Interpolación de Datos y Transformaciones de Variables

Técnicas para Ciencia de Datos en Economía

Análisis de Datos
2025-10-23

Contenidos

PARTE I: INTERPOLACIÓN DE DATOS

El Problema de los Datos Intercensales

Contexto: Los censos poblacionales se realizan cada 10 años

Problema: ¿Qué población asignamos a los años intermedios?

¿Por qué importa?

- Cálculo de tasas per cápita (PIB, gasto público, etc.)
- Proyecciones y planificación de políticas públicas
- Investigación demográfica y económica
- Análisis históricos consistentes

La interpolación NO es proyección - solo completa datos entre observaciones conocidas

Método 1: Interpolación Lineal

¿Qué hace?

- Conecta los puntos censales con líneas rectas
- Asume incremento absoluto constante cada año

¿Para qué sirve?

- Simple y transparente
- Útil cuando no hay supuestos sobre el proceso subyacente
- Buena primera aproximación

Limitaciones:

- **Poco realista demográficamente**: La población no crece linealmente
- Ignora que las tasas de crecimiento suelen cambiar
- Puede subestimar o sobreestimar sistemáticamente en períodos largos

Método 2: Interpolación Exponencial (Geométrica)

¿Qué hace?

- Asume tasa de crecimiento porcentual constante
- Interpola en escala logarítmica

¿Para qué sirve?

- Más realista para fenómenos demográficos y económicos
- Respeta el comportamiento de crecimiento compuesto
- Estándar en demografía oficial

Limitaciones:

- Asume tasa constante (poco realista en transición demográfica)
- No captura aceleraciones o desaceleraciones dentro del período
- Puede fallar si hay cambios estructurales

Método 3: Interpolación Logística

¿Qué hace?

- Crecimiento rápido al inicio, se desacelera hacia el final
- Modela la transición demográfica

¿Para qué sirve?

- Captura cambios en las tasas de crecimiento
- Útil en períodos de transición demográfica
- Más flexible que el modelo exponencial

Limitaciones:

- Requiere calibración del parámetro de curvatura (k)
- Puede no ajustarse bien a todos los períodos
- Más complejo de explicar y defender

Método 4: Spline Cúbica

¿Qué hace?

- Curva suave que pasa exactamente por todos los puntos censales
- Une segmentos cúbicos con continuidad en derivadas

¿Para qué sirve?

- Máxima suavidad visual
- Útil para visualizaciones
- Captura cambios graduales en tendencias

Limitaciones:

- **Puede crear oscilaciones artificiales** entre censos
- No tiene interpretación demográfica clara
- Sensible a outliers en los datos censales
- Puede producir valores irrealistas cerca de los extremos

Método 5: LOCF (Last Observation Carried Forward)

¿Qué hace?

- Mantiene el valor del último censo hasta el próximo
- "Escalones" en la serie temporal

¿Para qué sirve?

- Útil solo en contextos muy específicos (stocks, inventarios)
- Transparente: muestra claramente dónde hay datos reales

Limitaciones:

- NO RECOMENDADO para población
- Ignora completamente el crecimiento poblacional
- Genera saltos artificiales en análisis
- Sesga cualquier cálculo que dependa de población

Método 6: Interpolación con Tasa Variable

¿Qué hace?

- Primero interpola las tasas de crecimiento entre censos
- Luego aplica estas tasas interpoladas año a año

¿Para qué sirve?

- Incorpora información sobre cambios en tasas
- Más flexible que exponencial simple
- Útil cuando hay información adicional sobre tendencias

Limitaciones:

- Requiere calcular tasas de crecimiento (sensible a errores)
- Acumula errores período a período
- Más complejo computacionalmente
- Puede desviarse de valores censales si no se ajusta

Comparación de Métodos: ¿Cuál Elegir?

Para población (recomendado → menos recomendado):

1. **Exponencial/Geométrica** - estándar demográfico
2. **Logística** - si hay transición demográfica clara
3. **Lineal** - simplicidad, períodos cortos
4. **Tasa Variable** - si hay información auxiliar
5. **Spline** - solo para visualización
6. **LOCF** - nunca para población

Criterios de decisión:

- Interpretabilidad demográfica
- Longitud del período intercensal
- Disponibilidad de información auxiliar
- Propósito del análisis (investigación vs. visualización)

PARTE II: TRANSFORMACIONES DE VARIABLES

¿Por Qué Transformar Variables?

Problemas comunes en datos económicos:

- Variables en diferentes escalas (ventas en millones, empleados en unidades)
- Distribuciones asimétricas (salarios, ingresos)
- Outliers que distorsionan análisis
- Necesidad de comparar series heterogéneas

Las transformaciones permiten:

- Comparar variables comparables
- Mejorar la performance de modelos estadísticos
- Identificar y tratar observaciones atípicas
- Estabilizar varianzas

Estandarización (Z-score)

¿Qué hace?

- Transforma variables a media = 0 y desviación estándar = 1
- Fórmula: $Z = (X - \mu) / \sigma$

¿Para qué sirve?

- **Comparar variables en diferentes unidades** (ventas vs. empleados)
- Requisito para muchos algoritmos de machine learning
- Identificar observaciones inusuales ($|Z| > 3$)
- Análisis de componentes principales (PCA)

Limitaciones:

- **Sensible a outliers** (afectan media y desvío)
- Pierde la interpretación original de la variable
- Asume distribución aproximadamente normal
- No acota valores en un rango específico

Normalización (Min-Max Scaling)

¿Qué hace?

- Escala variables al rango [0, 1]
- Fórmula: X_norm = (X - X_min) / (X_max - X_min)

¿Para qué sirve?

- Redes neuronales y algoritmos sensibles a escala
- Cuando se necesita un rango acotado
- Comparaciones relativas dentro de una variable
- Visualizaciones con múltiples variables

Limitaciones:

- **Extremadamente sensible a outliers** (definen min y max)
- Valores futuros pueden caer fuera del rango [0,1]
- No mantiene relaciones de distancia entre observaciones
- Pierde interpretación económica

Detección de Outliers: Método IQR

¿Qué hace?

- Usa rango intercuartílico (Q3 - Q1)
- Outliers: valores fuera de [Q1 - 1.5×IQR, Q3 + 1.5×IQR]

¿Para qué sirve?

- Robusto a distribuciones asimétricas
- No asume normalidad
- Estándar en análisis exploratorio (boxplots)
- Detecta valores extremos relativos a la distribución

Limitaciones:

- Umbral de 1.5×IQR es arbitrario
- Puede marcar muchos valores legítimos como outliers
- No distingue entre errores y eventos reales extremos
- En muestras pequeñas puede ser poco confiable

Detección de Outliers: Método Z-score

¿Qué hace?

- Usa distancia en desviaciones estándar
- Outliers: |Z| > 3 (o 2.5 según contexto)

¿Para qué sirve?

- Variables aproximadamente normales
- Interpretación probabilística clara
- Complementa método IQR

Limitaciones:

- **Asume normalidad** - falla con distribuciones asimétricas
- Sensible a outliers (problema circular)
- Umbral de 3 es convención, no ley estadística
- Puede fallar en muestras pequeñas

Tratamiento de Outliers: Winsorización

¿Qué hace?

- Reemplaza valores extremos por percentiles específicos
- Típicamente: reemplaza valores < P1 y > P99

¿Para qué sirve?

- Reducir impacto de outliers sin eliminar observaciones
- Mantener el tamaño muestral
- Análisis más robusto de tendencias centrales
- Reducir varianza sin sesgar demasiado

Limitaciones:

- Elección de percentiles es subjetiva
- Puede distorsionar relaciones reales entre variables
- Reduce varianza artificialmente
- No es apropiado si los extremos son informativos

Tratamiento de Outliers: Dummies

¿Qué hace?

- Crea variable binaria: 1 = outlier, 0 = normal
- Mantiene el dato original en el modelo

¿Para qué sirve?

- Controlar por outliers en regresiones
- Mantener toda la información
- Testear si outliers cambian resultados
- Identificar patrones en outliers (¿siempre la misma sucursal?)

Limitaciones:

- Aumenta número de variables
- Requiere definir qué es outlier
- No reduce el impacto del valor extremo directamente
- Interpretación más compleja

Transformación Logarítmica

¿Qué hace?

- Aplica log() a la variable: Y_log = log(Y)
- Comprime valores altos, expande valores bajos

¿Para qué sirve?

- **Interpretar cambios como porcentajes** (elasticidades)
- Reducir asimetría en distribuciones
- Estabilizar varianza heterocedástica
- Modelar relaciones multiplicativas (Cobb-Douglas)
- Calcular tasas de crecimiento: $\Delta \log(Y) \approx$ % cambio

Limitaciones:

- Solo para valores positivos ($Y > 0$)
- Dificulta interpretación de niveles absolutos
- Puede exagerar importancia de valores pequeños

Tasas de Crecimiento: Log vs. Método Directo

Método Directo:

- % cambio = [(Y*t* - *Y*{t-1}) / Y_{t-1}] × 100

Método Logarítmico:

- % cambio ≈ [log(Y*t)* - *log(Y*{t-1})] × 100

¿Cuál usar?

- **Método log:** Mejor para acumulación en el tiempo, simetría
- **Método directo:** Más intuitivo, comunicación no técnica

Diferencia importante:

- Son aproximadamente iguales para cambios pequeños (<10%)
- Divergen significativamente para cambios grandes

Datos Faltantes: Interpolación en Series de Tiempo

Contexto: Datos faltantes en series temporales de ventas/economía

Métodos aplicables:

- **Interpolación lineal:** Simple, para gaps cortos
- **LOCF:** Si se asume persistencia (precios, tasas)
- **Spline:** Para series suaves
- **Modelos ARIMA:** Para series con estructura temporal

Consideraciones:

- ¿Missing at random (MAR) o sistemático?
- Longitud del gap (< 5% del período es manejable)
- Presencia de estacionalidad o tendencia
- Propósito del análisis posterior

Síntesis: Flujo de Trabajo Recomendado

1. Análisis Exploratorio

- Visualizar distribuciones
- Identificar outliers (IQR + Z-score)
- Evaluar datos faltantes

2. Tratamiento de Outliers

- Winsorización para análisis robusto
- Dummies para regresiones
- Nunca eliminar sin justificación

3. Transformaciones

- Log para variables económicas asimétricas
- Estandarización para machine learning
- Normalización para redes neuronales

Mensajes Clave

No existe "la mejor" técnica:

- Depende del contexto, propósito y datos
- Siempre documentar decisiones tomadas

Transparencia es fundamental:

- Reportar datos originales y transformados
- Análisis de sensibilidad (probar múltiples métodos)

Los outliers pueden ser informativos:

- No eliminar automáticamente
- Investigar causas antes de tratar

Las transformaciones tienen trade-offs:

- Ganan en un aspecto, pierden en otro
- La elección debe ser consciente y justificada