Universidad de El Salvador
Facultad de Ciencias Naturales y Matemática
Maestría en Estadística y Ciencia de Datos
Inferencia Estadística y Regresión

Control de lectura

Presentado por: Salvador Enrique Rodríguez Hernández (rh06006)
Fecha de entrega: 04 de diciembre de 2024

1. Transformaciones

1.1 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 λ y es expresada como:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda} \; si \; \lambda \neq 0, \\ \log(y) \; si \; \lambda = 0 \end{cases}$$

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

- Cuando λ se aproxima a 0, la transformación equivale a aplicar el logaritmo a la variable y.
- Si $\lambda > 1$, la variable transformada crece más lentamente que la original.
- Si $\lambda < 1$, la variable transformada crece más rápidamente que la original.

La elección del valor de λ 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 y.
- Si la variable explicativa muestra problemas similares, se transforma x.

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.

1.2 Estimación de la transformación de la respuesta por máxima verosimilitud

La estimación del parámetro λ 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 λ para modificar los datos y ajustarlos a un comportamiento más lineal. Dependiendo del valor de λ, 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 $L(\lambda)$ se define como:

$$L(\lambda) = -\frac{n}{2}\ln \text{VNE}\,(\lambda)$$

donde $\text{VNE}(\lambda)$ representa la varianza no explicada después de aplicar la transformación. El valor de λ 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 λ. Cada iteración implica ajustar un modelo de regresión, calcular la varianza residual asociada, y evaluar la función de verosimilitud. El intervalo de confianza de λ puede estimarse utilizando límites basados en la distribución χ^2, 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.

1.3 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 $\lambda = 1 - \alpha$, donde α 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 $\text{Var}(y|x) = k \cdot \text{E}(y|x)^{\alpha}$, entonces se transforma la respuesta con $\lambda = \alpha - 1$.

Si $\alpha = 1$, se obtiene la transformación logarítmica, $\lambda = 0$. Para estimar la relación entre la variabilidad y la media esperada, se realiza el siguiente procedimiento:

(1) Ordenar los valores de y según los valores crecientes de x.
(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.

Sea $\overline{y}_h$ la media del grupo h y R_h el rango del grupo. Si el gráfico $R_h = f(\overline{y}_h)$, es de la forma

$$R_h = k\overline{y}_h^{\alpha}$$

entonces debemos transformar la respuesta con $y(\lambda)$, donde $\lambda = 1 - \alpha$.

Ejemplo 3.6

Se va a analizar los datos del ejercicio 5.2 sobre la relación entre el número de trabajadores y el de supervisores. Los datos se encuentran en el fichero trabajadores.dat. El modelo estimado para los datos originales se presenta a continuación.

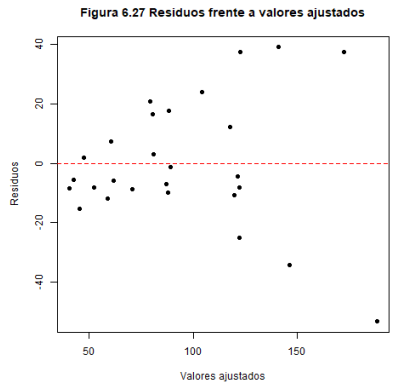
```
# Cargar el conjunto de datos
data <- read.table("trabajadores.dat", header = FALSE)
colnames(data) <- c("Supervisores", "Trabajadores")
# Modelo lineal inicial
modelo_inicial <- lm(Supervisores ~ Trabajadores, data = data)
summary(modelo_inicial)
```

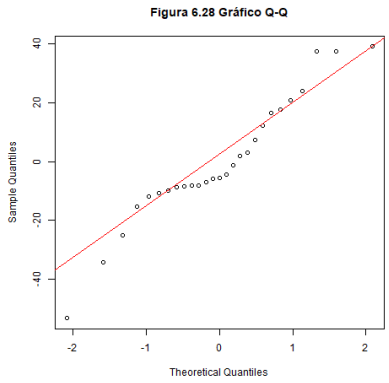
```
Call:
lm(formula = Supervisores ~ Trabajadores, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-53.294  -9.298  -5.579  14.394  39.119

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.44806    9.56201   1.511    0.143
Trabajadores  0.10536    0.01133   9.303 1.35e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.73 on 25 degrees of freedom
Multiple R-squared:  0.7759,    Adjusted R-squared:  0.7669
F-statistic: 86.54 on 1 and 25 DF,  p-value: 1.35e-09
```

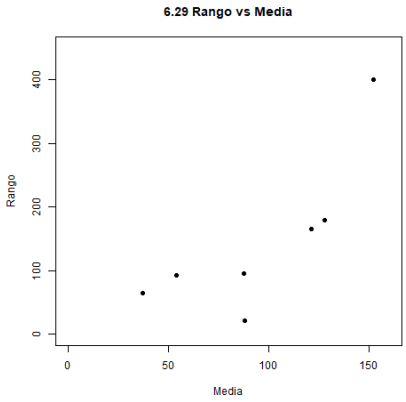
```
# Gráfico 6.27: Residuos frente a valores ajustados
png("figura_6_27.png")
plot(fitted(modelo_inicial), residuals(modelo_inicial),
     main = "Figura 6.27 Residuos frente a valores ajustados",
```

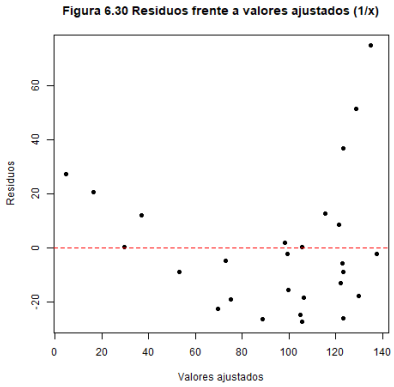
```
     xlab = "Valores ajustados", ylab = "Residuos", pch = 16, col =
"black")
abline(h = 0, col = "red", lty = 2, lwd = 1.5)
dev.off()
# Gráfico 6.28: Gráfico Q-Q
png("figura_6_28.png")
qqnorm(residuals(modelo_inicial), main = "Figura 6.28 Gráfico Q-Q")
qqline(residuals(modelo_inicial), col = "red", lwd = 1.5)
dev.off()
# Grupos y análisis de rangos y medias
grupos <- cut(data$Trabajadores, breaks = seq(0, max(data$Trabajadores),
length.out = 4), include.lowest = TRUE)
media_grupo <- tapply(data$Supervisores, grupos, mean)
rango_grupo <- tapply(data$Supervisores, grupos, function(x) max(x) -
min(x))
# Gráfico 6.29: Rangos frente a medias de grupo
png("figura_6_29.png")
plot(media_grupo, rango_grupo, main = "Figura 6.29 Rangos frente a medias
de grupo",
     xlab = "Media", ylab = "Rango", pch = 16, col = "black")
dev.off()
# Transformación inversa: y ~ 1/x
data$InvTrabajadores <- 1 / data$Trabajadores
modelo_inverso <- lm(Supervisores ~ InvTrabajadores, data = data)
summary(modelo_inverso)
# Gráfico 6.30: Residuos frente a valores ajustados (1/x)
png("figura_6_30.png")
plot(fitted(modelo_inverso), residuals(modelo_inverso),
     main = "Figura 6.30 Residuos frente a valores ajustados (1/x)",
     xlab = "Valores ajustados", ylab = "Residuos", pch = 16, col =
"black")
abline(h = 0, col = "red", lty = 2, lwd = 1.5)
dev.off()
```

Las figuras 6.27 y 6.28 muestran las gráficas de residuos frente a los valores previstos y la gráfica probabilística normal de los residuos. La gráfica 6.27 muestra claramente heterocedasticidad y no linealidad, y el 6.28, que la distribución de los residuos no es normal.










```
# Transformación log-log: log(y) ~ log(x)
data$LogSupervisores <- log(data$Supervisores)
data$LogTrabajadores <- log(data$Trabajadores)
modelo_log <- lm(LogSupervisores ~ LogTrabajadores, data = data)
summary(modelo_log)
```

Se puede intentar buscar una transformación que resuelva estos problemas simultáneamente. Haciendo grupos de cuatro observaciones y calculando la media y el rango del grupo se obtienen los resultados de la tabla siguiente:

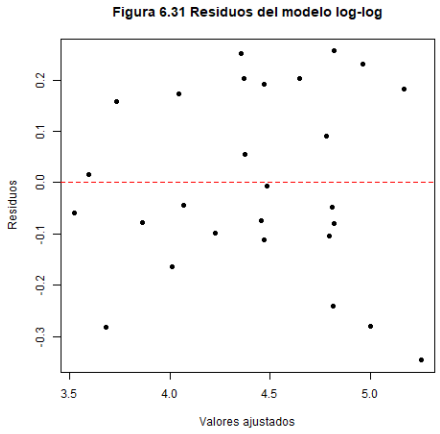
Grupos	1-4	5-8	9-12	13-16	17-20	21-24	25-27
Medias	37.00	53.75	87.50	88.00	121.00	127.75	152.30
Rangos	64	92	96	21	165	179	400

Que se representa en la figura 6.29. Se observa que la relación parece ligeramente no lineal, sobre todo por el último punto, y aproximadamente cuadrática. Suponer que es cuadrática y $\alpha=2$, llevaría a transformar con $\lambda=-1$, que es la transformación inversa. Se va a probar también la transformación logarítmica, que sería admitir linealidad en la relación, que es también consistente con los datos. Observemos que la decisión respecto a la transformación depende mucho de las coordenadas del último punto que se ha calculado con menos datos que los anteriores, por lo que conviene darle menos crédito. El gráfico de los

residuos de $= y(-1) = \frac{(y^{\lambda}-1)}{-1}$ frente a x se presenta en la figura 6.30, y el de los residuos de $\log y$ frente a $\log x$, en la figura 6.31.

La figura 6.30 muestra que la transformación inversa ha hecho desaparecer el crecimiento de la varianza, aunque ha acrecentado la no linealidad. La figura 6.31 presenta la gráfica de los residuos cuando transformamos ambas variables mediante el logaritmo. Se observa que ha desaparecido la heterocedasticidad, y la falta de linealidad también ha mejorado ostensiblemente.

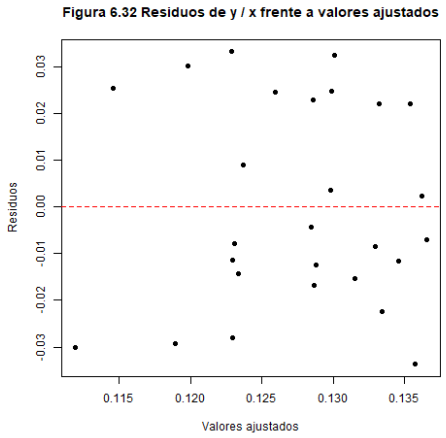
```
# Transformación log-log: log(y) ~ log(x)
data$LogSupervisores <- log(data$Supervisores)
data$LogTrabajadores <- log(data$Trabajadores)
modelo_log <- lm(LogSupervisores ~ LogTrabajadores, data = data)
summary(modelo_log)
# Gráfico 6.31: Residuos del modelo log-log
png("figura_6_31.png")
plot(fitted(modelo_log), residuals(modelo_log),
     main = "Figura 6.31 Residuos del modelo log-log",
     xlab = "Valores ajustados", ylab = "Residuos", pch = 16, col =
"black")
abline(h = 0, col = "red", lty = 2, lwd = 1.5)
dev.off()
```



Para linearizar la relación en logaritmos tenemos que transformar la variable x, ya que no queremos modificar la respuesta, que es ahora homocedástica. La transformación tiene que comprimir los valores de la x con relación al logaritmo, lo que supone que, por ejemplo, debemos tomar $\lambda = -0.5$. La figura 6.32 muestra los residuos de esta regresión y se observa que la relación es ahora aproximadamente lineal. A la hora de elegir un modelo conviene tener en cuenta su interpretación. El modelo estimado en logaritmos es $\log y = -1.48 + 0.909 \log x$. Este indica que un aumento del 10% en el número de trabajadores supone un aumento promedio del 9.09% en el número de supervisores. El $R^2 = 0.88$. El modelo con la transformación inversa tiene una interpretación más compleja.

```
# Gráfico 6.32: Residuos de y / x frente a valores ajustados
png("figura_6_32.png")
```

```
plot(fitted(modelo_y_por_x), residuals(modelo_y_por_x),
     main = "Figura 6.32 Residuos de y / x frente a valores ajustados",
xlab = "Valores ajustados", ylab = "Residuos", pch = 16, col = "black")
abline(h = 0, col = "red", lty = 2, lwd = 1.5)
dev.off()
```

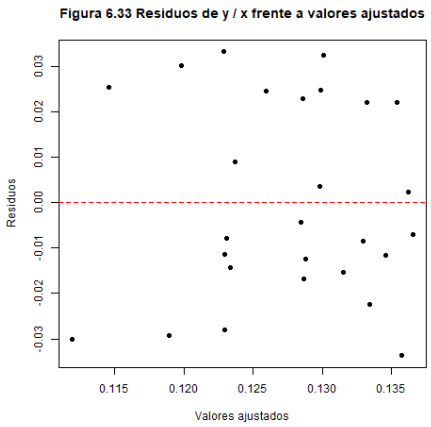


Otra forma posible de modelar estos datos es utilizar como respuesta la variable y/x, el número de supervisores por trabajador que será una variable homocedástica. La regresión de esta variable respecto a x conduce al modelo

$$\frac{y}{x} = 0.141 - 0.000018x$$
$$(14.57) \quad (-1.53)$$

donde ahora, por simplicidad, hemos indicado debajo de los coeficientes los estadísticos t. Se observa que la relación entre el número de supervisores y el de trabajadores es aproximadamente constante, alrededor de 0.14, es decir, aproximadamente 1.4 supervisores para cada 10 trabajadores, ya que la pendiente no es significativa. El p-valor de 0.138 para la t sugiere que es posible que al aumentar el número de trabajadores el número de supervisores decrezca algo. Esto es consistente con los resultados anteriores. La figura 6.33 muestra que no hay señales claras de error de especificación en el modelo.

```
# Gráfico 6.33: Residuos frente a valores ajustados (final)
png("figura_6_33.png")
plot(fitted(modelo_y_por_x), residuals(modelo_y_por_x),
     main = "Figura 6.33 Residuos de y / x frente a valores ajustados",
     xlab = "Valores ajustados", ylab = "Residuos", pch = 16, col = "black")
abline(h = 0, col = "red", lty = 2, lwd = 1.5)
dev.off()
```



1.4 Consecuencias de las transformaciones

Las transformaciones en las observaciones permiten analizar relaciones de manera lineal cuando se aplican logaritmos. Por ejemplo, la ecuación $y = ax^{\beta}u$ puede reescribirse como $\ln y = \ln a + \beta \ln x + \ln u$, lo que facilita trabajar con variables transformadas, definidas como y' y x'. Esta transformación afecta directamente la distribución del término de error u. Si originalmente sigue una distribución normal, el término transformado $\ln u$ seguirá una log-normal, lo que implica que la esperanza y la varianza asociadas a u pueden calcularse como $E[u] = e^{\sigma^2/2}$ y $D^2(u) = e^{\sigma^2}\left(e^{\sigma^2} - 1\right)$, respectivamente.

En la ecuación original, al tomar las expectativas, se obtiene que $E[y] = ax^{\beta}e^{\sigma^2/2}$. Este resultado introduce un sesgo en la predicción debido a la transformación aplicada. Específicamente, el estimador $\hat{y}$, derivado del modelo transformado, es sesgado al trabajar en la escala original. Este sesgo es proporcional a $\sigma^2/2$ y surge por la varianza del término $\ln u$. Aunque la presencia de este sesgo podría ser considerada una desventaja, se compensa con la reducción general del error cuadrático medio, lo que hace que las transformaciones sean una herramienta útil para mejorar la eficiencia predictiva.

Finalmente, investigaciones como las de Heien (1968) y Goldberger (1968) destacan la relevancia de abordar y corregir el sesgo introducido por estas transformaciones logarítmicas. Estas investigaciones proponen métodos alternativos para mejorar la precisión de las predicciones en la escala original, subrayando la importancia de las transformaciones en la modelización estadística.

Interpretación de los coeficientes

Cuando se interpretan los coeficientes en modelos de regresión transformados, es fundamental tener en cuenta las implicaciones de dichas transformaciones respecto a la relación estudiada entre las variables. Dependiendo de la transformación aplicada, los coeficientes adquieren diferentes significados:

1. En un modelo lineal simple de la forma:

$$y = \beta_0^{(1)} + \beta_1^{(1)}x,$$

el coeficiente $\beta_1^{(1)}$ representa el incremento absoluto en la variable dependiente y cuando la variable independiente x aumenta en una unidad.

2. En un modelo semi-logarítmico, expresado como:

$$\ln(y) = \beta_0^{(2)} + \beta_1^{(2)} x,$$

el coeficiente $\beta_1^{(2)}$ indica aproximadamente el cambio porcentual que experimenta y si x aumenta en una unidad. Esta interpretación se basa en la propiedad del logaritmo natural, donde $\ln(1+z) \approx z$ para valores pequeños de z.

3. En un modelo log-log ajustado, que adopta la forma:

$$\ln(y) = \beta_0^{(3)} + \beta_1^{(3)} \ln(x),$$

el coeficiente $\beta_1^{(3)}$ dividido por 100 se interpreta como el incremento porcentual en y asociado a un incremento del 1% en x.

4. En un modelo de doble logaritmo, descrito por:

$$\ln(y) = \beta_0^{(4)} + \beta_1^{(4)} \ln(x),$$

el coeficiente $\beta_1^{(4)}$ es conocido como elasticidad. Este término describe el cambio porcentual en y frente a un cambio porcentual en x. Es especialmente relevante en análisis económicos donde se busca capturar proporciones entre variables.

La elasticidad, en este contexto, mide la sensibilidad proporcional de y respecto a x. Por ejemplo, un valor de $\beta_1^{(4)} = 1.2$ indica que, por cada incremento del 1% en x, y aumentará un 1.2%.

Por lo tanto, cada coeficiente β en estos modelos tiene un significado particular y debe interpretarse teniendo en cuenta tanto el modelo como la transformación aplicada.

Interpretación de la varianza residual

Al analizar la varianza residual en un modelo logarítmico, se observa que esta adquiere una interpretación específica. En este contexto, la varianza de los errores se relaciona directamente con el error relativo o porcentual de la estimación de la variable dependiente sin transformar. Esto implica que, al trabajar con transformaciones logarítmicas, la varianza de la perturbación no depende del nivel de la variable transformada, manteniéndose constante para diferentes valores. Esta constancia contrasta con los modelos sin transformar, donde la variabilidad tiende a incrementarse a medida que aumenta el nivel de la variable dependiente.

En términos matemáticos, el error en el modelo logarítmico puede aproximarse como:

$$u_i = \ln\left(\frac{y_i}{\mu_i}\right),$$

lo que refleja la relación entre los valores observados y los valores ajustados del modelo. Además, la varianza de u_i, calculada como:

$$\text{Var}(u_i) = \frac{\sigma^2}{\mu_i^2},$$

resalta que la transformación logarítmica estabiliza la variabilidad relativa, igualándola aproximadamente a la de la variable transformada, independientemente del nivel absoluto de los datos originales. Esto subraya la utilidad de las transformaciones logarítmicas para controlar la heterocedasticidad en los modelos de regresión y proporcionar estimaciones más consistentes y robustas.

2. Regresión no paramétrica

Los métodos de regresión no paramétrica buscan estimar la forma de la esperanza condicional $m(x)$ directamente, sin requerir una estructura específica para el modelo. Esta

aproximación se obtiene promediando las observaciones de la respuesta en los valores de x cercanos al punto de interés, ponderadas según una función $w_n(x)$. La estimación resultante, $\widehat{m}(x)$, es una media local que depende del peso asignado a las observaciones cercanas y decrece a medida que aumenta la distancia al punto de interés.

Un procedimiento común para determinar los pesos es el método del núcleo, donde se utiliza una función de densidad K para asignar los pesos, y un parámetro h controla la suavidad de la función. Si h es grande, la función será más suave, pues se asigna un peso uniforme a un rango amplio de observaciones. Si h es pequeño, los valores cercanos a x predominan, y la función núcleo reproduce puntos observados casi exactamente.

La elección adecuada de h es crucial para garantizar una representación fiel de los datos. En el límite, cuando $h \to \infty$, la estimación converge al promedio global $\bar{y}$. Por el contrario, cuando $h \to 0$, la estimación converge al valor puntual y_i. Se recomienda probar distintos valores de h para encontrar un equilibrio entre suavidad y ajuste a los datos.
Los residuos $e_i = y_i - \widehat{m}(x_i)$ se calculan tras obtener $\widehat{m}(x)$ y permiten evaluar la adecuación del modelo no paramétrico, de manera similar a los análisis realizados en modelos lineales.

3. Predicción

3.1 Estimación de las medias condicionadas

Un modelo de regresión permite estimar la media de las distribuciones de la variable dependiente y para cada valor de la variable independiente x. Esta estimación se realiza a través de la ecuación de regresión, y los valores numéricos obtenidos para la media y la predicción de y son idénticos, aunque la precisión de cada uno puede diferir.

La precisión al estimar la media de la distribución condicionada de y se analiza utilizando un estimador centrado, que tiene como media el valor esperado de y. La varianza de este estimador depende del tamaño muestral y del efecto palanca h, que mide la influencia de un punto x_h respecto al promedio de x. Cuando los puntos están dentro del rango observado, el proceso se denomina interpolación, mientras que para puntos fuera del rango se conoce como extrapolación. En el caso de extrapolación, la precisión puede disminuir significativamente, ya que la varianza no necesariamente disminuye con el tamaño muestral.

Para manejar estas diferencias, se introduce el número equivalente de observaciones, que depende del inverso del efecto palanca. Este número ajusta la varianza para reflejar la información efectiva disponible en la estimación de la media condicionada.

Intervalos para las medias

Al construir intervalos de confianza para la media estimada m_h, se considera la distribución del error estandarizado del estimador, que sigue una distribución normal estándar. A partir de este enfoque, se calcula un estadístico t con $n-2$ grados de libertad. Este estadístico permite determinar un intervalo de confianza que incluye la varianza residual y el número equivalente de observaciones.

El intervalo de confianza proporciona un rango dentro del cual se espera que se encuentre el valor verdadero de m_h, considerando un nivel de confianza específico. Este procedimiento es esencial para evaluar la fiabilidad de las estimaciones en un modelo de regresión.

3.2 Predicción de una nueva observación

Para predecir el valor de y dado $x = x_h$, se sustituye x_h en la ecuación de regresión para obtener $\widehat{y_h}$, que representa la media de la distribución condicionada. Este valor se toma como la predicción óptima.

Estimación y predicción

La predicción y la estimación comparten principios, pero su enfoque conceptual difiere. La estimación busca minimizar la variabilidad del estimador alrededor de un parámetro desconocido y fijo, mientras que la predicción minimiza el error de predicción considerando la variabilidad de una variable aleatoria. Un ejemplo claro es el caso de estimar la media μ de una población fija, donde se minimiza el error cuadrático medio de estimación $(ECME)$, lo que lleva al uso de la media muestral $\hat{y} = \bar{y}$. Por otro lado, al predecir y para un individuo al azar en la población, se minimiza el error cuadrático medio de predicción $(ECMP)$, que se puede descomponer en términos de la varianza de la estimación de la media y la varianza de y alrededor de su media teórica.

Criterios de predicción

La predicción óptima se define según un criterio específico, generalmente buscando minimizar el valor esperado del error cuadrático medio de predicción $E[(y - \widehat{y_h})^2]$. Este enfoque asegura que el predictor óptimo sea igual a la esperanza matemática de la variable y dado x_h, lo que implica $\widehat{y_h} = E[y]$.

Intervalo de confianza

El predictor $\widehat{y_h}$ es centrado y su varianza se descompone como la suma de la varianza de y y la varianza de $\widehat{y_h}$. Esto permite derivar un intervalo de confianza para la predicción, expresado como:

$$y_h = \widehat{y_h} \pm t_\alpha s_R \sqrt{1 + \widehat{h_h^{-1}}}$$

Bandas de confianza

Se generan bandas de confianza al unir los extremos de los intervalos de confianza construidos para un nivel de significación α y cada valor de x. Estas bandas pueden aplicarse tanto para la predicción de valores y_h como para las medias condicionadas m_h. La amplitud de las bandas de confianza es mayor para la predicción de y_h que para m_h, ya que en el primer caso se incorpora la variabilidad de y_h con respecto a su media. A medida que nos alejamos del punto central $(\bar{x}, \bar{y})$, las bandas de confianza se amplían debido a las discrepancias entre la pendiente estimada $\widehat{\beta_1}$ y la pendiente real β_1, lo que aumenta los errores en predicciones más alejadas del rango observado.

Riesgos de extrapolación

Los intervalos de confianza obtenidos son válidos únicamente bajo el supuesto de que el modelo sea correcto. Extrapolar fuera del rango observado conlleva riesgos debido a que la relación entre las variables puede no ser lineal en esos puntos. Por ejemplo, relaciones como el número de directivos frente al número de trabajadores, la velocidad de un vehículo frente a su consumo, o las emisiones contaminantes frente a su concentración en el aire, suelen ser lineales en intervalos limitados, pero no lo son fuera de ellos. La extrapolación fuera del rango observado, especialmente para valores lejanos de x_h, puede generar errores significativos debido al aumento del efecto palanca y la falta de datos empíricos que soporten esas predicciones.