Regresión múltiple con datos de corte transversal.

Miguel Ángel Tarancón Morán & Consolación Quintana Rojo

Área de Estadística Económica y Empresarial

Departamento de Economía Aplicada 1

Universidad de Castilla – La Mancha

Contenidos.

1. Introducción.
2. Especificación.
3. Estimación.
4. Contraste y validación.
5. Contraste de hipótesis del modelo lineal.
6. Utilización del modelo.

Introducción.

- El **análisis de regresión múltiple** es una de las técnicas de análisis de **dependencias** más profusamente utilizadas.

- En el modelo de regresión múltiple la **variable dependiente** tiene escala **métrica**. Las variables explicativas pueden ser métricas o ser atributos.

- Según los datos de los que se alimenta el modelo, se aplicarán diferentes métodos de estimación, especificaciones y pruebas:

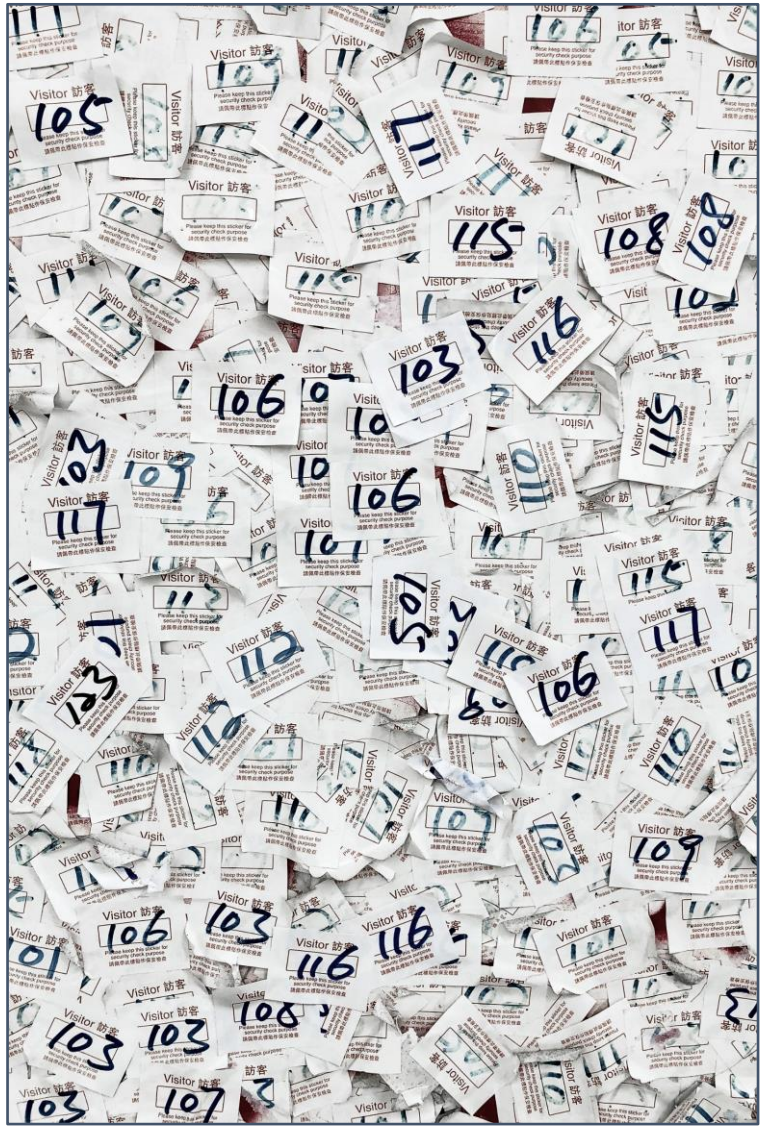
  - Series temporales.
  - Datos de corte transversal.
  - Paneles de datos.

- Nos centraremos en **datos de corte transversal** (cada caso es una empresa).

Introducción.

- La **construcción de un modelo de regresión** cuenta con una serie de etapas:
  - **Especificación** del modelo: establecer las variables que entrarán a formar parte del modelo (dependiente, explicativas).
  - **Estimación:** calcular el valor de los parámetros o coeficientes estructurales del modelo.
  - **Contraste y validación:** verificar si el modelo estimado cumple con las hipótesis que garantizan unas buenas propiedades y si es adecuado para representar la realidad.
  - **Utilización del modelo**: a efectos de previsión, análisis estructural o simulación de escenarios.

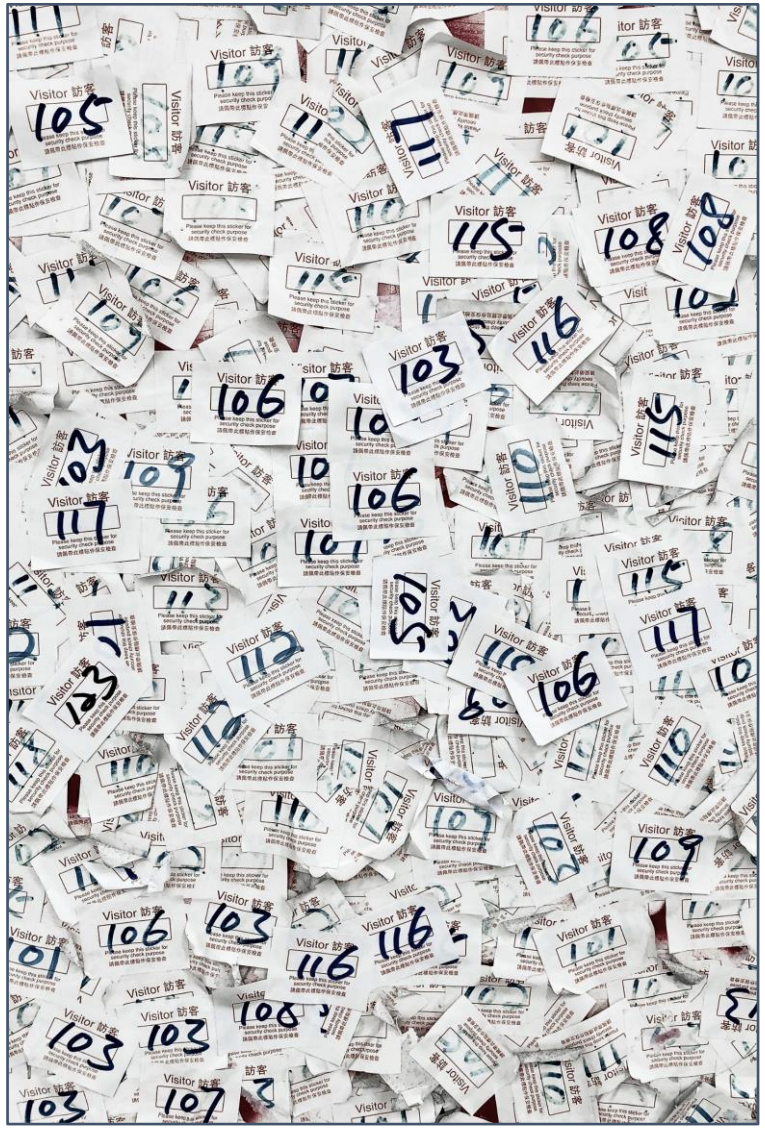
Especificación.



- En el **modelo básico de regresión** (MBR) vamos a suponer que existen:
  - Variable dependiente ***y*.**
  - k variables explicativas $\boldsymbol{x_j}$.
  - Variable o **perturbación aleatoria *u***, que recoge el efecto conjunto de todas aquellas variables que afectan al comportamiento de *y* pero que no están explicitadas en la especificación como variables x.
- El tamaño de la muestra es n.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik} + u_i \qquad i = 1,2,\ldots,n$$

Especificación.



- Notación matricial:

$$y_1 = \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{13} + \ldots + \beta_k x_{1k} + u_1$$
$$y_2 = \beta_1 x_{21} + \beta_2 x_{22} + \beta_3 x_{23} + \ldots + \beta_k x_{2k} + u_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_n = \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \ldots + \beta_k x_{nk} + u_n$$

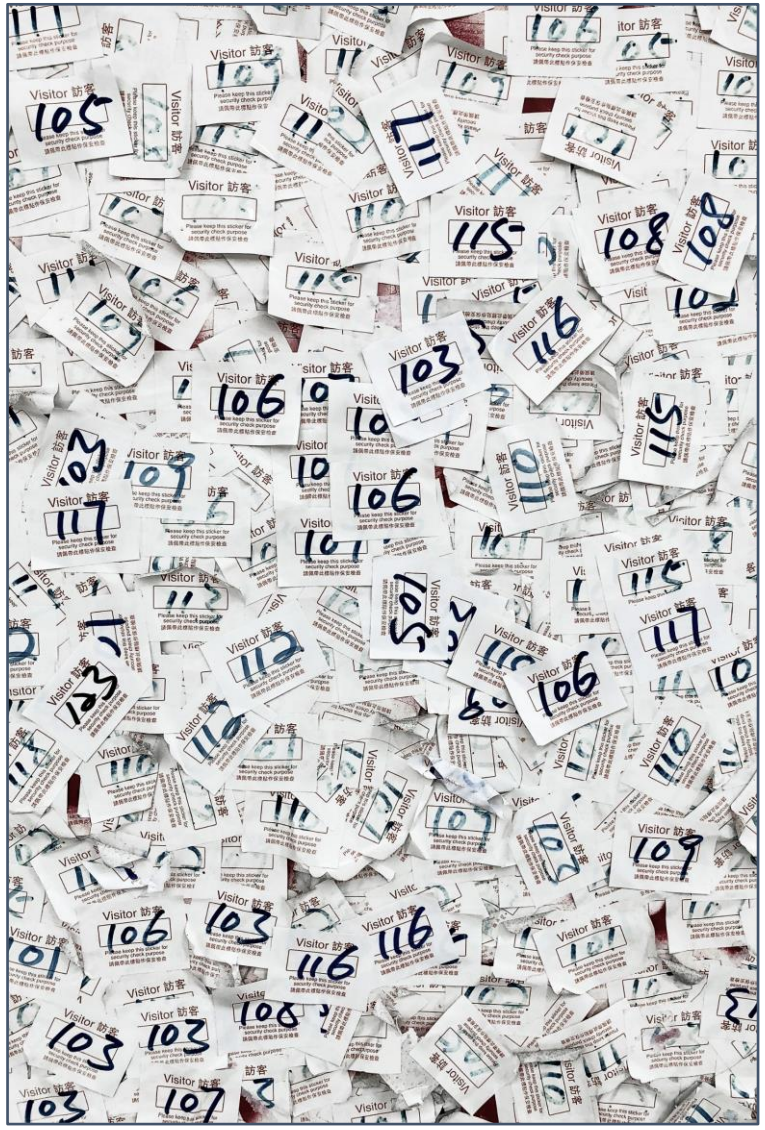
$$y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \qquad u = \begin{bmatrix} u_1 \\ u_2 \\ \ldots \\ u_n \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_k \end{bmatrix}$$

La primera columna o variable “x” suele ser un vector de 1, a fin de especificar un término independiente en la ecuación.

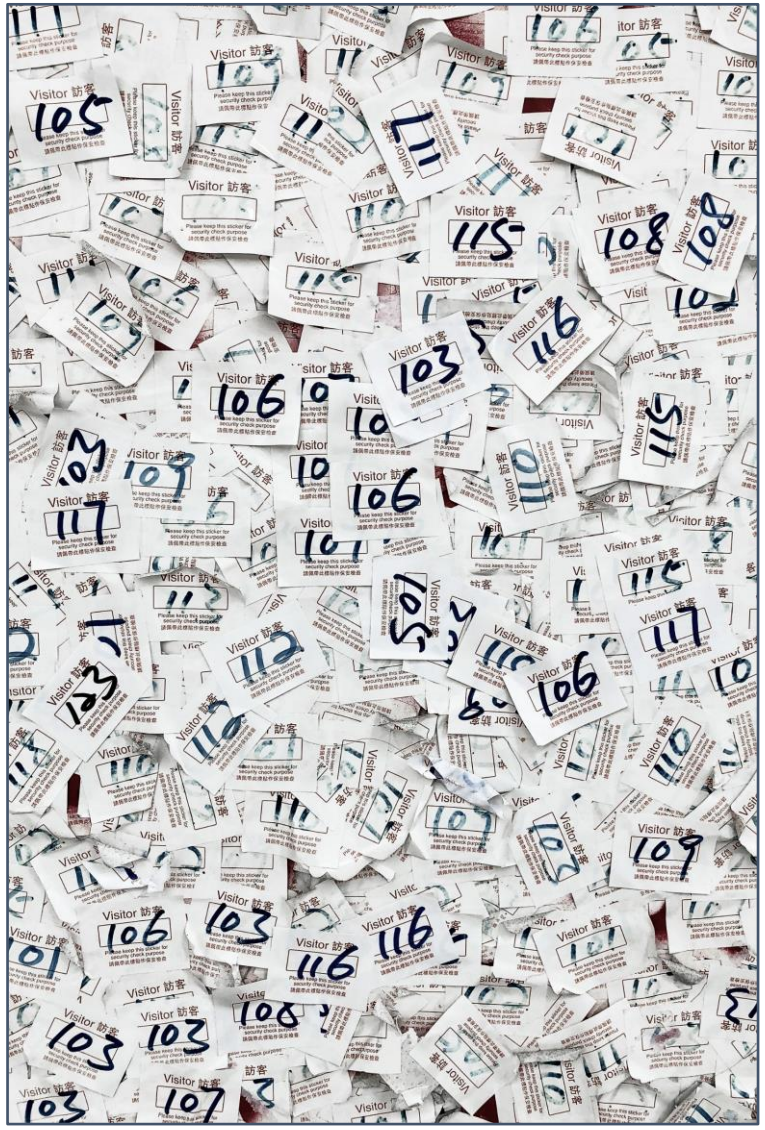
$$y = X\beta + u$$

Especificación.



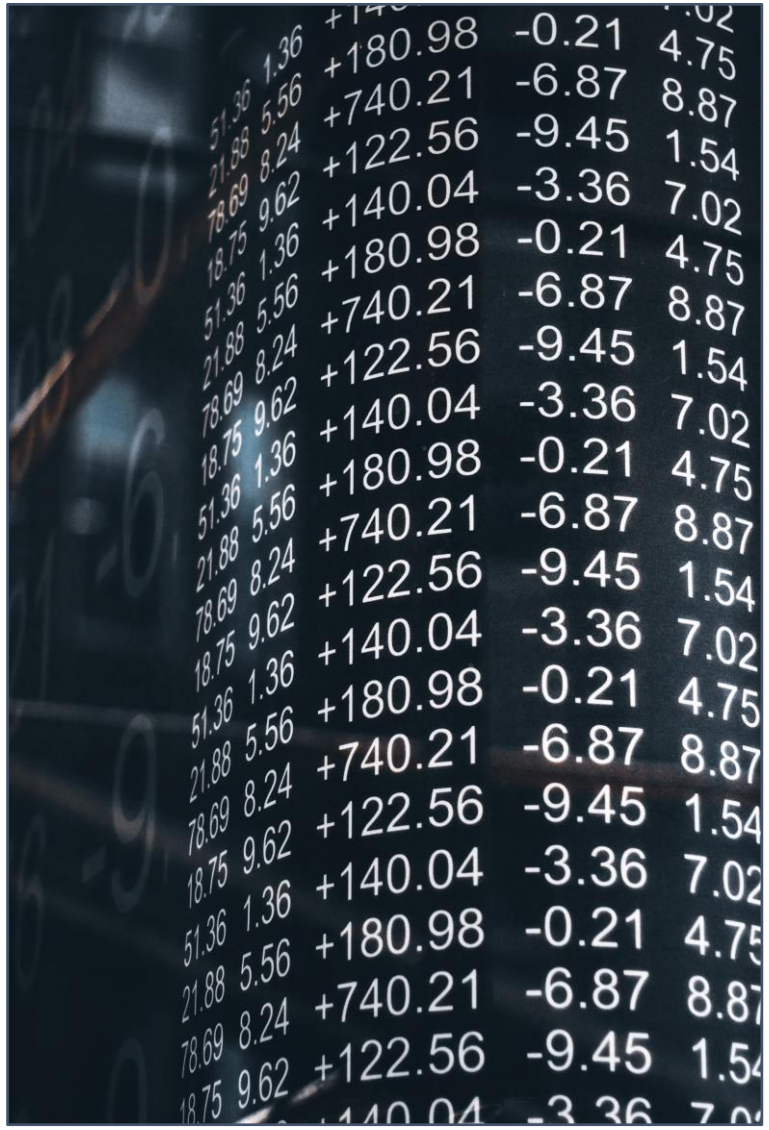
- En el modelo básico de regresión (MBR) el vector de la perturbación aleatoria ***u*** ha de cumplir con una serie de **hipótesis básicas**:
  - **Normalidad:** El vector debe seguir una distribución aproximadamente normal: $u \sim N$
  - **Media nula:** $E(u) = 0$
  - **Homoscedasticidad:** La varianza del vector de perturbaciones debe ser constante, es decir, $Var(u_i) = \sigma^2 \quad \forall i$
  - **Ausencia de autocorrelación serial:** El valor de la perturbación aleatoria para un caso no depende del valor que la perturbación aleatoria toma en otros casos, es decir, $cov(u_k, u_l) = 0 \quad \forall k \neq l$

Especificación.



- En el modelo básico de regresión (MBR), otros elementos diferentes a la perturbación aleatoria ***u*** también han de cumplir con otras **hipótesis básicas**:
  - **Permanencia Estructural:** El vector de parámetros β debe ser válido para representar el comportamiento de todos los elementos de la muestra (problema del cambio estructura).
  - **Regresores no-estocásticos:** las variables explicativas $\boldsymbol{x}_j$ no deben estar correlacionadas con el vector de perturbaciones ***u*** (problema de la endogeneidad).
  - **Rango pleno:** las variables explicativas $\boldsymbol{x}_j$ no deben tener altas correlaciones entre sí (problema de la multicolinealidad).

Estimación.



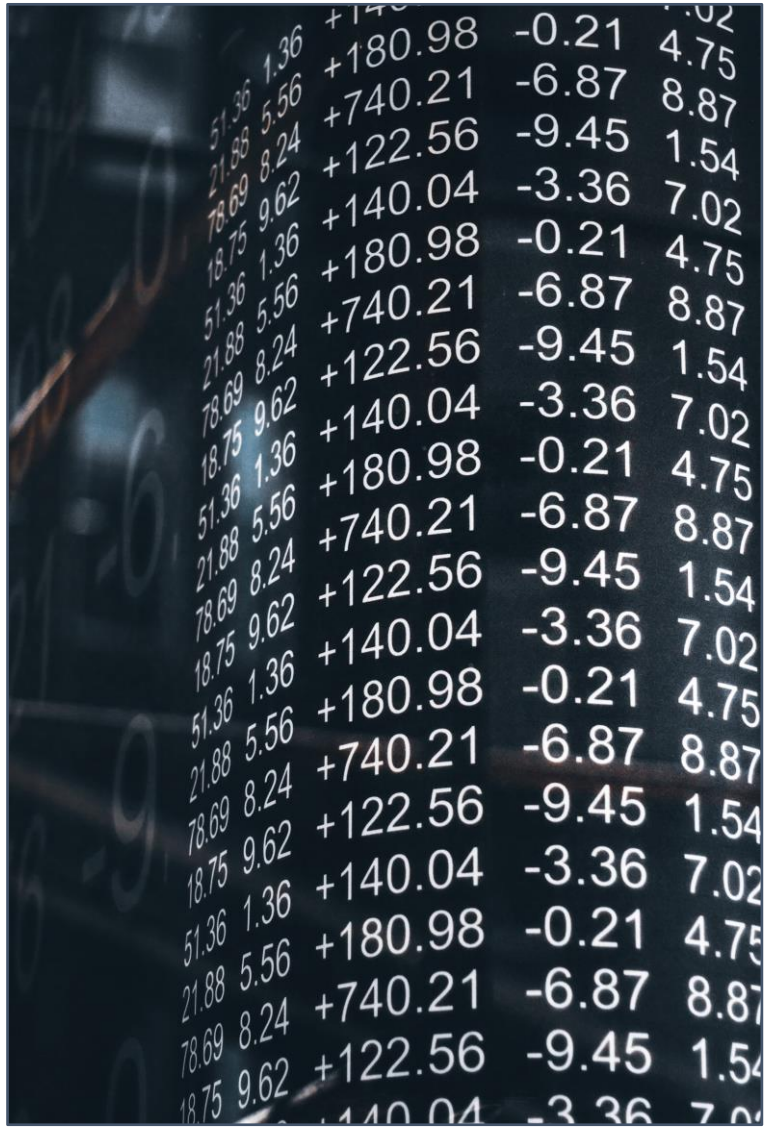
- El método de estimación básico de los parámetros estructurales β_j es el de mínimos cuadrados ordinarios (**MCO**).

- Los estimadores MCO ($\hat{\beta}_j$) tendrán las mejores propiedades (insesgadez, eficiencia, consistencia) en la medida en que se cumplan las hipótesis básicas del modelo lineal.

- En el caso de que algunas hipótesis no se cumplan, podrán aplicarse otros métodos de estimación, diversos métodos econométricos, o asumir que los estimadores carecen de algunas de las propiedades deseables.

- Estimación: $\hat{y}_i = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \ldots + \hat{\beta}_k x_{ik}$

Estimación.

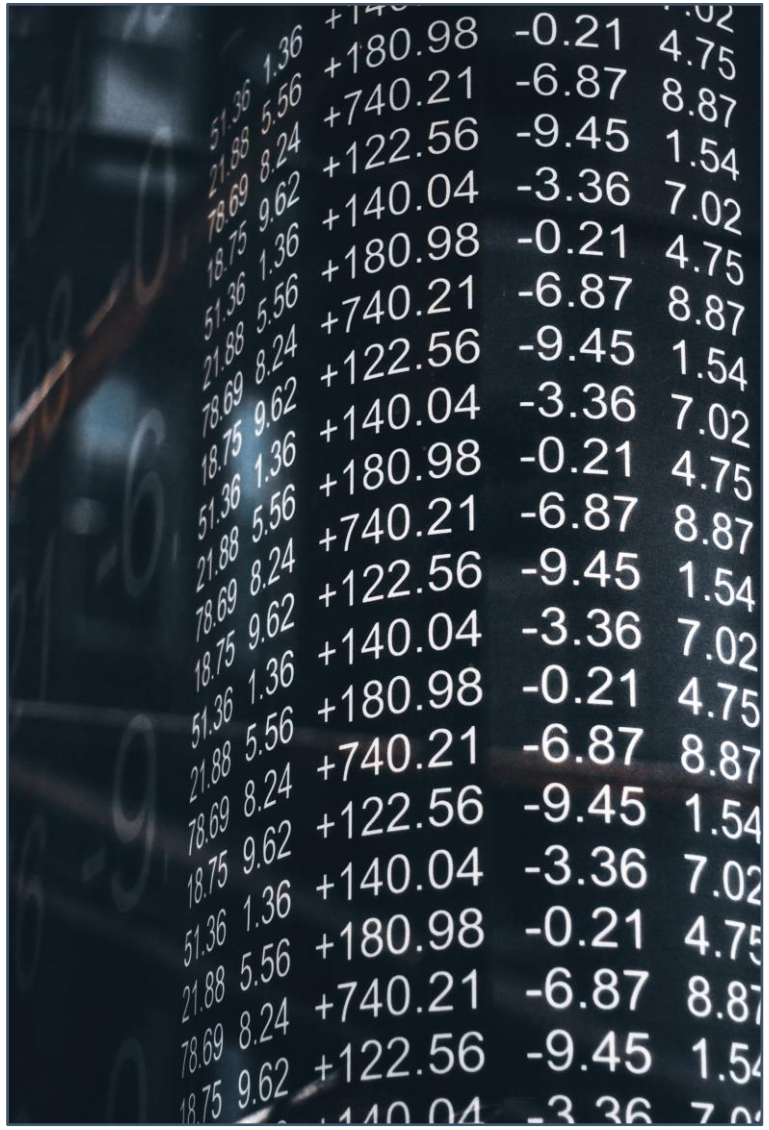


- Una vez estimados los parámetros, se podrá calcular, para cada elemento de la muestra, el **error o residuo**, como diferencia entre el valor de la variable dependiente para ese caso, y el valor estimado según la regresión:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \ldots + \hat{\beta}_k x_{ik}\right)$$

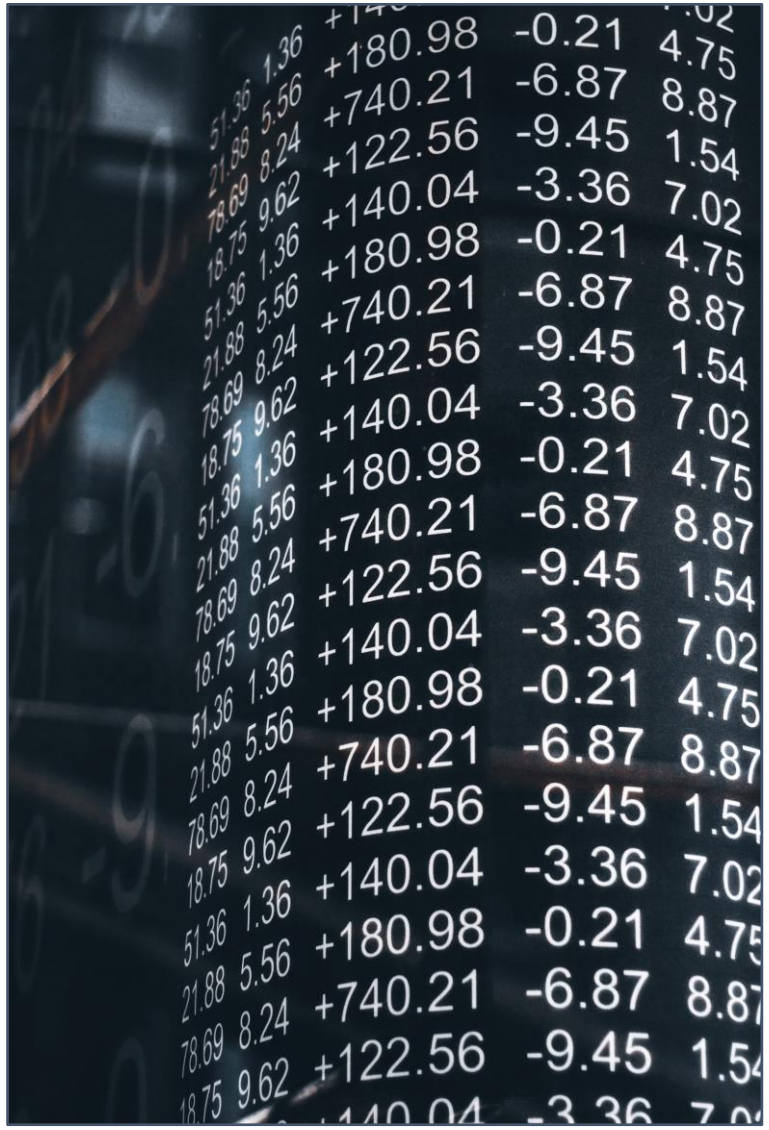
- Los residuos se consideran como las *estimaciones de las perturbaciones aleatorias*. Es por ello que el vector de residuos se utiliza para verificar el cumplimiento de las hipótesis del modelo básico referentes al comportamiento de la perturbación (normalidad, homoscedasticidad, ausencia de autocorrelación…)

Contraste y validación.



- Contraste de **significación individual** de los parámetros (contraste t).
  - H0: $\beta_j = 0$ H1: $\beta_j = 0$
- Contraste de **significación conjunta** de los parámetros (contraste t).
  - H0: $\beta_1 = \beta_2 = \cdots = \beta_j = \beta_k = 0$
  - H1: $\exists\ \beta_j / \beta_j \neq 0$

Contraste y validación.



- Bondad del ajuste: **coeficiente de determinación lineal (R^2)**. Proporción de la varianza de la variable dependiente recogida por la regresión. Mejor usar la versión corregida **($\overline{R}^2$)**.

- Si se han estimado regresiones alternativas para explicar la misma variable dependiente, y se encuentran en la misma situación en cuanto al cumplimiento de las hipótesis básicas, ¿qué especificación elegir? Se puede aplicar el **Criterio de Información de Akaike** (AIC).

- El AIC toma el siguiente valor: $AIC = 2k - 2\ln(L)$, con k= nº de parámetros estimados y ln(L) logaritmo de la función de verosimilitud de la muestra.

- A menor AIC, mejor especificación del modelo.

Contraste de hipótesis del modelo lineal.

- Cuando se estima una regresión múltiple con base en datos de corte transversal, se han de verificar prioritariamente las siguientes hipótesis (I):
  - **Linealidad** del modelo (forma funcional correcta):
    - Métodos: contrastes estadísticos (Ramsey-Reset).
    - Consecuencias del problema: estimadores sesgados, pérdida de eficiencia.
    - Posible solución: transformación previa de los datos (logarítmica, etc.) Inclusión de alguna variable omitida.

Contraste de hipótesis del modelo lineal.

- Cuando se estima una regresión múltiple con base en datos de corte transversal, se han de verificar prioritariamente las siguientes hipótesis (II):
  - **Rango pleno** (no multicolinealidad entre variables explicativas):
    - Métodos: factor de inflación de la varianza (vif).
    - Consecuencias del problema: pérdida "artificial" de significación estadística de los parámetros, "mezcla" de efectos de las variables: parámetros con valores inesperados.
    - Posible solución: transformación previa de los datos (logarítmica, etc.) Eliminación de variables.

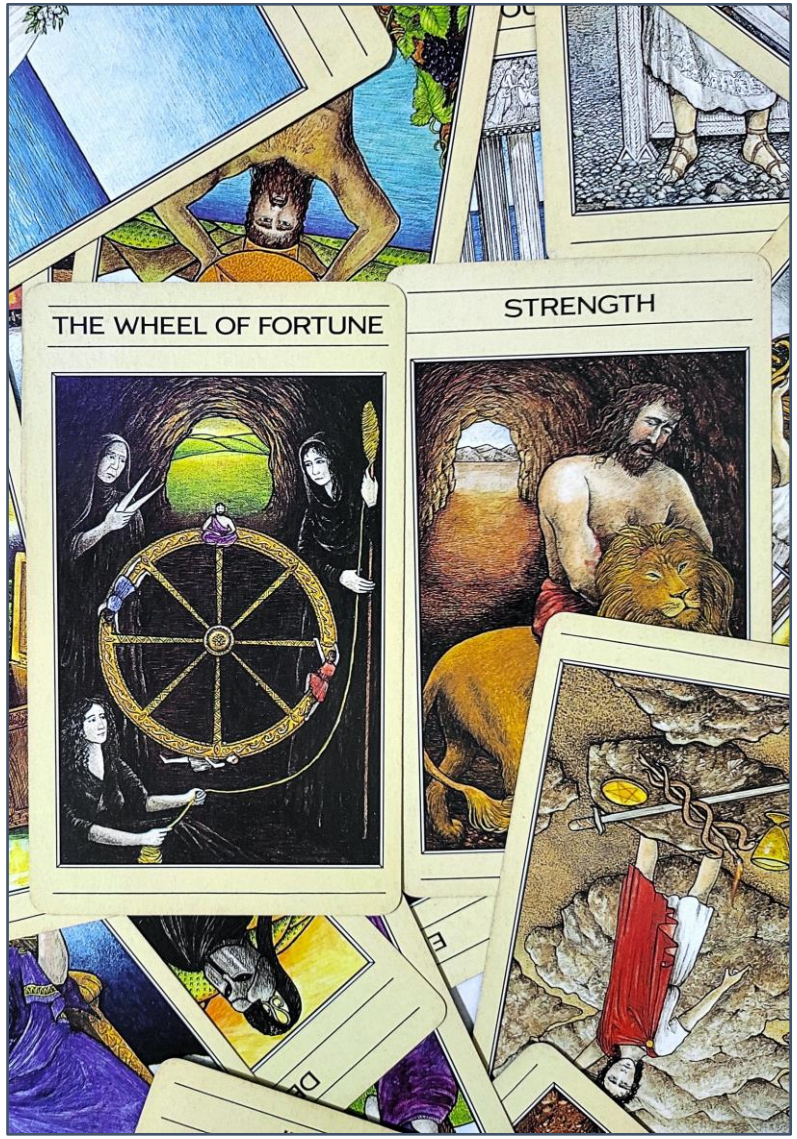
Contraste de hipótesis del modelo lineal.

- Cuando se estima una regresión múltiple con base en datos de corte transversal, se han de verificar prioritariamente las siguientes hipótesis (III):
  - **Normalidad** en el vector de perturbaciones aleatorias:
    - Métodos: gráfico (gráficos qq de los residuos) y contrastes estadísticos (Shapiro-Wilk).
    - Consecuencias del problema: pérdida de eficiencia en los estimadores.
    - Posible solución: transformación previa de los datos (logarítmica, etc.)

Contraste de hipótesis del modelo lineal.

- Cuando se estima una regresión múltiple con base en datos de corte transversal, se han de verificar prioritariamente las siguientes hipótesis (IV):
  - **Homoscedasticidad** en el vector de perturbaciones aleatorias:
    - Métodos: contrastes estadísticos (Breusch y Pagan).
    - Consecuencias del problema: pérdida de eficiencia en los estimadores.
    - Posible solución: transformación previa de los datos (logarítmica, deflactación de variables, etc.)

Utilización del modelo.



- Una vez estimado, validado y contrastado el modelo, puede ser utilizado principalmente de dos modos:
  - **Análisis Estructural: interpretar** el significado económico de los parámetros o coeficientes estimados. Tiene especial relevancia la interpretación de los **signos**. Si las variables no vienen transformadas en logaritmos: propensiones marginales. Si vienen expresadas en logaritmos: elasticidades.
  - **Previsión / simulación:** el modelo estima el **valor esperado** de la variable dependiente ante un **escenario** formado por valores propuestos para las variables explicativas.

¡Muchas gracias!