Regresión Lineal Simple - Semana 03

Johnatan Cardona Jiménez
jcardonj@unal.edu.co
Profesor Asistente - Departamento de Estadística
Universidad Nacional de Colombia, Sede Medellín

Semestre 02-2024

Validación de los supuestos sobre los errores ε_i del modelo de RLS

Recuerde que los supuestos sobre los errores asumidos en el modelo de RLS se pueden resumir como:

$$\varepsilon_i \overset{\text{iid.}}{\sim} N(0\ ,\ \sigma^2),\ \ i = 1, \ldots, n$$

donde, iid. es la abreviación de independiente e idénticamente distribuido.

Luego, para la validación del modelo se deben probar los supuestos:

- Los errores del modelo tienen media cero.
- Los errores del modelo tienen varianza constante.
- Los errores del modelo se distribuyen normal.
- Los errores del modelo son independientes.

Para ello se usan los residuales del modelo

$$e_i = Y_i - \widehat{Y}_i, \ \ i = 1, \ldots, n,$$

que pueden ser consideradas como estimaciones de los errores del modelo ε_i.

Los errores del modelo tienen media cero

Usando los residuales del modelo podemos probar que:

$$\begin{aligned}\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(Y_i - \widehat{Y}_i) = \sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i) \\ &= \sum_{i=1}^{n}\left[Y_i - \left(\bar{Y} - \widehat{\beta}_1\bar{X}\right) - \widehat{\beta}_1 X_i\right] \\ &= \sum_{i=1}^{n}\left[\left(Y_i - \bar{Y}\right) - \widehat{\beta}_1\left(X_i - \bar{X}\right)\right] \\ &= \underbrace{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)}_{=0} - \widehat{\beta}_1\underbrace{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)}_{=0} = 0\end{aligned}$$

por lo tanto, el supuesto de media cero de los errores siempre se cumple.

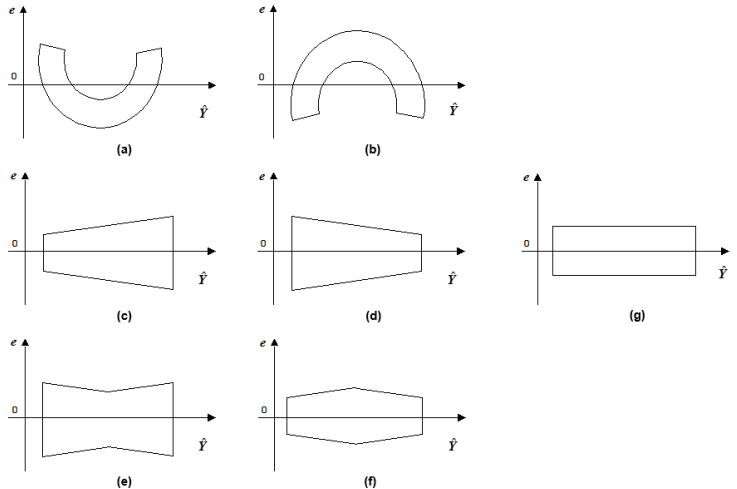
Los errores del modelo tienen varianza constante

El supuesto de varianza constante (homogeneidad de varianza) se puede validar a través de un gráfico de residuales vs. valores ajustados o predichos, donde se quiere probar:

$$\begin{cases} H_0 : V\left[\varepsilon_i\right] = \sigma^2 \\ \\ H_1 : V\left[\varepsilon_i\right] \neq \sigma^2 \end{cases}$$

La siguiente Figura muestra algunos patrones comunes de la nube de puntos en los gráficos de residuales, que sirven para detectar si este supuesto se cumple, incluso en algunas ocasiones sirven para detectar un mal ajuste del modelo lineal.

Patrones comunes en residuales: (a) y (b) Mala expeficifación del modelo (predictor cuadrático no considerado). (c), (d), (e) y (f) Varianza No Constante del error. (g) **Modelo Lineal y Varianza Constante**.



Otras opciones para evaluar el supuesto de varianza constante:

- Se puede recurrir a procedimientos inferenciales de Homogeneidad de Varianza, una de ellas es la prueba de Levene Modificada, que No depende del supuesto de normalidad.
- La prueba de Levene Modificada es aplicable cuando la varianza se incrementa o disminuye con X y los tamaños de muestra necesitan ser suficientemente grandes para que la dependencia entre los residuales pueda ser ignorada.
- **En este curso usaremos solo la Prueba Gráfica basada en el gráfico de Residuales vs. Valores Predichos.**

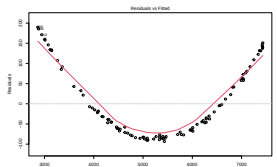
Algunas soluciones al problema de "Varianza No Homogénea"

1. Mínimos Cuadrados Ponderados cuando la Varianza del Error Varía de forma sistemática.
   En la función objetivo de mínimos cuadrados, las diferencias entre los valores observados y esperados de y_i se multiplican por pesos o factores de ponderación ω_i, tomados en forma inversamente proporcional a la varianza de y_i, esto es, la función de mínimos cuadrados considerada es:

$$S(\beta_0\ ,\ \beta_1) = \sum_{i=1}^{n} \omega_i (Y_i - \beta_0 + \beta_1 X_i)^2$$

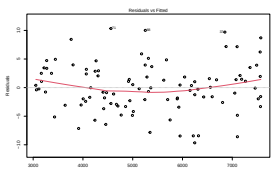
2. Usar Transformaciones en Y que estabilicen la varianza. En muchas ocasiones emplear transformaciones logarítmicas puede ayudar a estabilizar la varianza no constante presente en el modelo con la variable original.
3. Emplear una relación funcional adecuada. Por ejemplo, las situaciones (a) y (b) del gráfico anterior se podría solucionar el problema de varianza no constante adicionando un término cuadrático. **Ejemplo:**

```
Datos = read.csv("DataExamp1.csv")
x = Datos$X
y = Datos$Y
modelo = lm(y ~ x)
plot(modelo, 1)
```



En el modelo anterior el gráfico de resudiales parece indicar la ausencia de un término cuadrático. Elevando al cuadrado la variable independiente obtenemos el siguiente resultado:

```
Datos = read.csv("DataExamp1.csv")
x2 = (Datos$X)^2
y = Datos$Y
modelo = lm(y ~ x2)
plot(modelo, 1)
```



Los errores del modelo se distribuyen normal

En la validación del supuesto de normalidad se evalúa:

$$\begin{cases} H_0 : \varepsilon_i \sim \text{ Normal} \\ \\ H_1 : \varepsilon_i \not\sim \text{ Normal} \end{cases}$$

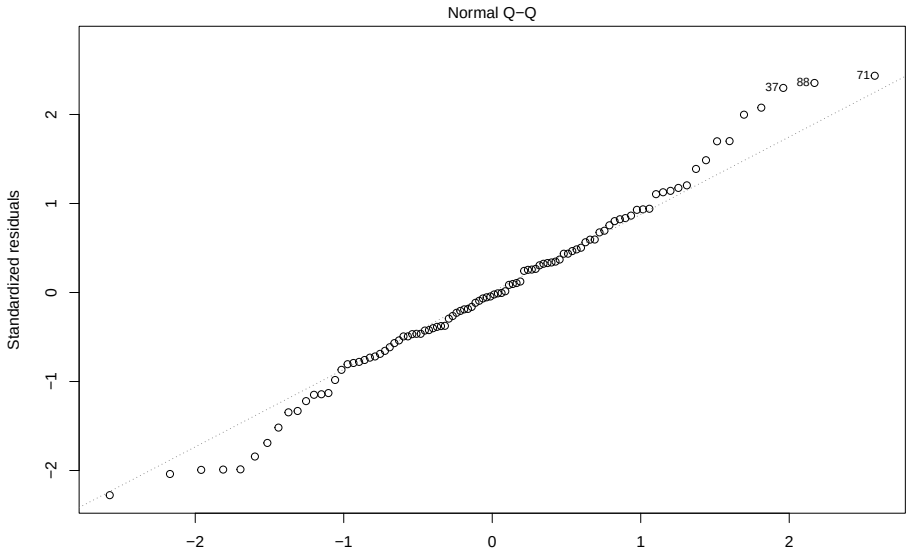
Esto se puede realizar bien sea examinando los **p-valores** arrojados por una prueba específica de normalidad, como la prueba de Shapiro-Wilk, o mediante un gráfico de normalidad, en el cual se evalúa si la nube de puntos en la escala normal se puede ajustar por una línea recta.

Ejemplo:

```
Datos = read.csv("DataExamp1.csv")
x2 = (Datos$X)^2
y = Datos$Y
modelo = lm(y ~ x2)
shapiro.test(modelo$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  modelo$residuals
## W = 0.9877, p-value = 0.486
```

```r
plot(modelo, 2)
```



Algunas soluciones al problema de “No Normalidad”

- La desviación del supuesto de normalidad **frecuentemente va de la mano** con la no homogeneidad de la varianza, por ello, a menudo una misma transformación de los valores de Y, logra estabilizar la varianza y una aproximación a la normalidad. En estos casos se debe usar primero una transformación que estabilice la varianza y evaluar si el supuesto de normalidad se cumple para los datos transformados.
- Otra solución es trabajar con métodos no paramétricos de regresión, los cuales no asumen algún modelo probabilístico en particular sobre la variable respuesta.

Los errores del modelo son independientes

- Evaluar este supuesto no es posible en muchas situaciones prácticas, pues para probar el supuesto de independencia **sería necesario conocer el orden de las observaciones en el tiempo**. En tal caso, se podría analizar el supuesto a través de un gráfico de residuales vs. el tiempo u orden de recolección de los datos. Se buscaría patrones sistemáticos como ciclos, rachas, o cualquier otro comportamiento que indique correlación entre los valores de la serie o secuencia de los residuales.
- El contexto en el cual los datos fueron obtenidos también nos podría dar información sobre el alcance de este supuesto. Por ejemplo, en el caso de un experimento controlado donde las ejecuciones se realizan de forma aleatoria se podría dar por sentado que este supuesto se ha cumplido, pues dada la elatoriedad en la ejecución se espera que las observaciones sean independientes.

Algunas soluciones al problema “no-independencia de los errores”

1. Trabajar con modelos con errores correlacionados.
2. Adicionar variables de tendencia, estacionalidad.
3. Trabajar con primeras diferencias.

Temática abordada en detalle en el curso de Estadística III.

Prueba de Falta de Ajuste

- Se define **Falta de Ajuste** en un modelo de regresión lineal cuando la relación funcional definidida entre la variable respuesta (Y) y la variable predictoria (X) no es la más apropiada. Se puede dar el caso de tener un modelo de regresión significativo, pero que presente falta de ajuste.
- La presencia de este problema puede identificarse a través del gráfico de **Residuales vs. Valores Predichos o versus Valores de la Variable Predictora**, de manera que cuando ocurre esta violación, el gráfico exhibe un patrón en el cual los residuales se desvían de cero en forma sistemática, por ejemplo, cuando la nube de puntos de estos gráficos presentan una forma de U o una forma de U invertida, como se observó en la Figura de los patrones de gráficos de residuales, partes (a) y (b).

Otra forma de probar la **falta de ajuste**, es mediante la prueba de Falta o Carencia de Ajuste, la cual prueba que un tipo específico de función de regresión ajusta adecuadamente a los datos.

Para el caso de la regresión lineal simple bajo una relación funcional de primer orden, se podría probar:

$$\begin{cases} H_0: \ E(Y_i) = E(Y|X_i) = \beta_0 + \beta_1 X_i \\ \\ H_1: \ E(Y_i) = E(Y|X_i) \neq \beta_0 + \beta_1 X_i \end{cases}$$

La prueba asume que los valores de Y dado X son:

- Independientes.
- Se distribuyen en forma normal.
- Tienen Varianza Constante.

Para esta prueba se requiere que en al menos un valor de X se haya tomado más de una observación de Y, ie. **que se tengan réplicas**.

Para explicar en qué consiste esta prueba, es necesario establecer una nueva notación, así:

- m: El número de valores distintos de X, denominados *niveles*.
- n_i: El número de observaciones de Y tomadas en el i-ésimo nivel de X. Por tanto, el número total de observaciones n tomadas cumple que $n = \sum_{i=1}^{m} n_i$.
- Y_{ij}: La j-ésima observación de la respuesta Y en el i-ésimo nivel de X, $i = 1, \ldots, m$, $j = 1, \ldots, n_i$.
- X_i: El i-ésimo nivel de X, $i = 1, \ldots, m$.
- $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$: promedio muestral de las n_i observaciones de Y tomadas en en el i-ésimo nivel de X, $i = 1, \ldots, m$.

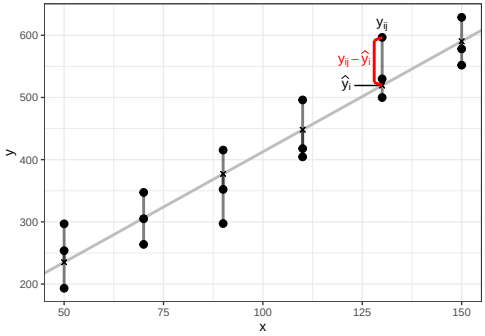
Para entender el significado de esta prueba, considere en la tabla ANOVA una nueva partición de la variabilidad, esta vez, del término del error, representada por la suma de cuadrados del error

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \widehat{y}_i)^2 ,$$

en dos componentes: **una debida a la falta de ajuste (LOF) y otra debida a lo que denominará un error puro (PE)**.

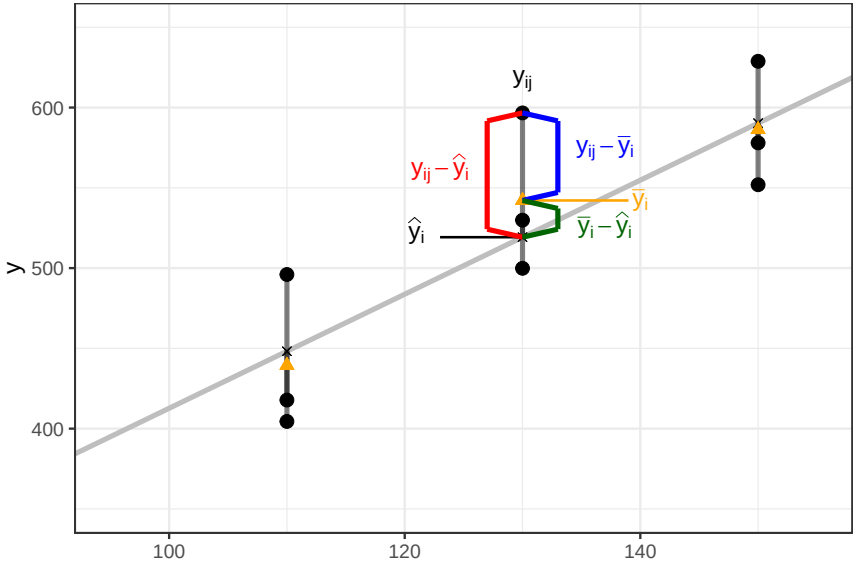
Veamos gráficamente como se da esta nueva partición de la variabilidad, para ello en la nueva notación consideremos las desviaciones $y_{ij} - \widehat{y}_i$.

Variabilidad en Y al incluir el modelo de RLS



Observe que: $m = 6$ y $n_i = 3\ \forall_i,\ \ i = 1, \ldots, m$.

Ilustración de la nueva descomposición de la variabilidad



De ahí que podamos escribir cada diferencia $y_{ij} - \widehat{y}_i$ como:

$$y_{ij} - \widehat{y}_i = (\bar{y}_i - \widehat{y}_i) + (y_{ij} - \bar{y}_i)$$

y reemplazando en la SSE, se obtiene:

$$\begin{aligned}
\text{SSE} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \widehat{y}_i)^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n_i}\Big[(\bar{y}_i - \widehat{y}_i) + (y_{ij} - \bar{y}_i)\Big]^2 \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i - \widehat{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 \\
&= \sum_{i=1}^{m} n_i\,(\bar{y}_i - \widehat{y}_i)^2 + \sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2
\end{aligned}$$

Tarea: comprobar que $2\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\bar{y}_i - \widehat{y}_i)(y_{ij} - \bar{y}_i) = 0$

Así, **la suma de cuadrados del error SSE queda expresada mediante la suma de dos componentes**, a saber:

- $\sum_{i=1}^{m} n_i \left(\bar{y}_i - \widehat{y}_i\right)^2$, que está relacionada con las diferencias entre los promedios de Y en cada nivel de la predictora X y los valores ajustados por el modelo de regresión, y que representan el desajuste del modelo de primer orden, al cual se le conoce como **Suma de Cuadrados de la Falta de Ajuste, abreviado SSLOF**.
- $\sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(y_{ij} - \bar{y}_i\right)^2$, que está relacionada con las diferencias entre las observaciones de la respuesta y los promedios de Y en cada nivel de la predictora X, por lo que a esta componente se le conoce como **Suma de Cuadrados del Error Puro, abreviado SSPE**.

De donde, se obtiene que: $\mathbf{SSE} = \mathbf{SSLOF} + \mathbf{SSPE}$.

Cada una de estas sumas de cuadrados tiene asociados unos grados de libertad (g.l):

- Se sabe que SSE tiene $n-2$ g.l.
- Analizando la expresión para SSPE, se tienen las mismas n observaciones y se estiman m medias de Y (una en cada nivel de la predictora X, y así SSPE tiene $n-m$ g.l.
- Finalmente, SSLOF tiene m observaciones (los promedios estimados) y se estiman los dos parámetros del modelo, de donde SSLOF tiene $m-2$ g.l.

Acá los grados de libertad (g.l) de las sumas de cuadrados también forman una identidad, así:

$$\begin{array}{ccccc} \text{g.l(SSE)} & = & \text{g.l(SSLOF)} & + & \text{g.l(SSPE)} \\ (n-2) & = & (m-2) & + & (n-m) \end{array}$$

A continuación, se definen **los cuadrados medios como la razón entre las sumas de cuadrados y sus respectivos grados de libertad**. Esto es,

- $\text{MSLOF} = \text{SSLOF}/\text{g.l(SSLOF)} = \text{SSLOF}/(m-2)$.
- $\text{MSPE} = \text{SSPE}/\text{g.l(SSPE)} = \text{SSPE}/(n-m)$.

Se puede demostrar que:

- $E\left[\text{MSPE}\right] = \sigma^2$.
- $E\left[\text{MSLOF}\right] = \sigma^2 + \frac{\sum_{i=1}^{m} n_i \left[E(Y_i) - \beta_0 - \beta_1 x_i\right]^2}{m-2}$.

Note que, bajo H_0 tanto MSLOF como MSPE son estimaciones independientes de σ^2.

De lo anterior, **se considera el siguiente estadístico de prueba**:

$$F_0 = \frac{\text{MSLOF}}{\text{MSPE}} = \frac{\text{SSLOF}/(m-2)}{\text{SSPE}/(n-m)} \sim F_{m-2\ ,\ n-m}$$

que bajo la hipótesis nula $H_0: \ E(Y_i) = \beta_0 + \beta_1 x_i$, se distribuye como una F con $(m-2)$ y $(n-m)$ grados de libertad.

Así, a un nivel de significancia α **se rechaza la hipótesis nula de que el modelo de primer orden es adecuado** (en favor de la hipótesis de que el modelo lineal tiene falta de ajuste) si $F_0 > F_{\alpha,m-2,n-m}$.

En la tabla ANOVA, presentada en clases anteriores, se puede incluir la prueba de falta de ajuste que descompone el SSE del modelo, así:

Análisis de varianza que incorpora la prueba de falta de ajuste en el modelo de RLS

Fuente de Variación	Suma de Cuadrados	Grados de Libertad	Cuadrado Medio	F Calculado
Regresión	SSR	1	$MSR = \frac{SSR}{1} = SSR$	$F_0 = \frac{MSR}{MSE}$
Error	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$	
Falta de Ajuste	$SSLOF$	$m-2$	$MSLOF = \frac{SSLOF}{m-2}$	$F_0 = \frac{MSLOF}{MSPE}$
Error Puro	$SSPE$	$n-m$	$MSPE = \frac{SSPE}{n-m}$	
Total	SST	$n-1$		

Ejemplo:

```r
library(model)
Datos = read.csv("DataExamp2.csv")
#Modelo de primer orden
modelo1 = lm(Y ~ X, data = Datos)
lack_fit_test(modelo1)
```

```
## Lack of fit test - Anova Table
##                Sum Sq  Df   Mean Sq F value    Pr(>F)
## Regression  548553612   1 548553612 91125.3 < 2.2e-16 ***
## Residuals     1793893 298      6020
## Lack of fit   1790682  98     18272  1138.4 < 2.2e-16 ***
## Pure error       3210 200        16
## Total       550347505 299
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

NOTAS:

- En general, en el cálculo de la SSPE sólo se utilizan aquellos niveles i de X en los cuales hay réplicas.
- En general, la prueba de falta de ajuste puede aplicarse a otras funciones de regresión, sólo se requiere modificar los grados de libertad del SSLOF, que en general corresponden a $m - p$, donde p es el número de parámetros en la función de regresión. Para el caso específico de la regresión lineal simple, $p = 2$.
- Cuando se concluye que el modelo de regresión en H_0 es apropiado, la práctica usual es usar el MSE y no el MSPE como un estimador de la varianza, debido a que el primero tiene más grados de libertad.
- Cualquier inferencia sobre los parámetros del modelo lineal, por ejemplo la prueba de significancia de la regresión, solo debe llevarse a cabo luego de haber probado que el modelo lineal es apropiado.

Algunas soluciones al problema “el modelo de regresión de primer orden no es apropiado”

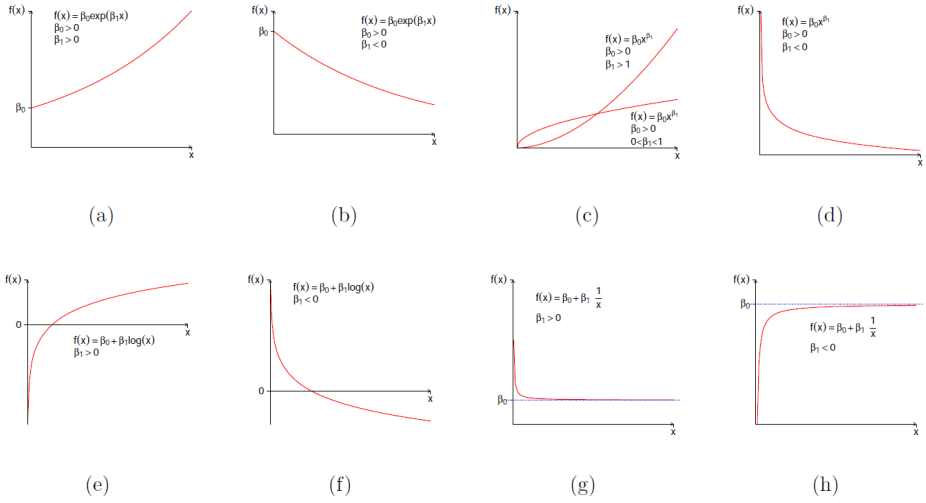
- Abandonar el modelo de regresión de primer orden y desarrollar un modelo más apropiado.
- Emplear alguna transformación en los datos de manera que el modelo de regresión lineal sea apropiado a los datos transformados (modelos intrínsecamente lineales).
- Se pueden usar curvas de regresión no paramétricas también llamadas curvas suavizadas, para explorar y/o confirmar la forma de la función de regresión, por ejemplo el método LOESS. En este caso la curva suavizada se grafica junto con las bandas de confianza del modelo de regresión; si la primera cae entre las segundas, entonces se tiene evidencia de que el modelo ajustado es apropiado.

Transformaciones: Modelos Intrínsecamente Lineales

Un modelo de regresión se considera lineal en los parámetros, por ello las transformaciones en las variables no implican modelos no lineales. Modelos intrínsecamente lineales son aquellos que relacionan Y con X por medio de una transformación en Y o en X, originando un modelo de la forma $Y^* = \beta_0 + \beta_1 X^* + \varepsilon$, donde Y^* y X^* son las variables transformadas.

Casos comunes de modelos Intrínsecamente Lineales

Modelo	Denominación	Transformación
$Y = \beta_0 e^{\beta_1 X} \varepsilon$	Modelo exponencial multiplicativo	Se ajusta $Y^* = \beta_0^* + \beta_1 X + \varepsilon^*$ con $Y^* = \ln(Y)$
$Y = \beta_0 X^{\beta_1} \varepsilon$	Modelo potencial multiplicativo	Se ajusta $Y^* = \beta_0^* + \beta_1 X^* + \varepsilon^*$ con $Y^* = \ln(Y)$ y $X^* = \ln(X)$
$Y = \beta_0 + \beta_1 \ln(X) + \varepsilon$	Modelo logarítmico	Se ajusta $Y = \beta_0 + \beta_1 X^* + \varepsilon$ con $X^* = \ln(X)$
$Y = \beta_0 + \beta_1 (1/X) + \varepsilon$	Modelo recíproco	Se ajusta $Y = \beta_0 + \beta_1 X^* + \varepsilon$ con $X^* = 1/X$



Gráficas de las funciones intrínsecamente lineales dadas en la Tabla 3.3: (a) y (b) en modelo exponencial; (c) y (d) en modelo de potencia;(e) y (f) en modelo logarítmico; (g) y (h) en modelo recíproco.

Ejemplo:

```r
library(model)
#Ejemplo transformaciones logaritmicas
Datos = read.csv("DatosExemp3.csv")
#str(Datos)
#Variables originales
#plot(Y ~ X, data = Datos)
#Variable transformada
#plot(LogY ~ X, data = Datos)
#Ajusto del modelo
modelo = lm(LogY ~ X, data = Datos)
#Tabla anova
anova_table_lm(modelo)
```

```
## Anova Table
##            Sum Sq Df Mean Sq F value    Pr(>F)
## Regression 4751.2  1  4751.2    4504 < 2.2e-16 ***
```

NOTAS:

1. Los modelos exponenciales y de potencia aditivos: $Y = \beta_0 e^{\beta_1 X} + \varepsilon$, $Y = \beta_0 X^{\beta_1} + \varepsilon$ no son intrínsecamente lineales.
2. El supuesto necesario es que cuando el término de error ε es transformado, esta variable transformada deberá ser iid $N(0, \sigma^2)$, por ello deben examinarse los residuales del modelo transformado.
3. En los casos con modelos exponenciales y potenciales multiplicativos, si σ es pequeño se puede obtener un intervalo de confianza aproximado para la respuesta media tomando antilogaritmos sobre los límites del intervalo hallado para la respuesta media de Y^*. Sin embargo, cuando hacemos esto, en términos generales, estamos hallando un intervalo de confianza para la mediana de Y (recordar la distribución lognormal).