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: 07 de diciembre de 2024

1. Transformaciones

1.1 Transformaciones para Linearizar Relaciones No Lineales

Las transformaciones se utilizan para convertir relaciones no lineales en lineales y abordar problemas asociados con la heterocedasticidad, es decir, la variabilidad no uniforme de los errores. Entre estas transformaciones, la de Box-Cox es una herramienta útil. La transformación Box-Cox se define para un parámetro λ y es expresada como:

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

Cuando el parámetro λ se aproxima a 0, la transformación Box-Cox equivale a aplicar el logaritmo a la variable y. Si λ es mayor que 1, la variable transformada crece más lentamente que la original, mientras que, si λ es menor que 1, la variable transformada crece más rápidamente que la original.

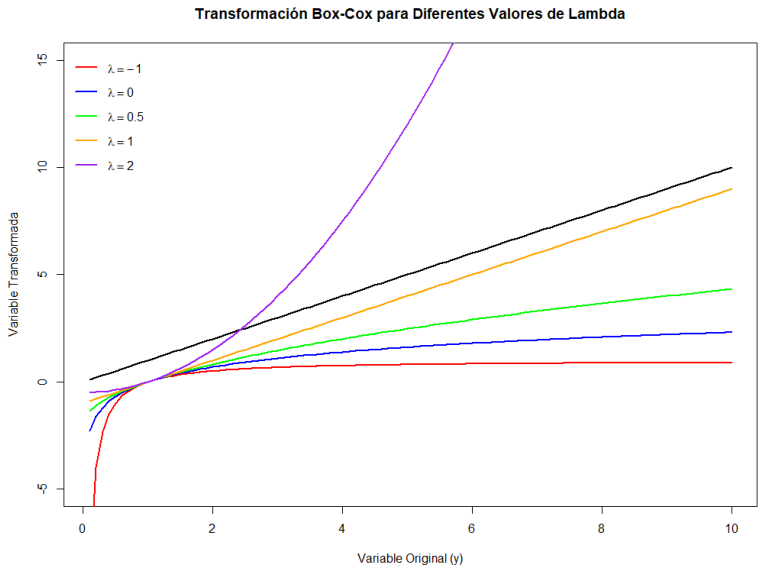


Ilustración 1. Gráfico de la Transformación Box-Cox

La elección del valor de λ tiene como objetivo transformar los datos para que cumplan mejor con las suposiciones de los modelos estadísticos, ajustándolos a una distribución que sea más compatible con los análisis requeridos. En este contexto:

- Si los residuos presentan heterocedasticidad (varianza no constante) o no siguen una distribución normal, se transforma la variable dependiente y para estabilizar la varianza o aproximar la normalidad.
- Si la variable explicativa x muestra problemas similares, se transforma para corregir estos inconvenientes y mejorar la linealidad o la relación entre las variables.

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 $\lambda < 1$, una transformación adecuada comprime el rango de valores, haciendo que la variable transformada crezca más lentamente que la original, y ajustando así la relación para ser más lineal. Este efecto es más pronunciado cuanto menor sea el valor de λ.

Por el contrario, si $\lambda > 1$, la transformación expande los incrementos pequeños en comparación con los mayores, haciendo que la variable transformada crezca más rápidamente que la original. Cuando $\lambda = 0$, se aplica una transformación logarítmica, lo que estabiliza la varianza y aproxima la normalidad de los datos.

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 linearización al elegir parámetros específicos para la transformación que compensen dichas variaciones. Por ejemplo, si la relación entre las variables muestra residuos con una tendencia creciente hacia valores más altos de la variable dependiente, un $\lambda > 1$ puede ser más adecuado. En cambio, si los residuos muestran una tendencia decreciente, un $\lambda < 1$ puede ser más efectivo.

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 de la transformación de la respuesta por máxima verosimilitud

La estimación del parámetro λ para realizar transformaciones tiene como objetivo optimizar la linealidad y homocedasticidad de los datos, así como reducir la varianza no explicada por el modelo. Este proceso se realiza comúnmente mediante el método de máxima

verosimilitud, que permite encontrar el valor de λ más adecuado para ajustar los datos a un modelo lineal. Para ello, se utiliza una fórmula general definida de la siguiente manera:

$$z(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda y^{\lambda - 1}} \text{ si } \lambda \neq 0 \\ \hat{y} \ln y \text{ si } \lambda = 0 \end{cases}$$

En esta fórmula, el parámetro λ es utilizado para modificar los datos y ajustarlos a un comportamiento más lineal.

La función de verosimilitud $L(\lambda)$ se define como:

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

donde VNE(λ) 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. Por lo que el intervalo de confianza aproximado para λ se puede calcular utilizando:

$$L_{\text{máx}}(\lambda) - \frac{1}{2}\chi^2_{1,\alpha}$$

En este caso, $\chi^2_{1,\alpha}$ representa el valor crítico de una distribución χ^2 con un grado de libertad para un nivel de significancia α.

El método también puede extenderse para manejar transformaciones simultáneas en varias variables. Por ejemplo, si λ_1 es el parámetro de la variable respuesta y y λ_2 el de la variable explicativa x, se fija un valor para λ_2 y se realiza el proceso anterior para encontrar λ_1. Este procedimiento se repite para distintos valores de λ_2, obteniendo una colección de máximos de verosimilitud $L_{\text{máx}}(\lambda_1|\lambda_2)$. El máximo absoluto entre estos representa el mejor ajuste del modelo para ambas variables. Este enfoque no solo mejora la linealidad y homocedasticidad, sino que también produce un modelo más robusto frente a problemas de heterocedasticidad o no linealidad.

1.2 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 $\text{Var}(y|x) = k \cdot \text{E}(y|x)^{\alpha}$, 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 la gráfica $R_h = f(\overline{y}_h)$, es de la forma:

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

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

Ejemplo 6.3

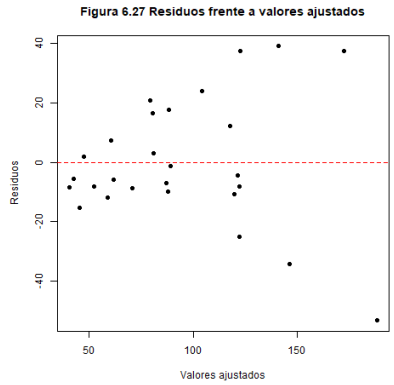
Se va a analizar los datos del ejercicio 5.2 sobre la relación entre el número de trabajadores y el de supervisores. Los datos se encuentran en el fichero trabajadores.dat. El modelo estimado para los datos originales se presenta a continuación.

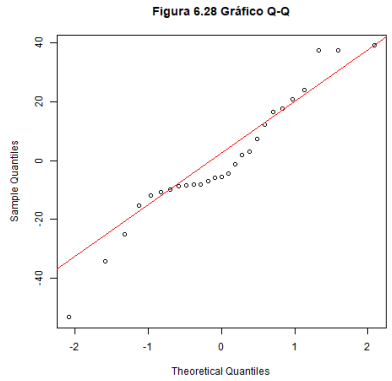
```
# Cargar el conjunto de datos
data <- read.table("trabajadores.dat", header = FALSE)
colnames(data) <- c("Supervisores", "Trabajadores")
# Modelo lineal inicial
modelo_inicial <- lm(Supervisores ~ Trabajadores, data = data)
summary(modelo_inicial)
```

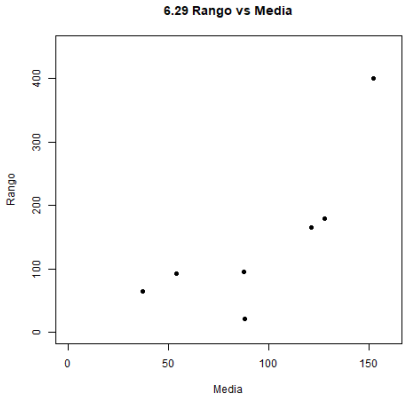
```
Call:
lm(formula = Supervisores ~ Trabajadores, data = data)

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
```

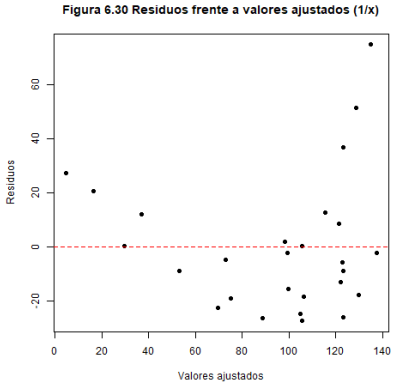
```
# Gráfico 6.27: Residuos frente a valores ajustados
png("figura_6_27.png")
plot(fitted(modelo_inicial), residuals(modelo_inicial),
     main = "Figura 6.27 Residuos frente a valores ajustados",
     xlab = "Valores ajustados", ylab = "Residuos", pch = 16, col =
"black")
abline(h = 0, col = "red", 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:

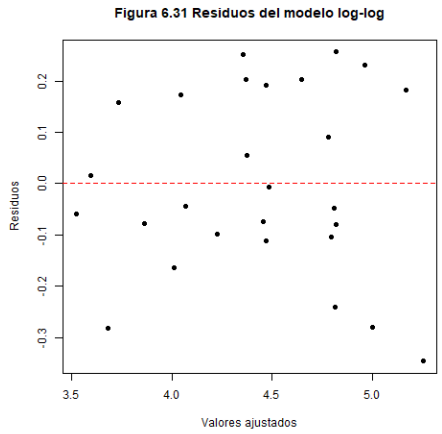
Grupos	**1-4**	**5-8**	**9-12**	**13-16**	**17-20**	**21-24**	**25-27**
Medias	37.00	53.75	87.50	88.00	121.00	127.75	152.30
Rangos	64	92	96	21	165	179	400

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.

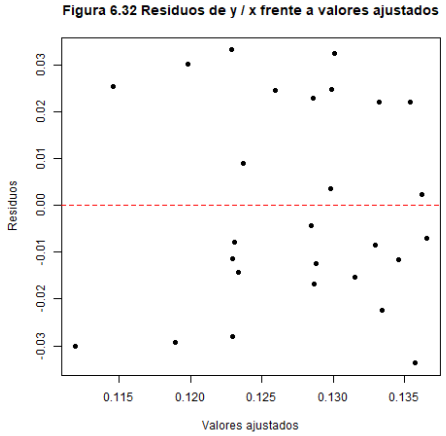
```
# Transformación log-log: log(y) ~ log(x)
data$LogSupervisores <- log(data$Supervisores)
data$LogTrabajadores <- log(data$Trabajadores)
modelo_log <- lm(LogSupervisores ~ LogTrabajadores, data = data)
summary(modelo_log)
# Gráfico 6.31: Residuos del modelo log-log
png("figura_6_31.png")
plot(fitted(modelo_log), 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()
```



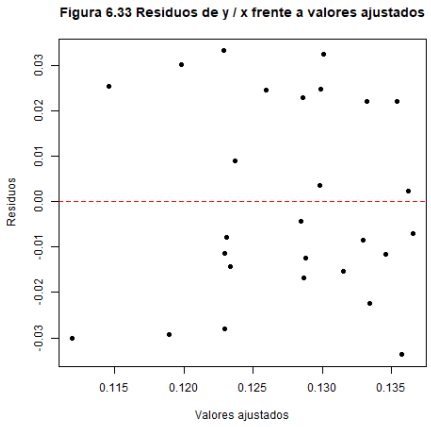
Otra forma posible de modelar estos datos es utilizar como respuesta la variable y/x, el número de supervisores por trabajador que será una variable homocedástica. La regresión de esta variable respecto a x conduce al modelo

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

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()
```



1.3 Consecuencias de las transformaciones

Sesgos en la predicción

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. El modelo de transformación logarítmica en x está definido como:

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

En este modelo, el coeficiente $\beta_1^{(3)}$ representa el cambio en y asociado a un incremento logarítmico en x. Por ejemplo, Δy puede ser expresado como:

$$\Delta y = \beta_1^{(3)} \ln\left(\frac{x_2}{x_1}\right),$$

donde $\frac{\beta_1^{(3)}}{100}$ puede interpretarse como el incremento en y cuando x aumenta en un 1\%.

Este modelo es útil cuando la variable x tiene una distribución amplia y es necesario capturar su efecto relativo (proporcional) sobre y. Además, permite ajustar relaciones donde x tiene un impacto no lineal pero puede representarse como lineal al aplicar el logaritmo.

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

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

En el modelo log-log, la relación se ajusta según la siguiente ecuación:

$$\ln\left(\frac{y_2}{y_1}\right) = \beta_1^{(4)} \ln\left(\frac{x_2}{x_1}\right) = \beta_1^{(4)} \ln\left(1 + \frac{1}{x_1}\right),$$

que implica, aproximadamente:

$$\frac{\Delta y}{y_1} = \beta_1^{(4)} \frac{1}{x_1}.$$

En este modelo, el coeficiente $\beta_1^{(4)}$ representa el incremento porcentual de y cuando x aumenta un 1%. Este tipo de modelo se utiliza frecuentemente en análisis económicos, donde se busca capturar relaciones proporcionales entre variables. Al coeficiente $\beta_1^{(4)}$ se le denomina elasticidad, ya que mide la sensibilidad proporcional de y con 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.

5. 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. La expresión de la estimación resultante es:

$$\widehat{m}(x) = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}$$

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.

6. Predicción

6.1 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 $\widehat{y_h}$, donde:

$$\widehat{y_h} = \bar{y} + \beta_1(x_h - \bar{x}),$$

y:

$$E[\widehat{y_h}] = \beta_0 + \beta_1\bar{x} + \beta_1(x_h - \bar{x}) = \beta_0 + \beta_1 x_h = m_h.$$

La varianza de este estimador está dada por:

$$\mathrm{Var}(\widehat{y_h}) = \frac{\sigma^2}{h_h},$$

donde $\hat{n}_h$ es una medida del efecto palanca de una observación, que depende de la posición de x_h respecto al promedio de x.

Cuando los puntos x_h están dentro del rango observado en la muestra, el proceso se denomina interpolación, y h_h satisface $\frac{1}{n} \leq h_h \leq 1$. En este caso, la precisión de la estimación mejora con el tamaño de la muestra. En contraste, cuando x_h está fuera del rango observado, se habla de extrapolació. En este caso, la precisión de la estimación puede no mejorar con el aumento del tamaño de la muestra, ya que la varianza puede ser considerablemente mayor.

Para manejar estas diferencias, se introduce el concepto de número equivalente de observaciones, $(\widehat{n_h})$, definido como:

$$\widehat{n_h} = \frac{1}{h_h}.$$

Este número representa la cantidad efectiva de información disponible para la estimación en x_h. La varianza de $\widehat{y_h}$ puede reescribirse como:

$$\mathrm{Var}(\widehat{y_h}) = \frac{\sigma^2}{\widehat{n_h}}.$$

Esta expresión indica que la varianza de $\widehat{y_h}$ tiene una forma similar a la de la varianza marginal $\frac{\sigma^2}{n}$, pero depende del efecto palanca h_h del punto considerado.

Intervalos para las medias

Para construir intervalos de confianza para la media estimada m_h, se utiliza la distribución del error estandarizado del estimador, que sigue una distribución normal estándar. A partir de este enfoque, se define el siguiente estadístico:

$$t_{n-2} = \frac{\hat{y}_h - m_h}{\hat{S}_R/\sqrt{\hat{n}_h}}$$

donde $\hat{y}_h$ es la media estimada, m_h es el valor verdadero de la media, S_R es el error estándar residual, y $\hat{n}_h$ representa el número equivalente de observaciones. El estadístico t_{n-2} tiene $n-2$ grados de libertad y permite calcular un intervalo de confianza que incluye la varianza residual y el número equivalente de observaciones. De esta forma, se obtiene un intervalo de confianza al nivel de significancia α:

$$m_h = \hat{y}_h \pm t_{\alpha/2}\hat{S}_R/\sqrt{\hat{n}_h}$$

Este intervalo proporciona un rango dentro del cual se espera encontrar el valor verdadero de m_h, considerando un nivel de confianza específico.

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 ($\widehat{y_h} = \widehat{\beta_0} + \widehat{\beta_1}x_h$) para obtener $\widehat{y_h}$, que representa la media de la distribución condicionada. Este valor se toma como la predicción óptima.

Estimación y predicción

Aunque la predicción y la estimación comparten principios conceptuales, su enfoque es distinto. La estimación se centra en minimizar la variabilidad del estimador alrededor de un parámetro desconocido pero fijo. Esto implica minimizar el error cuadrático medio de estimación ($ECME$):

$$ECME(\hat{y}) = E[(\hat{y} - \mu)^2],$$

lo que lleva a la elección de la media muestral como el estimador óptimo, es decir, $\hat{y} = \bar{y}$. La media muestral minimiza la varianza de la distribución muestral del estimador.

Por otro lado, la predicción considera minimizar el error cuadrático medio de predicción $ECMP$:

$$ECMP(\hat{y}) = E[(y - \hat{y})^2],$$

donde y es una variable aleatoria. El predictor óptimo para minimizar este error es $\hat{y} = E(y)$, lo que generalmente se aproxima por la media muestral $\bar{y}$. En este caso, el $ECMP$ puede descomponerse en términos de la varianza de la estimación de la media ($E[(\bar{y} - \mu)^2]$) y la varianza de la variable alrededor de su media teórica ($E[(y - \mu)^2]$):

$$ECMP(\bar{y}) = E[(y - \bar{y})^2] = E[(\bar{y} - \mu)^2] + E[(y - \mu)^2]$$

Esta descomposición refleja cómo la precisión en la estimación de la media y la variabilidad inherente de la variable aleatoria contribuyen conjuntamente al error de predicción. Este enfoque permite evaluar tanto la confiabilidad del modelo como su capacidad predictiva en el contexto del análisis estadístico.

Criterios de predicción

La predicción óptima depende del criterio elegido, que generalmente busca minimizar el valor esperado del error cuadrático medio de predicción $E[(y - \widehat{y_h})^2]$. Si $\widehat{y_h}$ es un predictor cualquiera para y dado $x = x_h$, el error de predicción, definido como $e_h = y_h - \widehat{y_h}$, se minimiza tomando $\widehat{y_h}$ igual a la esperanza matemática de y condicionado a x_h, lo que implica $\widehat{y_h} = E[y]$.

Más generalmente, para $\alpha > 1$, se busca minimizar $E[(y - \widehat{y_h})^{\alpha}]$. Este criterio determina una clase de estimadores, y en el caso específico del error cuadrático medio, implica tomar la media de la distribución condicionada, es decir, $\widehat{y_h} = E[y|x_h]$. Este resultado fue demostrado formalmente en los años treinta por Wiener y Kolmogorov.

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:

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

El predictor $\hat{y}_h$ es centrado, lo que implica que:

$$E[\hat{y}_h] = E\left[\widehat{\beta_0} + \widehat{\beta_1} x_h\right] = \beta_0 + \beta_1 x_h = E[y_h].$$

Esto asegura que el error de predicción sea cero en promedio. La varianza de la predicción se descompone en dos términos principales: la varianza de y_h y la varianza de $\hat{y}_h$. Utilizando esta relación, el error cuadrático medio de predicción se expresa como:

$$E[(y_h - \hat{y}_h)^2] = \mathrm{Var}(y_h) + \mathrm{Var}(\hat{y}_h).$$

Dado que $\mathrm{Var}(y_h) = \sigma^2$ y $\mathrm{Var}(\hat{y}_h)$ está dada por $\sigma^2 \hat{n}_h^{-1}$, sustituimos para obtener:

$$E[(y_h - \widehat{y_h})^2] = \sigma^2 \left[1 + \hat{n}_h^{-1}\right].$$

A partir de esta expresión, el intervalo de confianza para y_h a un nivel de confianza α está dado por:

$$y_h = \widehat{y_h} \pm t_{\alpha/2} \hat{s}_R \sqrt{1 + \hat{n}_h^{-1}},$$

donde s_R es el error estándar residual y $t_{\alpha/2}$ corresponde al valor crítico de la distribución t de Student con los grados de libertad adecuados.

Este intervalo combina la incertidumbre de la estimación de y_h y la variabilidad inherente de los datos, proporcionando una herramienta robusta para evaluar la precisión de la predicción.

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.

Las bandas de confianza se generan 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.

Comparando las expresiones de los intervalos de confianza, es evidente que la amplitud de las bandas de confianza es mayor para la predicción de $y_h$que para m_h. Esto se debe a que, en el primer caso, se incluye la variabilidad de $y_h$con respecto a su media m_h. Por otro lado, al alejarse 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. Estas discrepancias incrementan los errores de predicción conforme nos alejamos 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 importantes debido a la posibilidad de que la relación entre las variables no sea lineal en esos puntos. En muchos casos, es esperable que la relación cambie significativamente, lo que compromete la validez de las predicciones.

Por ejemplo, relaciones como el número de directivos frente al número de trabajadores en una organización pueden ser lineales para empresas de tamaño mediano, pero no para empresas muy pequeñas o muy grandes. Asimismo, la relación entre la velocidad de un vehículo y su consumo puede ser lineal a velocidades moderadas, pero a velocidades cercanas al límite de potencia del motor, el consumo aumenta exponencialmente. De manera similar, las emisiones contaminantes frente a su concentración en el aire suelen ser lineales en intervalos limitados, pero no lo son fuera de ellos.

Cuando se extrapola fuera del rango observado, especialmente para valores lejanos de x_h, se incrementa el efecto palanca de los puntos extremos, lo que puede producir errores significativos en las predicciones. Además, al alejarse del rango observado, el número equivalente de observaciones para predecir estos puntos extremos tiende a cero, lo que implica la falta de una base empírica que respalde estas predicciones. Por tanto, la extrapolación debe realizarse con cautela, considerando las posibles limitaciones y los riesgos asociados a la falta de linealidad y a la insuficiencia de datos.