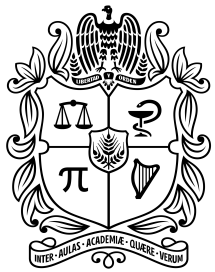
Estadística II - 3006915
Regresión Lineal Simple

Mateo Ochoa Medina

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín

Periodo académico 2023-2S



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Contenido

1. Validación de supuestos sobre los errores
2. Supuesto de linealidad del modelo: prueba de falta de ajuste
3. Referencias

Contenido

1. Validación de supuestos sobre los errores
2. Supuesto de linealidad del modelo: prueba de falta de ajuste
3. Referencias

Validación de los supuestos sobre los errores ε_i

Se sabe que los supuestos sobre los errores asumidos en el modelo de regresión lineal simple se refieren a:

1. Tener distribución normal.
2. Tener media cero.
3. Tener varianza constante.
4. Ser estadísticamente independientes.

Se sabe además que los supuestos 1-4 se pueden resumir como:

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

Las posibles desviaciones del modelo en relación con los supuestos en (1) pueden ser estudiadas a través de los residuales, $e_i = y_i - \hat{y}_i$, que son "seudo" estimaciones de los errores del modelo.

Los errores del modelo tienen media cero

Usando los residuales del modelo:

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) \\
&= \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right] \\
&= \sum_{i=1}^{n} (y_i - \bar{y}) - \hat{\beta}_1 \sum_{i=1}^{n} (x_i - \bar{x}) \\
&= 0.
\end{aligned}$$

Por lo tanto, se puede afirmar que 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 del gráfico de residuales vs. valores ajustados o vs. predictor. Si la función de regresión es adecuada, las gráficas de residuos vs. x, y residuos vs. $\hat{y}$, deben mostrar una dispersión homogénea alrededor de la recta horizontal en cero, que representa la media de los errores del modelo.

Otra forma de evaluar el supuesto es mediante un test de homogeneidad de varianza. Para el modelo de regresión lineal simple, existe el test de Levene modificado también conocido como el test Brown-Forsythe, el cual no depende del supuesto de normalidad; es aplicable cuando la varianza se incrementa o disminuye con X, y además se cuenta con un tamaño de muestra suficientemente grande para que las dependencias entre los residuales pueda ser ignorada.

Gráficos con prototipos de residuales cuando la varianza es constante

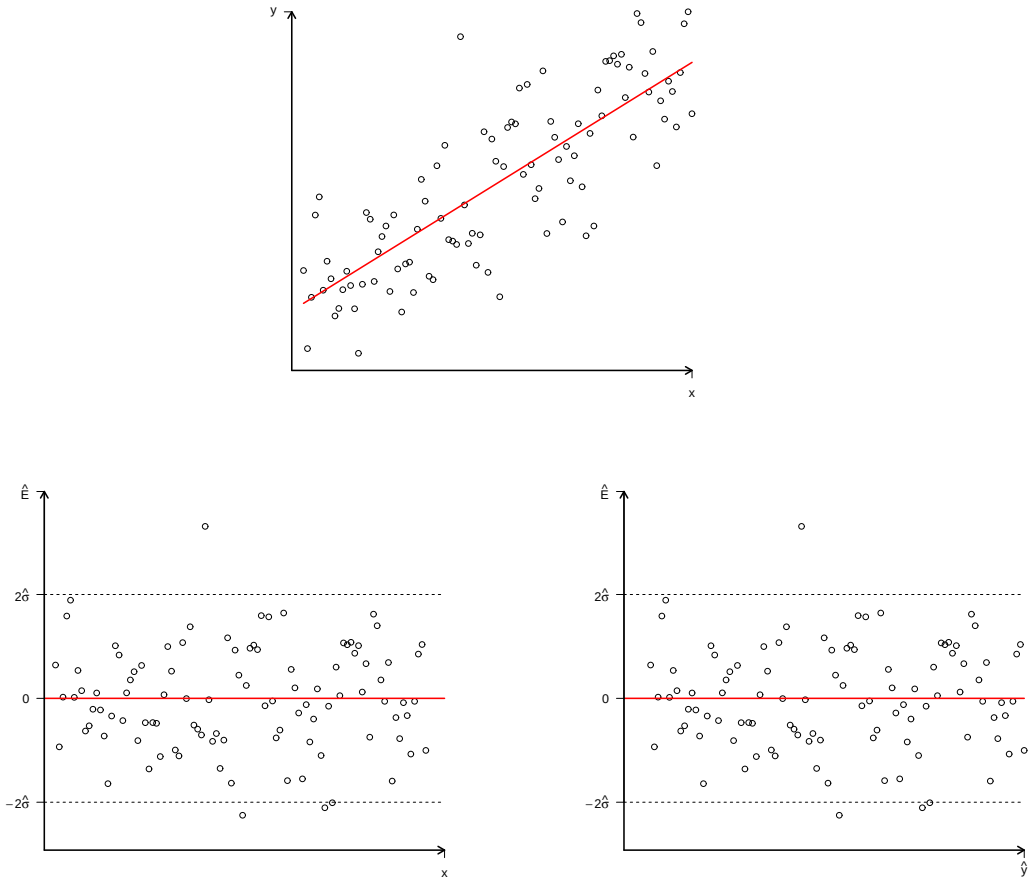


Figura 1: Ilustración de un modelo lineal adecuado con varianza constante: Arr. Gráfico de dispersión; Aba.Izq. Gráfico de residuos vs. x; Aba.Der. Gráfico de residuos vs. $\hat{y}$.

Gráficos con prototipos de residuales cuando la varianza no es constante

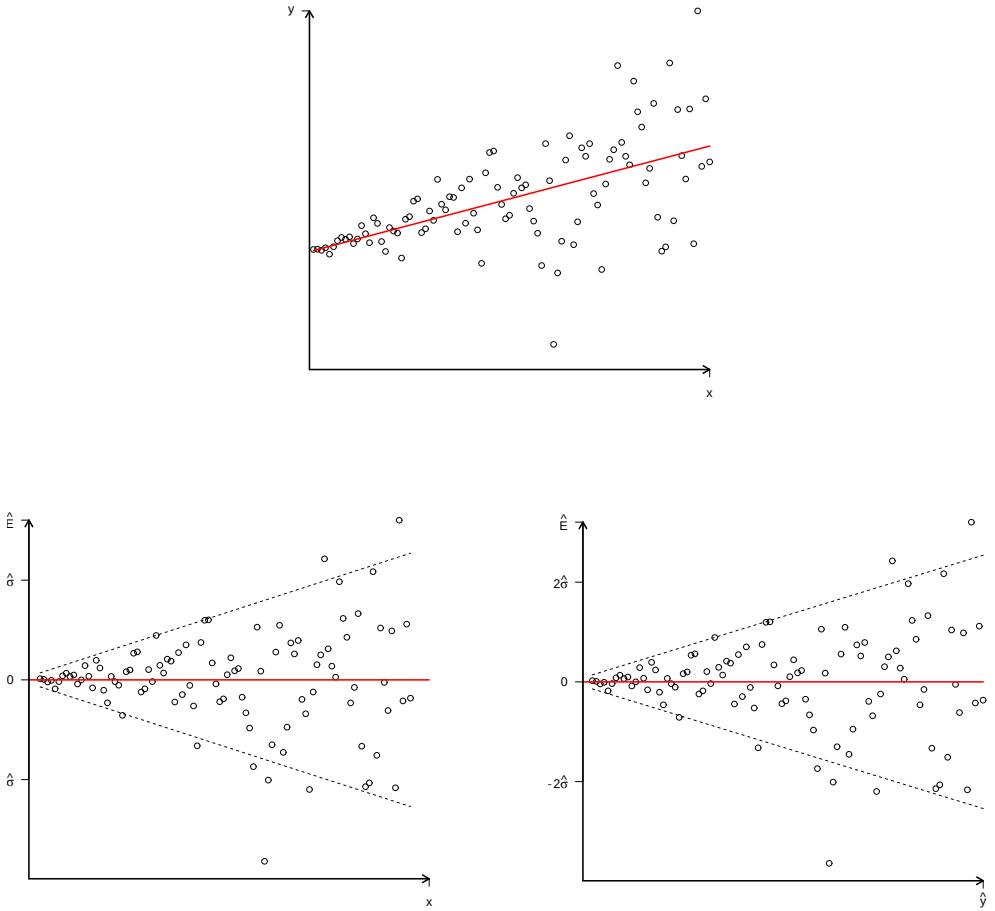


Figura 2: Ejemplos de patrones donde el modelo lineal es correcto pero la varianza no es constante. Patrón de embudo: Arr. Gráfico de dispersión con recta ajustada; Aba.Izq. Residuos vs. x; Aba.Der. Residuos vs. $\hat{y}$.

Gráficos con prototipos de residuales cuando la varianza no es constante

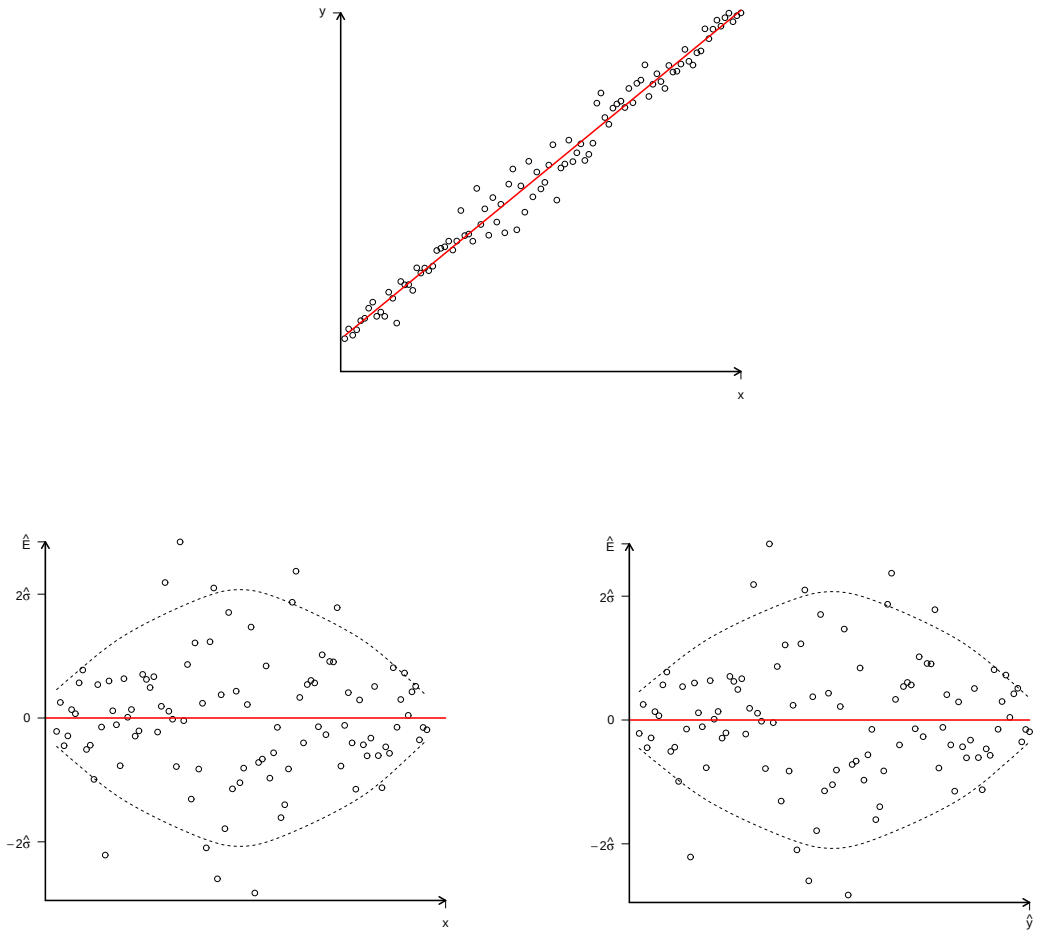


Figura 3: Ejemplos de patrones donde el modelo lineal es correcto pero la varianza no es constante. Patrón de balón de fútbol americano: Arr. Gráfico de dispersión con recta ajustada; Aba.Izq. Residuos vs. x; Aba.Der. Residuos vs. $\hat{y}$.

Soluciones al problema de "no homogeneidad de varianza"

1. Ajuste por mínimos cuadrados ponderados cuando la varianza del error varía de forma sistemática. En este método de ajuste, en la función objetivo de sumas de cuadrados de los errores del modelo, cada uno de ellos es multiplicado por un peso o factor de ponderación ω_i, el cual es inversamente proporcional a la varianza de Y_i. Particularmente, para el modelo de regresión lineal simple, la función de suma de cuadrados a minimizar es de la forma $S(\beta_0, \beta_1) = \sum_{i=1}^{n} \omega_i \varepsilon_i^2 = \sum_{i=1}^{n} \omega_i (Y_i - \beta_0 - \beta_1 x_i)^2$.
2. Otra posibilidad, es usar transformaciones sobre Y que estabilicen la varianza. En algunos casos, la asimetría y la varianza del error se incrementan con la respuesta media. Cuando la transformación apropiada es logarítmica, a veces es necesario sumar una constante a los valores de Y, específicamente cuando existen valores negativos. Se debe tener en cuenta también que cuando la varianza no es constante pero la relación de regresión es lineal, no es suficiente transformar a Y, pues en ese caso aunque se estabilice la varianza, también cambiará la relación lineal a una curvilínea y por ende, se requerirá también una transformación en X; sin embargo, este caso puede manejarse también usando mínimos cuadrados ponderados.

Los errores del modelo son independientes

Para evaluar si hay evidencia en contra de la independencia, es necesario conocer el orden de las observaciones en el tiempo. Sea t el índice que indica el orden de observación en el tiempo. En tal caso, podemos analizar el supuesto a través del gráfico de residuales vs. el tiempo u orden de recolección de los datos. Buscamos patrones sistemáticos como ciclos, rachas, y cualquier otro comportamiento que indique correlación entre los valores de la serie o secuencia de los residuales.

También existen pruebas para incorrelación como la prueba de Durbin Watson para autocorrelación de orden 1. Por lo que, esta prueba sólo detecta correlación entre observaciones sucesivas.

Notas:

- Incorrelación no implica independencia estadística, pero independencia estadística implica incorrelación, sin embargo, si el par de variables incorrelacionadas se distribuyen conjuntamente en forma normal, entonces son independientes.
- Si no conocemos el orden en que fueron tomadas las observaciones, no podemos aplicar Durbin Watson ni cualquier otra prueba de incorrelación y asumimos como válido el supuesto de independencia.

Soluciones al problema de "no independencia de los errores"

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

Los errores del modelo se distribuyen normal

El supuesto de normalidad puede evaluarse desde que sea válido que el conjunto de observaciones sobre el cual se construirá los resultados para la prueba, provienen de una muestra aleatoria, esta última condición implica que es requerido constatar como mínimo la incorrelación antes del test de normalidad. El test de normalidad se realiza sobre los errores del modelo de regresión:

$$H_0 : \varepsilon_i \sim N\left(0, \sigma^2\right) \quad \text{vs.} \quad H_1 : \varepsilon_i \nsim N\left(0, \sigma^2\right). \tag{2}$$

La prueba puede realizarse bien sea examinando los valores P arrojados por una prueba específica de normalidad, como el test de Shapiro Wilk, o bien, mediante un gráfico de normalidad en cual se evalúa si la nube de puntos en la escala normal se puede ajustar por la línea recta del modelo de los cuantiles normales. Si los errores son normales, se espera que con la gráfica de probabilidad normal obtenida con los residuos de ajuste, no se encuentre desviaciones significativas con respecto a la recta de probabilidad normal, y además que cualquier test de bondad de ajuste no rechace el supuesto de normalidad.

Nota:

Para una prueba de bondad de ajuste, rechace $H_0 : \varepsilon_i \sim N\left(0, \sigma^2\right)$ si el valor P es pequeño.

Patrón correcto en gráfico de probabilidad normal

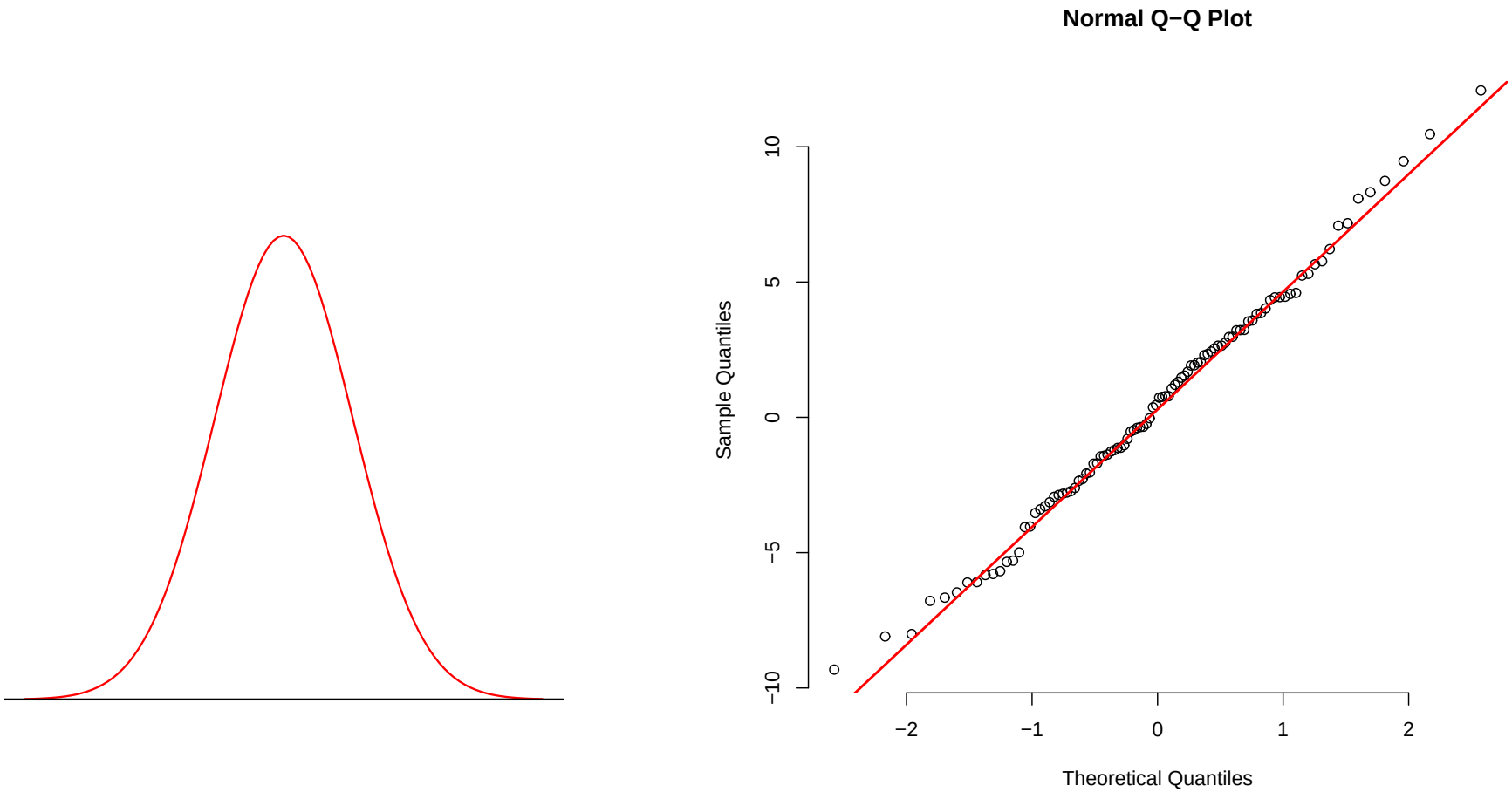
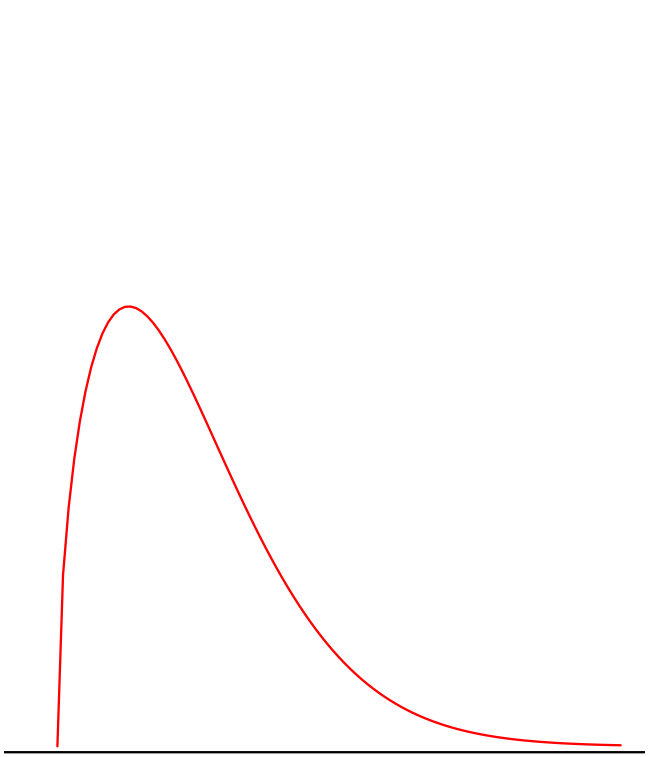


Figura 4: Izq. Densidad de la distribución poblacional (distribución normal de media cero). Der. Patrón en gráficos de probabilidad normal.

Patrón con desvío de la normalidad en gráfico de probabilidad normal



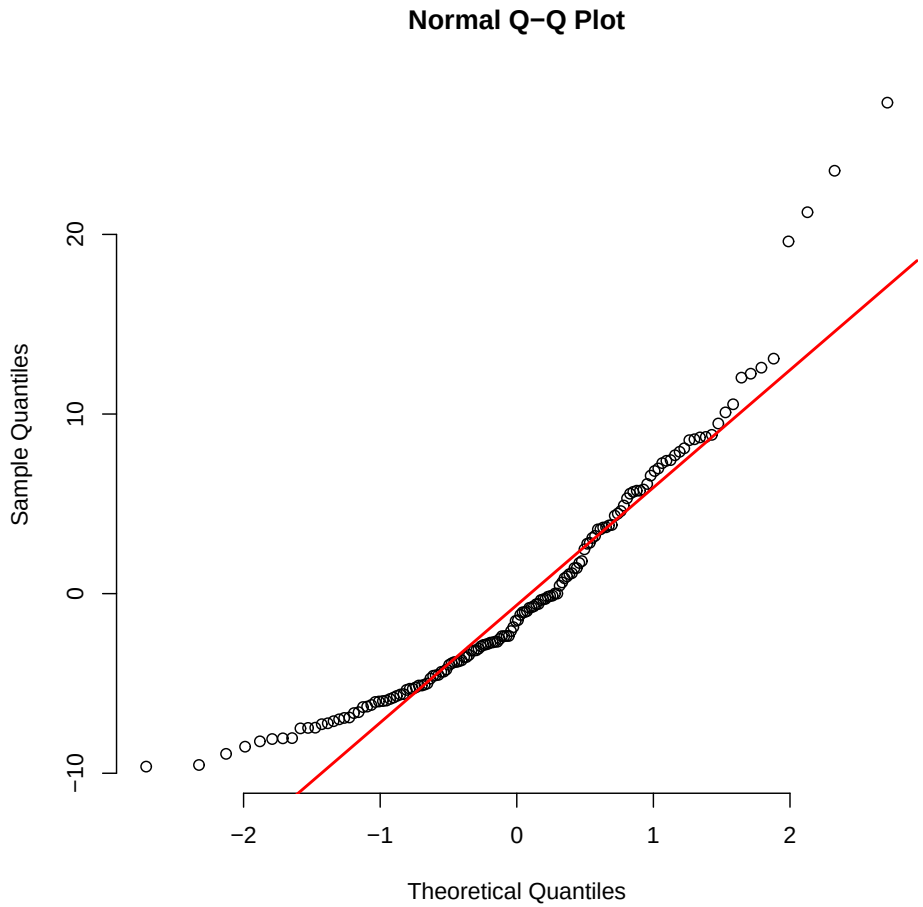
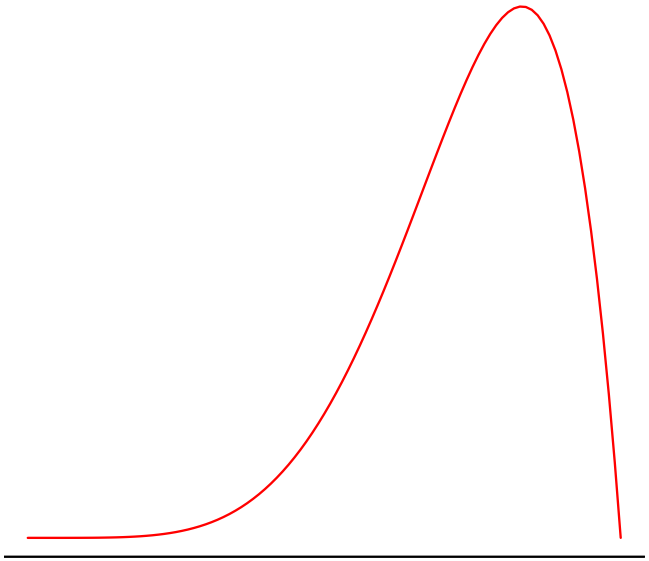


Figura 5: Izq. Densidad de la distribución poblacional (distribución no normal y asimétrica a derecha). Der. Patrón en gráficos de probabilidad normal.

Patrón con desvío de la normalidad en gráfico de probabilidad normal



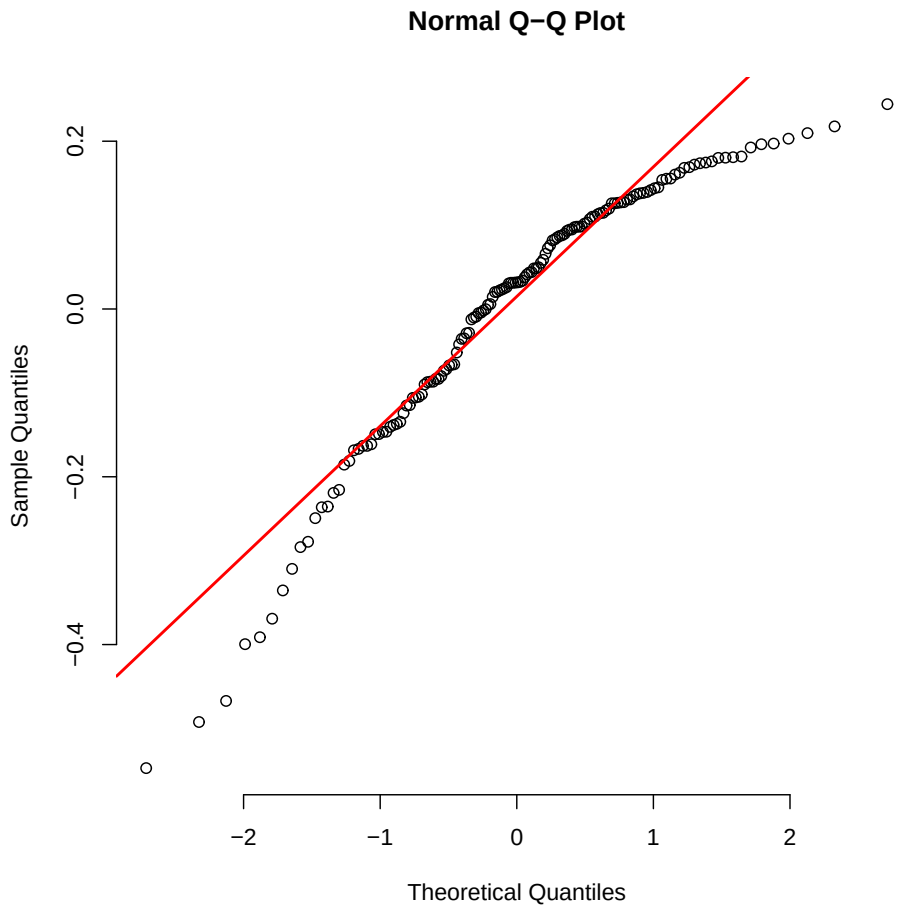


Figura 6: Izq. Densidad de la distribución poblacional (distribución no normal asimétrica a izquierda). Der. Patrón en gráficos de probabilidad normal.

Patrón con desvío de la normalidad en gráfico de probabilidad normal

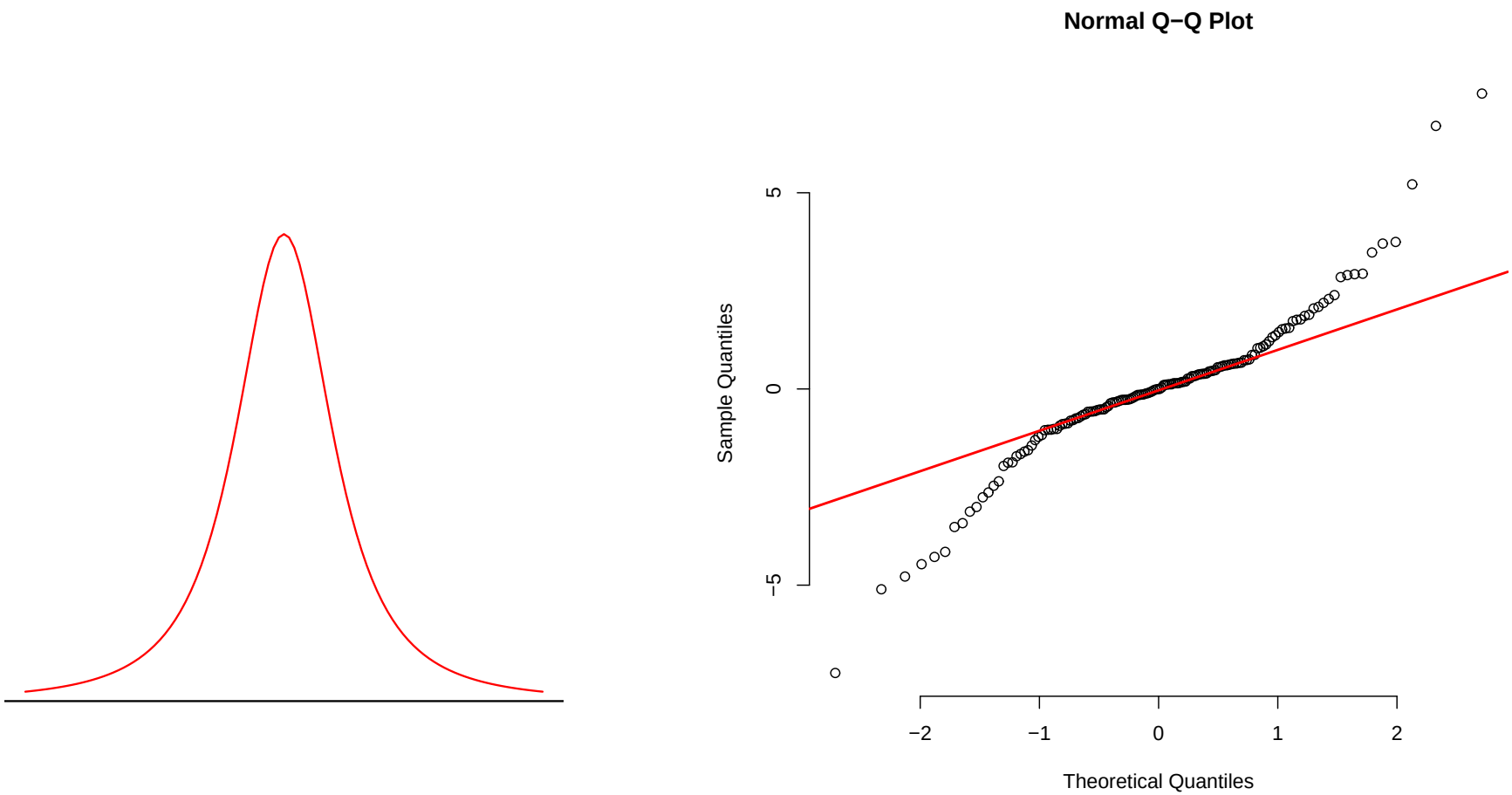


Figura 7: Izq. Densidad de la distribución poblacional (distribución no normal, simétrica pero de colas pesadas). Der. Patrón en gráficos de probabilidad normal.

Soluciones a no normalidad

1. La no normalidad frecuentemente va de la mano con la no homogeneidad de la varianza, por ello, a menudo una 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. Entre las transformaciones que logran corregir la no normalidad junto con varianza no constante, se tienen las transformaciones de potencia Box-Cox: Y^{λ}, que comprende la transformación de logaritmo natural (caso $\lambda = 0$).
2. Otra solución es trabajar con métodos de regresión no paramétricos.

Contenido

1. Validación de supuestos sobre los errores
2. Supuesto de linealidad del modelo: prueba de falta de ajuste
3. Referencias

Prueba de carencia o falta de ajuste

El modelo de regresión lineal simple asume implícitamente que el modelo real de regresión entre la variable respuesta y la variable predictora es lineal en los parámetros del modelo. La violación de este supuesto puede identificarse gráficamente a través del gráfico de residuales vs. x o del gráfico de residuales vs. $\hat{y}$, de manera que cuando ocurre esta violación, los gráficos presentan patrones en los cuales 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 de cualquier tipo de tendencia.

Gráficos con formas sistemáticas

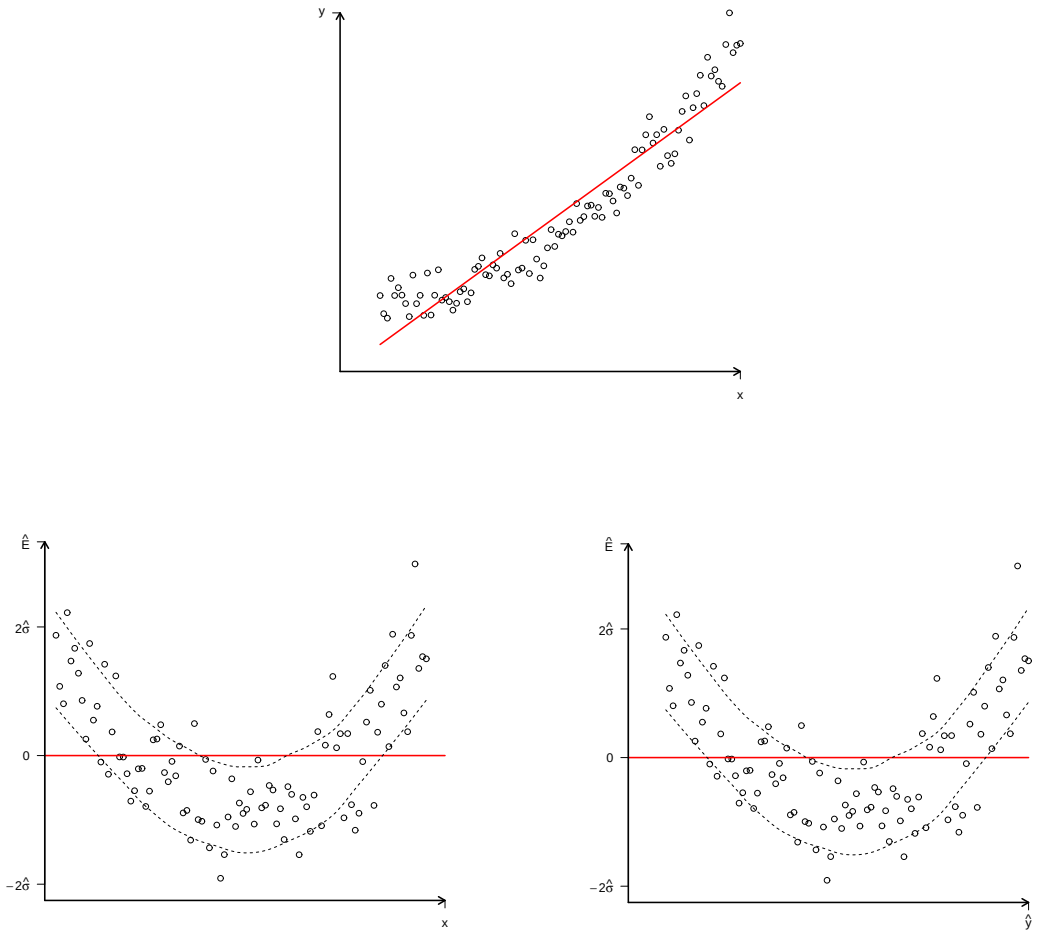


Figura 8: Ejemplo del caso donde el modelo lineal entre y y x no es adecuado, pero la varianza es constante. Arr. Gráfico de dispersión con recta ajustada. Aba.Izq. residuos vs. x. Aba.Der. residuos vs. $\hat{y}$.

Gráficos con formas sistemáticas

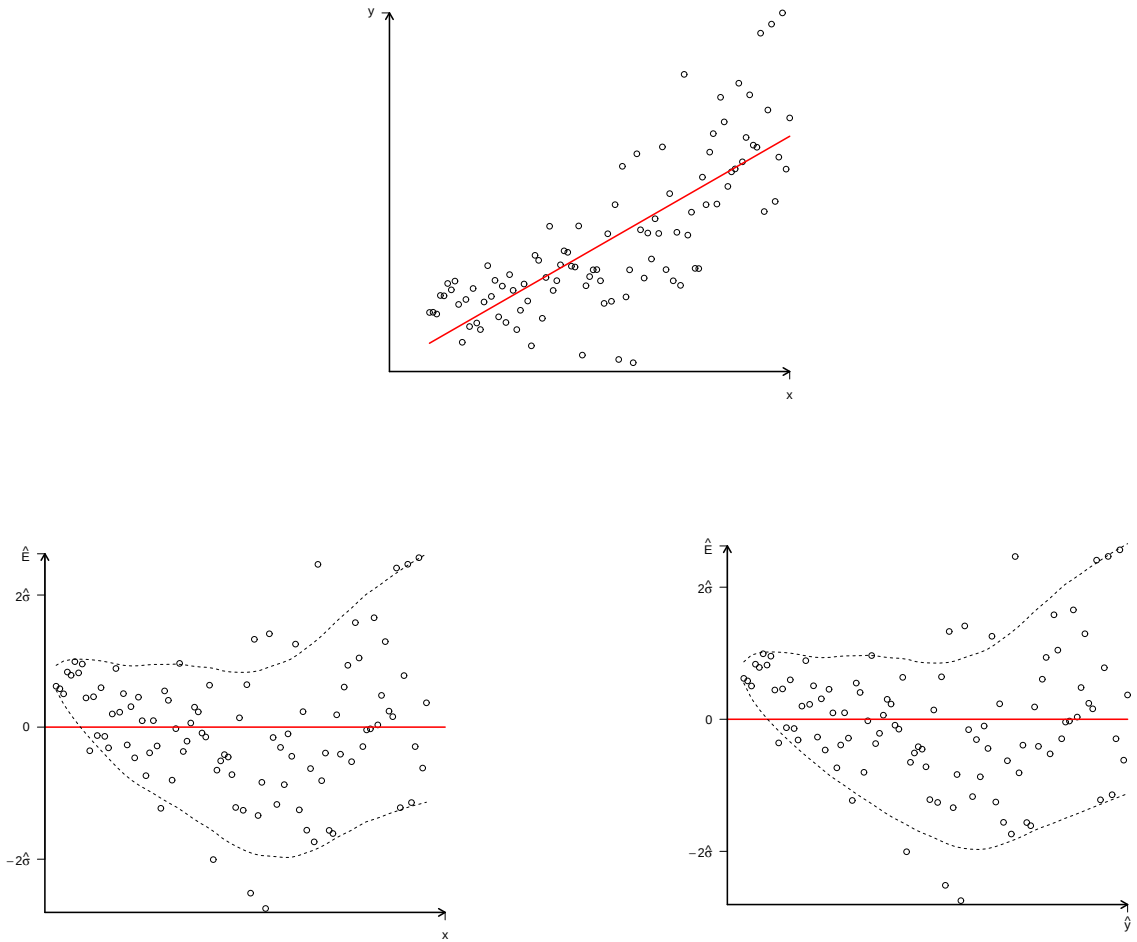


Figura 9: Ejemplo del caso donde el modelo lineal entre y y x no es adecuado, ni la varianza es constante Arr. Gráfico de dispersión con recta ajustada. Aba.Izq. residuos vs. x. Aba.Der. residuos vs. $\hat{y}$.

Prueba de carencia o falta de ajuste

Otra forma de probar la no linealidad del modelo, es mediante el test de falta o carencia de ajuste. Este test prueba que un tipo específico de función de regresión ajusta adecuadamente a los datos. El test asume que los valores de Y dado X son:

- independientes,
- se distribuyen en forma normal,
- tienen varianza constante.

Esta prueba requiere que en uno o más valores de X haya más de una observación de Y. Los ensayos repetidos de manera independiente para el mismo nivel de la variable predictora son denominados *réplicas*.

Prueba de carencia o falta de ajuste

Para explicar en qué consiste la prueba, es necesario modificar la notación usada de la siguiente manera, asumiendo que tenemos réplicas de la respuesta en un valor o nivel dado de X:

- Y_{ij}, la respuesta j-ésima en el i-ésimo nivel de X;
- x_i, i-ésimo nivel de X; supondremos $i = 1, 2, \ldots, k$;
- n_i, número de observaciones de Y tomadas en el i-ésimo nivel de X, Por tanto, el total de observaciones n corresponde a

$$n = \sum_{i=1}^{k} n_i.$$

Prueba de carencia o falta de ajuste

El test define el modelo denominado *modelo lineal general* que corresponde a

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \text{ con } \varepsilon_{ij} \overset{iid}{\sim} N\left(0, \sigma^2\right), \forall\, i, j,\ i = 1, \ldots, k,\ 1, \ldots, n_i, \quad (3)$$

donde $\mu_i = E\left(Y_{ij}\right)$, es decir, es la media de la variable respuesta en el i-ésimo nivel de X. Para este modelo, los estimadores de máxima verosimilitud corresponden a $\hat{\mu}_i = \bar{Y}_{i\bullet}$, donde $\hat{\mu}_i = \bar{Y}_{i\bullet}$ es la media muestral de Y en el nivel i de X, es decir, $\hat{Y}_{ij} = \bar{Y}_{i\bullet}$. La suma de cuadrados de los residuos del modelo lineal general es denominada *suma de cuadrados de error puro* y es denotada por $SSPE$,

$$SSPE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\bullet}\right)^2, \quad (4)$$

cuyos grados de libertad son $n - k = \sum_{i=1}^{k} n_i - k$.

Prueba de carencia o falta de ajuste

Se define el modelo lineal de regresión $\beta_0 + \beta_1 X$ como el modelo de la respuesta media bajo la hipótesis nula de la prueba. El test prueba para $\mu_i = E(Y \mid X = x_i)$ que,

$$H_0 : \mu_i = \beta_0 + \beta_1 x_i \quad \text{vs.} \quad H_1 : \mu_i \neq \beta_0 + \beta_1 x_i. \tag{5}$$

Es decir, H_0 postula que μ_i está relacionado linealmente con x_i mediante la ecuación $\mu_i = \beta_0 + \beta_1 x_i$. Por tanto, el modelo reducido, es decir, bajo H_0, para Y_{ij} es el modelo de regresión lineal para el cual la suma de cuadrados de los residuos es el SSE, con

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{Y}_{ij} \right)^2, \text{ con } \hat{Y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{6}$$

Los grados de libertad de SSE corresponden a $n - 2 = \sum_{i=1}^{k} n_i - 2$.

Prueba de carencia o falta de ajuste

Todas las observaciones de Y en el mismo nivel i de X tienen igual valor ajustado el cual podemos denotar por $\hat{Y}_{i\bullet}$, de ahí que se pueda escribir la ecuación del SSE como,

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(Y_{ij} - \hat{Y}_{i\bullet} \right)^2. \tag{7}$$

Bajo los supuestos sobre el error y la hipótesis nula $H_0 : \mu_i = \beta_0 + \beta_1 x_i$, el SSE de la regresión se descompone de la siguiente manera:

$$SSE = SSPE + SSLOF, \tag{8}$$

donde $SSLOF$ es la *suma de cuadrados de carencia de ajuste*. De la misma manera, se tiene una descomposición de los grados de libertad, así:

$$gl\,(SSE) = gl\,(SSPE) + gl\,(SSLOF)\,. \tag{9}$$

Prueba de carencia o falta de ajuste

Si despejamos a $SSLOF$, obtenemos, $SSLOF = SSE - SSPE$, por tanto,

$$SSLOF = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\bar{Y}_{i\bullet} - \hat{Y}_{i\bullet} \right)^2 = \sum_{i=1}^{k} n_i \left(\bar{Y}_{i\bullet} - \hat{Y}_{i\bullet} \right)^2, \qquad (10)$$

del mismo modo, obtenemos los grados de libertad de $SSLOF$:

$$gl\left(SSLOF\right) = gl\left(SSE\right) - gl\left(SSPE\right) = \left(n - 2\right) - \left(n - k\right) = k - 2.$$

Prueba de carencia o falta de ajuste

El estadístico de prueba para el test $H_0 : \mu_i = \beta_0 + \beta_1 x_i$ vs. $H_1 : \mu_i \neq \beta_0 + \beta_1 x_i$ y su distribución bajo H_0 y validez de los supuestos sobre los errores, es

$$F_0 = \frac{SSLOF/(k-2)}{SSPE/(n-k)} \sim f_{k-2,\,n-k}. \tag{11}$$

El criterio de decisión a un nivel de significancia α es rechazar H_0 si $F_0 > f_{\alpha,\,k-2,\,n-k}$, donde $f_{\alpha,\,k-2,\,n-k}$ es el percentil $(1-\alpha)\,100\,\%$ de la distribución $f_{k-2,\,n-k}$. También puede decidir mediante valor P: Rechazar H_0 si $P\left(f_{k-2,\,n-k} > F_0\right)$ es pequeño.

Nota:

Al rechazar H_0, se concluye que el modelo de regresión no es lineal en X. Observe además que para realizar la prueba de carencia de ajuste se requieren más de dos niveles de valores en X, es decir, que $k > 2$, y también es necesario que $k < n$, ya que la distribución $f_{k-2,\,n-k}$ debe tener grados de libertad no nulos en el numerador y denominador.

Prueba de carencia o falta de ajuste

El test de carencia de ajuste se realiza mediante un test ANOVA, que se resume en una tabla (ver Tabla 1).

Tabla 1: ANOVA para modelo de regresión y carencia de ajuste.

Fuente	Suma de cuadrados	Grados de libertad	Cuadrados medios	F calculada
Regresión	SSR	1	$MSR = SSR/1$	$F_0 = MSR/MSE$
Error	SSE	$n-2$	$MSE = SSE/(n-2)$	
Carencia de ajuste	$SSLOF$	$k-2$	$MSLOF = SSLOF/(k-2)$	$F_0 = MSLOF/MSPE$
Error puro	$SSPE$	$n-k$	$MSPE = SSPE/(n-k)$	
Total	SST	$n-1$	$MST = SST/(n-1)$	

Consideraciones del análisis de varianza

Notas sobre el ANOVA:

- En general, en el cálculo del $SSPE$ sólo se utilizan aquellos niveles x_i de X en los cuales hay replicas.
- La prueba de carencia de ajuste puede aplicarse con cualquier función de regresión lineal, es decir, diferentes a la del modelo de regresión lineal simple, sólo se requiere modificar los grados de libertad del $SSLOF$, que en general corresponden a $k - p$, donde p es el número de parámetros en la función de regresión propuesta. Para el caso específico de la regresión lineal simple, $p = 2$.
- Sin importar cuál sea la verdadera función de regresión considerada, se cumple que $E(MSPE) = \sigma^2$.
- $E(MSLOF) = \sigma^2$ sólo si la función de regresión propuesta es correcta, de lo contrario, $E(MSLOF) > \sigma^2$, específicamante, $E(MSLOF) = \sigma^2 + \frac{\sum_{i=1}^{k} n_i[\mu_i - (\beta_0 + \beta_1 x_i)]^2}{k-2}$.

Consideraciones del análisis de varianza

Notas sobre el ANOVA:

- Cuando se concluye que es apropiado la función de regresión propuesta (es decir, cuando no hay carencia de ajuste), la práctica usual es usar el *MSE* de la regresión 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, sólo debe llevarse a cabo luego de haber probado que el modelo de regresión lineal es apropiado.
- *SSLOF* representa la falta de ajuste entre el modelo general y el modelo lineal propuesto, esto es, un *SSLOF* pequeño indica que los dos modelos son muy cercanos, esto es, el modelo propuesto realmente es lineal, pero un *SSLOF* grande indica que hay diferencias significativas entre estos modelos, lo que implica que el modelo propuesto realmente no es lineal.

Soluciones al problema "el modelo de regresión lineal no es apropiado"

1. Abandonar el modelo de regresión lineal y desarrollar un modelo más apropiado.
2. Emplear alguna transformación en los datos de manera que el modelo de regresión lineal sea apropiado en los datos transformados.
3. 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.

Contenido

1 Validación de supuestos sobre los errores

2 Supuesto de linealidad del modelo: prueba de falta de ajuste

3 Referencias

Referencias

Kutner, M. H., Nachtsheim, C. J., Neter, J., y Li, W. (2005). *Applied Linear Statistical Models*. McGraw-Hill Irwin, New York, quinta edición.

Álvarez, N. G. (2022). Notas de Clase Análisis de Regresión - 3006918, Capítulo 2: Regresión Lineal Simple. Notas no publicadas.

Álvarez, N. G. y Gómez, C. M. L. (2018). Notas de Clase - Estadística II (3006918): Análisis de Regresión Lineal e Introducción al Muestreo.