Regresión lineal múltiple

José Gabriel Palomo Sánchez
gabriel.palomo@upm.es

E.U.A.T.
U.P.M.

Julio de 2011

Índice I

Índice II

El modelo de regresión múltiple I

- El modelo de regresión múltiple es la extensión a k variables explicativas del modelo de regresión simple.

- La estructura del modelo de regresión múltiple es la siguiente:

$$y = f(x_1, \ldots, x_k) + E.$$

El modelo de regresión múltiple II

Donde:

- y es la variable explicada, dependiente o respuesta.
- $x_1, \ldots, x_k$ son las variables explicativas, regresores o variables independientes.
- $y = f(x_1, \ldots, x_k)$ es la parte determinista del modelo.
- E representa el error aleatorio. Contiene el efecto sobre y de todas las variables distintas de $x_1, \ldots, x_k$.

El modelo de regresión lineal múltiple I

- El modelo de regresión lineal múltiple tiene la forma:

$$\underbrace{y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{Hiperplano}} + E$$

El Modelo de regresión lineal múltiple II

- El modelo de regresión lineal mútiple se utiliza cuando:

1. La variable dependiente, Y, depende linealmente de cada una de las variables explicativas, $X_1, \ldots, X_k$.
2. Un regresor no basta para explicar suficientemente la variabilidad de Y.

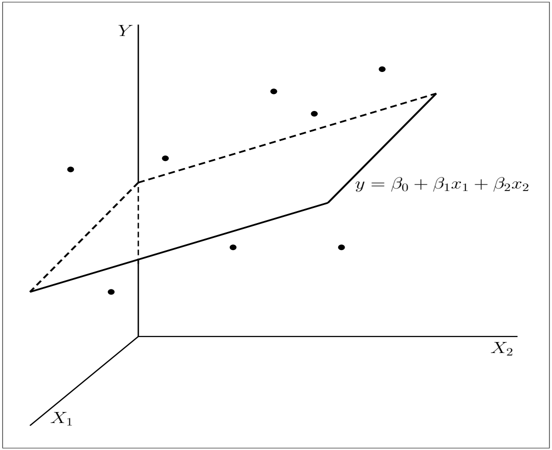
Ejemplo: El modelo de regresión lineal con dos regresores I

- En el caso particular en que haya dos regresores, $k = 2$, el modelo tendría la forma:

$$\underbrace{y = \beta_0 + \beta_1 x_1 + \beta_2 x_2}_{\text{Plano}} + E$$

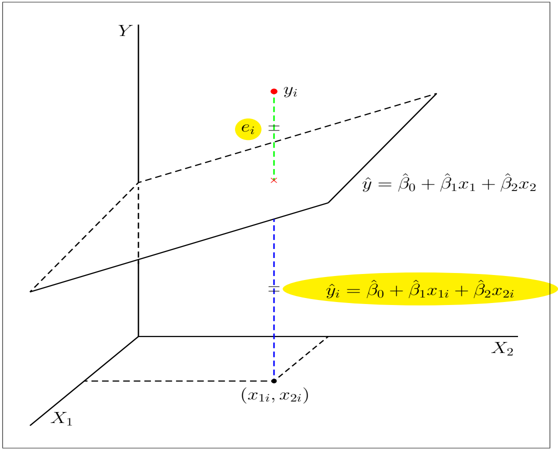
Ejemplo: El modelo de regresión lineal con dos regresores II

Gráficamente, el modelo de regresión lineal con dos regresores supone calcular la ecuación de un plano que describa la relación de Y con X_1 y X_2,



Ejemplo: El modelo de regresión lineal con dos regresores III

De forma análoga al caso simple, calculado el plano, el valor de y_i en cualquier punto (x_{1i}, x_{2i}, y_i) se puede descomponer en parte determinista y aleatoria,

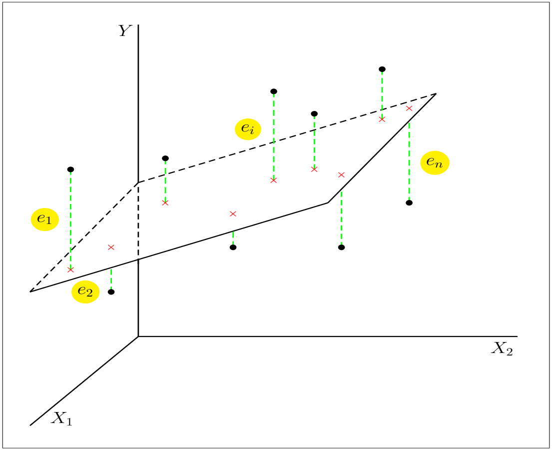


Ejemplo: El modelo de regresión lineal con dos regresores IV

La estimación por mínimos cuadrados de los parámetros del modelo consiste en calcular la ecuación del plano que haga mínimo el valor de

$$\sum e_i^2,$$

con $e_i = y_i - \hat{y}_i$.



El modelo general de regresión. Hipótesis del modelo I

Generalizando, al ajustar un modelo de regresión lineal múltiple se supondrá que se verifican las siguientes hipótesis:

1. Fijados los valores $x_{1i}, \ldots, x_{ki}$ de las variables $X_1, \ldots, X_k$, se tiene que $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + e_i$
2. Cada error $e_i \approx N\left(0, \sigma^2\right)$.
3. Cualquier par de errores e_i y e_j son independientes.
4. Las variables explicativas son, algebraicamente, linealmente independientes.
5. El número de datos es mayor o igual que $k + 2$.

El modelo general de regresión. Hipótesis del modelo II

Observaciones

- Las tres primeras hipótesis del modelo se justifican igual que en regresión simple.
- La condición de la independencia lineal algebraica de los regresores tiene por objeto ajustar la dimensión del problema, ya que si no se cumpliese se podrían eliminar regresores del modelo.
- El número de datos debe ser mayor o igual que $k + 2$ para poder estimar todos los parámetros del modelo.

Consecuencias de las hipótesis del modelo I

Las hipótesis impuestas al comportamiento del error del modelo tienen las siguientes consecuencias:

1. Para cada conjunto de valores, $x_{1i}, \ldots, x_{ki}$ de $X_1, \ldots, X_k$, la variable aleatoria $(Y|X_1 = x_{1i}, \ldots, X_k = x_{ki})$ tiene una distribución:

$$(Y|X_1 = x_{1i}, \ldots, X_k = x_{ki}) \approx N(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}, \sigma^2)$$

2. Las observaciones y_i de la variable Y son independientes.

Consecuencias de las hipótesis del modelo II

- Consecuentemente,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$$

representa la esperanza de la variable Y condicionada por los valores $x_{1i}, \ldots x_{ki}$ de las variables $X_1, \ldots, X_k$, respectivamente.

- Además, todas las variables Y_i tienen la misma varianza, σ^2. Es decir, son homocedásticas.

Estimación de los parámetros del modelo de regresión múltiple por mínimos cuadrados I

Supóngase que para estimar los parámetros del modelo

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

se dispone del conjunto de datos:

	X_1	$\cdots$	X_k	Y
Individuo 1	x_{11}	$\cdots$	x_{k1}	y_1
Individuo 2	x_{12}	$\cdots$	x_{k2}	y_2
	$\vdots$			$\vdots$
Individuo n	x_{1n}	$\cdots$	x_{kn}	y_n

Estimación de los parámetros del modelo de regresión múltiple por mínimos cuadrados II

Como

$$e_i = y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}),$$

resulta que el módulo del vector error es función de $\beta_0, \ldots, \beta_k$:

$$\sum e_i^2 = S(\beta_0, \ldots, \beta_k).$$

Estimación de los parámetros del modelo de regresión múltiple por mínimos cuadrados III

Para que S sea mínimo deberá ser:

$$\left.\begin{array}{c} \dfrac{\partial S}{\partial \beta_0} = 0 \\ \\ \dfrac{\partial S}{\partial \beta_1} = 0 \\ \vdots \\ \dfrac{\partial S}{\partial \beta_k} = 0 \end{array}\right\}$$

Estimación de los parámetros del modelo de regresión múltiple por mínimos cuadrados IV

Llamando

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix} \text{ e } Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

Estimación de los parámetros del modelo de regresión múltiple por mínimos cuadrados V

Resulta que:

$$\hat{\beta} = (X'X)^{-1}X'Y,$$

donde X' representa la matriz transpuesta de X.

Propiedades de los estimadores de los parámetros del modelo de regresión I

1. El parámetro $\hat{\beta}_i$, en regresión múltiple, representa el efecto del aumento de una unidad del regresor X_i sobre la respuesta, Y, cuando el resto de los regresores permanecen constantes.

2. Si los regresores están incorrelados, $\rho_{ij} = 0$, para todo i, j, los estimadores de los coeficientes de regresión estimados en el modelo múltiple y en los distintos modelos simples coinciden.

Propiedades de los estimadores de los parámetros del modelo de regresión II

Se puede demostrar que:

1. $\hat{\beta}_i$ sigue una distribución normal, para todo $i = 0, \ldots, k$.

2. Para todo $\hat{\beta}_i$, con $i = 0, 1, \ldots, k$, se cumple que $E(\hat{\beta}_i) = \beta_i$. Es decir $\hat{\beta}_i$ es un estimador centrado de β_i, para todo i.

3. La matriz de varianzas y covarianzas de $\hat{\beta}_0, \ldots, \hat{\beta}_k$ viene dada por la expresión:

$$COV(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Estimador de la varianza del error. La varianza residual I

- Análogamente al caso de regresión simple, la realización de inferencia sobre los parámetros del modelo requiere una estimación de σ^2.
- Como en el caso simple, el estimador máximo verosímil de la varianza es:

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n},$$

que no es un estimador centrado de σ^2.

Estimador de la varianza del error. La varianza residual II

La resolución del sistema,

$$\frac{\partial S}{\beta_0} = 0, \ldots, \frac{\partial S}{\beta_k} = 0,$$

que se emplea para calcular los estimadores de los parámetros β_i, pone de manifiesto las siguientes relaciones entre los residuos:

$$\left.\begin{array}{r} \sum e_i = 0 \\ \sum e_i x_{1i} = 0 \\ \vdots \\ \sum e_i x_{ki} = 0 \end{array}\right\} (k+1) \text{ restricciones.}$$

Estimador de la varianza del error. La varianza residual III

- De forma similar al caso simple se define la varianza residual como:

$$\hat{s}_R^2 = \frac{\sum e_i^2}{n-k-1},$$

que será el estimador habitual de σ^2.

Estimador de la varianza del error. La varianza residual IV

1. $\hat{s}_R^2$ es un estimador centrado de σ^2. Esto es:

$$E(\hat{s}_R^2) = \sigma^2.$$

2. Además:

$$\frac{(n-k-1)\hat{s}_R^2}{\sigma^2} \longrightarrow \chi^2_{n-k-1}$$

- Esta distribución permite realizar inferencia respecto de σ^2.

INFERENCIA RESPECTO DE LOS COEFICIENTES DE REGRESIÓN I

- En consecuencia con lo anterior, si la diagonal de la matriz, conocida, $(X'X)^{-1}$ tiene la forma:

$$D(X'X)^{-1} \longrightarrow \begin{pmatrix} d_{00} & & & & & \\ & d_{11} & & & & \\ & & \ddots & & & \\ & & & d_{ii} & & \\ & & & & \ddots & \\ & & & & & d_{kk} \end{pmatrix}$$

- La varianza de $\hat{\beta}_i$ será $\sigma^2 d_{ii}$.

Inferencia respecto de los coeficientes de regresión II

Por lo tanto,

$$\hat{\beta}_i \approx N(\beta i, \sigma\sqrt{d_{ii}}),$$

de donde,

$$\frac{\hat{\beta}_i - \beta i}{\sigma\sqrt{d_{ii}}} \longrightarrow N(0,1),$$

y

$$\frac{\hat{\beta}_i - \beta i}{\hat{s}_R\sqrt{d_{ii}}} \longrightarrow t_{n-k-1}.$$

Inferencia respecto de los coeficientes de regresión III

- La última expresión permite realizar, para todo $i = 0, \ldots k$, el contraste individual de regresión (test de la t):

$$H_0 : \beta_i = 0 \quad \text{frente a} \quad H_1 : \beta_i \neq 0,$$

ya que si $\beta_i = 0$,

$$\frac{\hat{\beta}_i}{\hat{s}_R \sqrt{d_{ii}}} \longrightarrow t_{n-k-1}.$$

Interpretación del test de la t en regresión múltiple

- Análogamente al caso simple, la aceptación de la hipótesis nula, $\beta_i = 0$, puede ser debida a que:
  1. X_i e Y sean independientes.
  2. Entre X_i e Y haya una relación de dependencia no lineal.
- En el primer caso, la variable X_i debe ser eliminada del modelo.
- En el segundo, se debe intentar una transformación que linealice la relación entre X_i e Y.

EL TEST DE LA F EN REGRESIÓN MÚLTIPLE I

Si se denomina $(X'X)_0^{-1}$ a la matriz resultante de eliminar la primera fila y la primera columna de la matriz

$$(X'X)^{-1} = \begin{pmatrix} d_{00} & & & & & \\ & d_{11} & & & & \\ & & \ddots & & & \\ & & & d_{ii} & & \\ & & & & \ddots & \\ & & & & & d_{kk} \end{pmatrix},$$

El test de la F en regresión múltiple II

se tiene que la matriz:

$$\sigma^2 (X'X)_0^{-1} = \sigma^2 \begin{pmatrix} d_{11} & & & & \\ & \ddots & & & \\ & & d_{ii} & & \\ & & & \ddots & \\ & & & & d_{kk} \end{pmatrix}$$

es la matriz de varianzas y covarianzas de $\hat{\beta}_1, \ldots, \hat{\beta}_k$.

El test de la F en regresión múltiple III

Llamando b al vector:

$$b = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix}$$

Se puede demostrar que:

$$\frac{\hat{b}'(X'X)_0\hat{b}}{k\hat{s}_R^2} \longrightarrow F_{(k,n-k-1)}.$$

El test de la F en regresión múltiple IV

- Esta distribución permite la realización del contraste fundamental de regresión múltiple:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

frente a:

$$H_1 : \text{Existe algún } \beta_i \text{ con } i = 1, \ldots, k \text{ tal que } \beta_i \neq 0.$$

Interpretación del test de la F I

- La aceptación de la hipótesis nula del test de la F,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0,$$

puede ser debida a:

  - Independencia de la variable explicada frente a todos los regresores.
  - Existe dependencia no lineal de la variable explicada frente a algún regresor.

Interpretación del test de la F II

- El rechazo de la hipótesis nula del test de la F significa que la variable explicada depende linealmente de alguno de los regresores.
    - Para saber cuál o cuáles de los regresores explican significativamente a la variable dependiente es necesario atender a los contrastes individuales de la t.

Interpretación conjunta del test de la F y de los contrastes individuales de la t I

La siguiente tabla contiene el conjunto de los casos posibles al realizar el test de la F y los contrastes de la t en regresión múltiple:

Caso	Test de la F	Contrastes individuales
1	Significativo	Todos significativos
2	Significativo	Alguno significativo
3	Significativo	Ninguno significativo
4	No significativo	Todos significativos
5	No significativo	Alguno significativo
6	No significativo	Ninguno significativo

INTERPRETACIÓN CONJUNTA DEL TEST DE LA F Y DE LOS CONTRASTES INDIVIDUALES DE LA t II

CASO 1: Cuando el contraste de la F es significativo y todos los contrastes de la t también lo son, se interpreta que todos los regresores influyen significativamente en la variable explicada.

CASO 2: Si el contraste de la F es significativo y sólo algunos de los regresores lo son, se interpreta que los regresores no significativos deben ser eliminados del modelo, o bien transformados si se intuye relación de dependencia no lineal entre la variable dependiente y alguno de ellos.

Interpretación conjunta del test de la F y de los contrastes individuales de la t III

Caso 3: Cuando el test de la F es significativo y ninguno de los contrastes individuales lo es, se da una situación paradójica que, frecuentemente, se origina por un problema denominado **multicolinealidad**. Su análisis y tratamiento se explica más adelante.

Casos 4 y 5: Si el test de la F es no significativo y todos o algunos de los contrastes individuales sí lo son, se origina igualmente una situación paradójica que responden a casos particulares de multicolinealidad.

Caso 6: Si el test de la F no es significativo y ninguno de los contrastes individuales lo es, no se detecta relación de dependencia lineal entre la variable explicada y los regresores.

Multicolinealidad. Detección y tratamiento I

Los casos 3, 4, y 5 citados anteriormente se deben habitualmente al problema de multicolinealidad.

- La multicolinealidad es consecuencia de que todos o una parte de los regresores $X_1, \ldots, X_k$ están fuertemente correlados.
- La detección de la multicolinealidad se realiza a través de:
  - La matriz de correlación de las variables explicativas.
  - La diagonal de la inversa de esta matriz.
  - Los autovalores de la matriz $X'X$.

Multicolinealidad. Detección y tratamiento II

El tratamiento de la multicolinealidad consiste básicamente en:

- Eliminar regresores del modelo que tengan alta correlación con el resto, lo que disminuye el número de parámetros que hay que estimar. (Esta es la solución más sencilla, cuando se puede utilizar.)
- Incluir información externa a los datos.

El coeficiente de determinación I

Una vez estimado el modelo de regresión múltiple,

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k,$$

se puede completar el conjunto de datos con la nueva columna $\hat{Y}$:

	X_1	$\cdots$	X_k	Y	$\hat{Y}$
Individuo 1	x_{11}	$\cdots$	x_{k1}	y_1	$\hat{y}_1$
Individuo 2	x_{12}	$\cdots$	x_{k2}	y_2	$\hat{y}_2$
	$\vdots$				$\vdots$
Individuo n	x_{1n}	$\cdots$	x_{kn}	y_n	$\hat{y}_n$

Donde $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_k x_{ki}$.

EL COEFICIENTE DE DETERMINACIÓN II

- Se define el coeficiente de determinación como el coeficiente de correlación lineal entre las variables Y e $\hat{Y}$.
- El coeficiente de determinación es una medida de bondad del ajuste del modelo y se representa por R^2.
- La eficacia de R^2 como medida de la bondad de ajuste depende de la relación entre el número de regresores, k y el tamaño muestral, n, siendo más fiable cuanto menor sea el cociente k/n.

El coeficiente de determinación III

El coeficiente de determinación tiene las siguientes propiedades:

- $0 \leq R^2 \leq 1$. Cuando $R^2 = 1$, la relación entre la variable explicada y los regresores es exacta.
- $R^2 \times 100$ representa el porcentaje de variabilidad de Y explicada por el modelo. Un valor de $R^2 \times 100$ bajo puede ser debido a la omisión de variables explicativas relevantes en el modelo.
- R^2 aumenta siempre con la introducción de nuevas variables en el modelo. Para evitar este inconveniente se calcula el coeficiente de determinación corregido por grados de libertad. (Ver Peña 2002)
- En regresión simple, el coeficiente de determinación es el cuadrado del coeficiente de correlación lineal.

Diagnosis y validación del modelo I

- Al igual que en el caso de la regresión simple, antes de emplear un modelo de regresión es necesario verificar las hipótesis básicas del modelo.

- Esta verificación (diagnosis) se realiza a través del análisis de los residuos.

Diagnosis y validación del modelo II

En particular:

- La normalidad del error se analiza con la representación de los residuos en papel probabilístico normal, o con algún test de normalidad.
- Las hipótesis de linealidad, homocedasticidad e independencia se verifican a través del gráfico de residuos frente a los valores previstos y frente a los valores de los regresores.
- La conveniencia de introducir una nueva variable en el modelo se puede analizar por medio del gráfico de los residuos frente a esta nueva variable.

La interpretación de los gráficos es similar a la que se da a los mismos en regresión simple.

Predicción en regresión múltiple I

Una vez estimado y validado el modelo de regresión, se puede emplear éste para hacer predicciones.

1. Se puede emplear $\hat{y}(x_{1i}, \ldots, x_{ki})$ para predecir el valor de $E(Y|X_1 = x_{1i}, \ldots, X_k = x_{ki})$.
2. También se puede emplear $\hat{y}(x_{1i}, \ldots, x_{ki})$ para predecir el valor de un individuo de la variable $(Y|X_1 = x_{1i}, \ldots, X_k = x_{ki})$.

Predicción en regresión múltiple II

- Obsérvese que los dos valores se estiman por el mismo número.
- Igual que en regresión simple, la estimación de la media se realiza con mayor precisión que el valor de un individuo concreto.
  Pueden consultarse los detalles del cálculo de intervalos de confianza, para ambos casos, en Peña (2002).

Los valores atípicos en regresión múltiple I

- Conceptualmente las ideas de punto atípico e influyente coinciden con las explicadas en regresión simple.
- La deteccción de puntos atípicos en regresión múltiple es más compleja que en regresión simple, debido a la dimensionalidad de los datos.
- Se emplearán los gráficos de residuos frente a las variables explicativas y a los valores previstos por el modelo para detectar puntos atípicos, aunque estos gráficos no siempre permiten encontrar estos puntos.

Los valores atípicos en regresión múltiple II

- Información sobre la construcción de estadísticos que permitan detectar atípicos puede encontrarse en Peña (2002).
- El tratamiento de los atípicos, una vez identificados, será similar al expuesto en regresión simple.

Ejemplo I. Modelos de regresión linealizables

Se exponen a continuación algunos ejemplos de relaciones no lineales, que se pueden linealizar mediante transformaciones adecuadas.

- Modelo polinómico de segundo grado con una variable independiente:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

- Haciendo la transformación $x_2 = x^2$, se obtiene el modelo lineal:

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + u$$

Ejemplo II. Modelos de regresión linealizables

- Modelo polinómico de segundo grado con dos variables independientes:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + u$$

- Haciendo las transformaciones $x_3 = x_1^2,\ x_4 = x_2^2,\ x_5 = x_1 x_2$, se obtiene el modelo lineal:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_3 + \beta_{22} x_4 + \beta_{12} x_5 + u$$

Ejemplo III. Modelos de regresión linealizables

- $y = \beta_0 + \beta_1 \left(\frac{1}{x_1}\right) + \beta_2 \ln x_2 + \beta_3 \sqrt{x_3} + u$

- Haciendo las transformaciones

  $z_1 = \left(\frac{1}{x_1}\right), \; z_2 = \ln x_2, \; z_3 = \sqrt{x_3}$, se obtiene el modelo lineal:

$$y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + u$$

Ejemplo IV. Modelos de regresión linealizables

- Modelo multiplicativo:

$$y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta} \epsilon,$$

donde ϵ representa el error aleatorio.

- Tomando logaritmos en ambos miembros, se tiene que:

$$\ln y = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2 + \delta \ln x_3 + ln\epsilon,$$

que es el modelo lineal:

$$Y = \beta_0 + \beta z_1 + \gamma z_2 + \delta z_3 + u$$

Ejemplos V. Modelos de regresión no linealizables

Por otro lado, cabe señalar que existen modelos de relación que no se pueden linealizar mediante funciones elementales, como por ejemplo:

- $$y = \beta_0 + \beta_1 e^{-\beta_2 X} + u$$
- $$y = \beta_0 + \beta_1 X + \beta_2 (\beta_3)^X + u$$