media y el rango para cada grupo.

Sea la media del grupo y el rango del grupo. Si el gráfico es de la forma

entonces debemos transformar la respuesta con , donde .

Resumen de Regresión Lineal y Transformaciones

6.4.2 Transformaciones para conseguir homocedasticidad

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.
4. Graficar la media contra el rango.

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**: