Flexibilización de los supuestos del modelo clásico

Multicolinealidad

Multicolinealidad perfecta

- Si se tiene una regresión con *k* variables explicativas, $X_1, X_2, \ldots, X_n$, existe una relación lineal exacta si:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0$$

Donde $\lambda_1, \lambda_2, \ldots, \lambda_k$ son constantes tales que no todas son simultáneamente cero.

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki}$$

Multicolinealidad casi perfecta

- Si se tiene una regresión con *k* variables explicativas, $X_1, X_2, \ldots, X_n$, existe una relación lineal exacta si:

$$\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k + v_i = 0$$

Donde v_i es un término de error estocástico.

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \cdots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i$$

Grados de multicolinealidad

X_2	X_3	X_3^*	X_4
10	50	52	100
15	75	75	225
18	90	97	324
24	120	129	576
30	150	152	900

Problemas de la multicolinealidad

Si es perfecta, los coeficientes de las regresoras son indeterminados, con errores estándar infinitos.

Si es casi perfecta, los coeficientes tienen errores estándar muy grandes (los intervalos de confianza se hacen muy amplios; poca precisión en la).

Consecuencias de la multicolinealidad

La multicolinealidad no viola los supuestos básicos de la regresión.

El problema es conseguir errores estándar pequeños.

Consecuencias en la práctica

1. Aunque los estimadores de MCO sigan siendo MELI, tienen varianzas y covarianzas grandes. → poca precisión en la estimación.

2. Intervalos de confianza mucho más amplios.

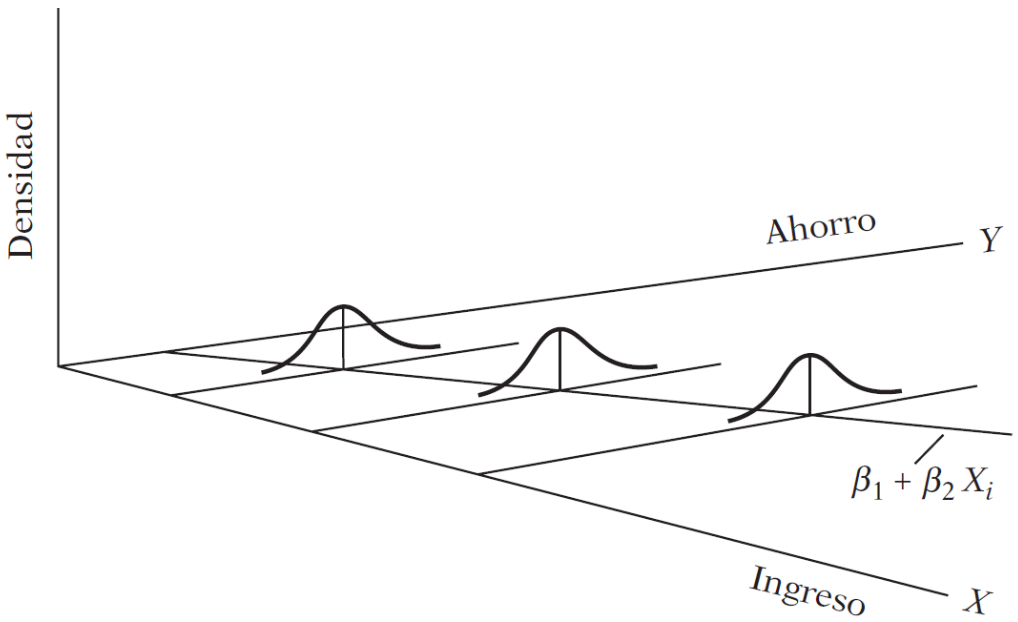
3. Razón *t* de uno o más coeficientes tiende a ser no significativa.

4. R^2 alta (mayor a 0.8) a pesar de tener razones *t* no significativas.
   - Valores *t* no significativos, R^2 elevada y valor de *F* significativo.

5. Estimadores de MCO y errores estándar sensibles a cambios en la muestra.

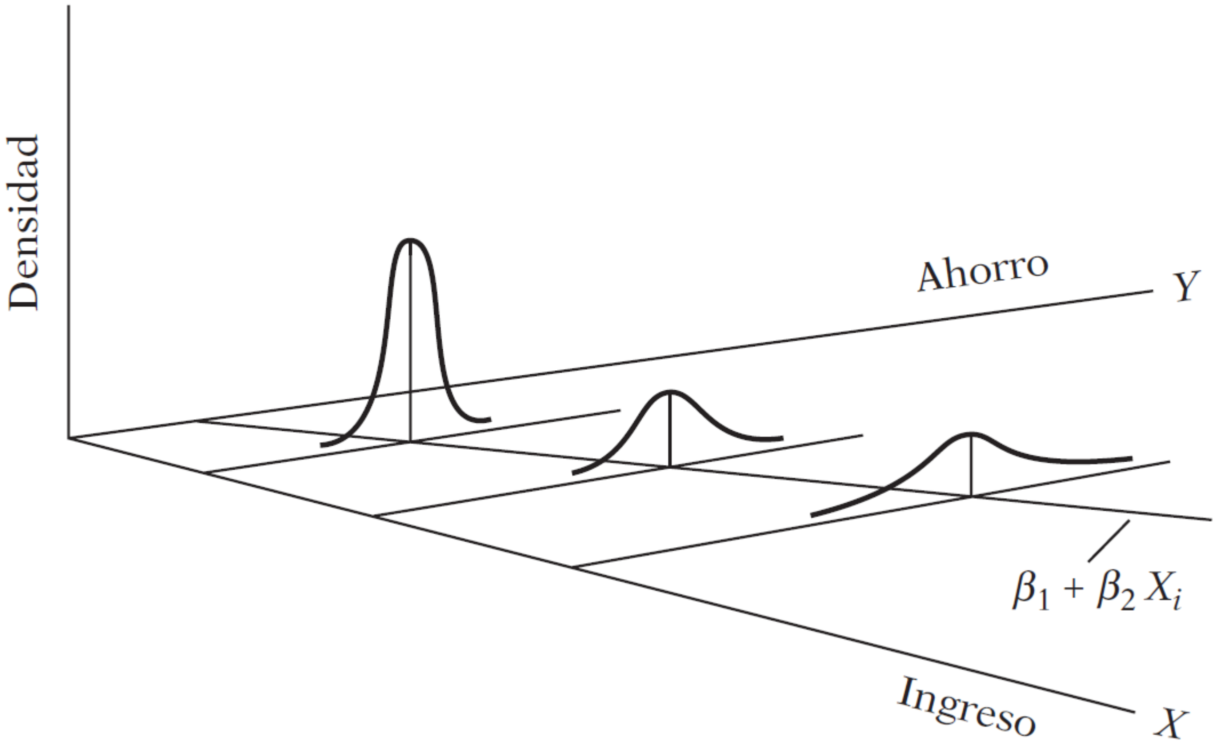
Heteroscedasticidad

Homoscedasticidad



- Recordando…

$$E\left(u_i^2\right) = \sigma^2 \quad para\ i = 1, 2, \ldots, n$$



Heteroscedasticidad

$$E(u_i^2) = \sigma_i^2 \quad para\ i = 1, 2, \ldots, n$$

Razones de encontrar heteroscedasticidad

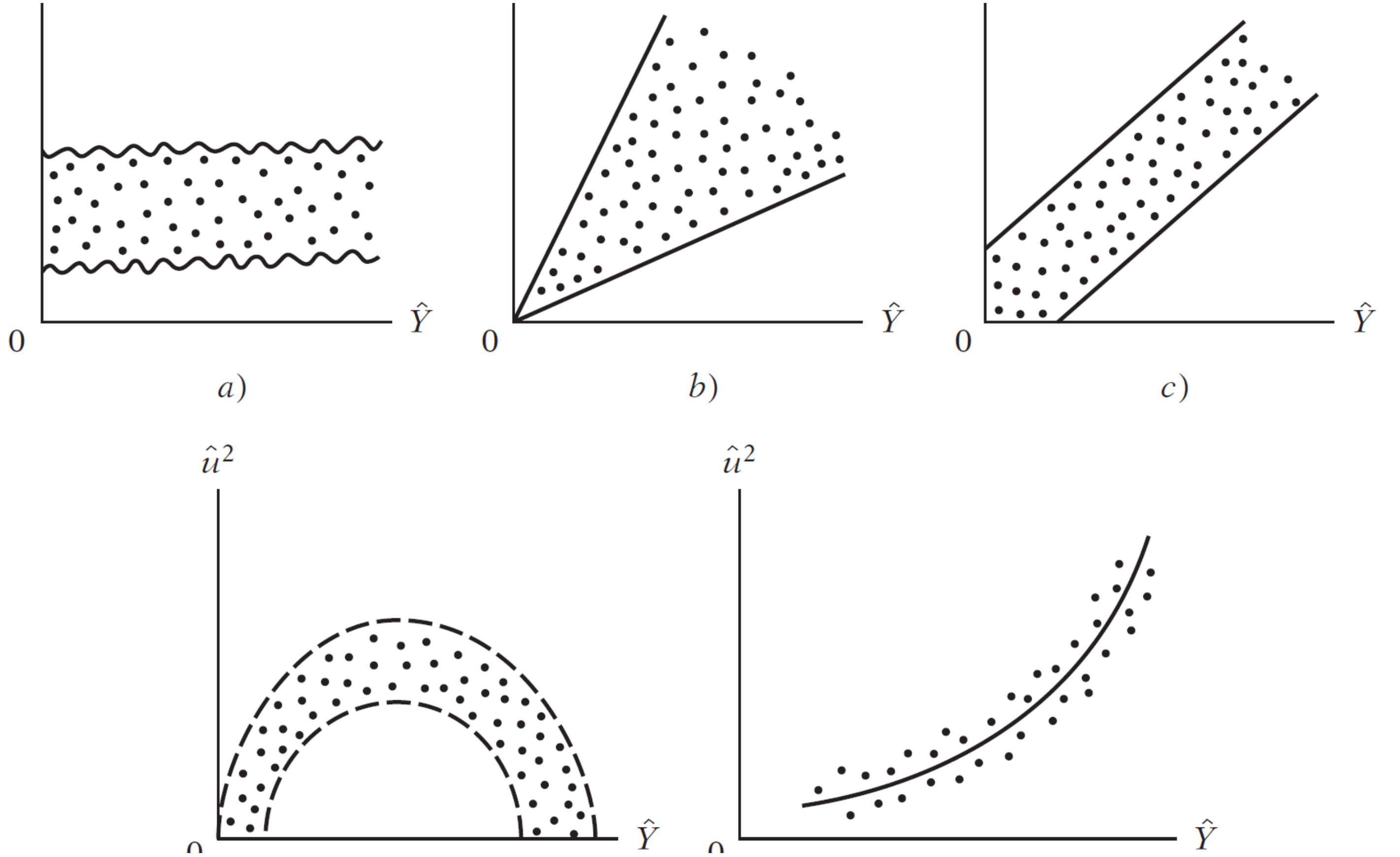
1. Modelos de aprendizaje del error
   - Conforme la gente aprende, se reducen sus errores con el tiempo.
2. Conforme aumenta el ingreso, la gente puede disponer más de su dinero y puede haber mayor variación en el consumo.
3. Si mejoran las técnicas de recolección de datos, σ_i^2 probablemente se reducirá.
4. Puede haber heteroscedasticidad cuando ∃ datos atípicos.
5. Al violar el supuesto 9 (modelo correctamente especificado).
   - Omisión de variables importantes.
   - Forma funcional incorrecta.
6. Asimetría en la distribución de regresoras.
   - P. ej. Ingreso, riqueza o escolaridad.

Consecuencias de usar MCO con heteroscedasticidad

- La mayoría de las veces no se puede establecer intervalos de confianza y probar hipótesis con pruebas t y F (la varianza de las βs son demasiado grandes).
- Las conclusiones o inferencias pueden estar muy equivocadas.

Detección de la heteroscedasticidad

- $\exists$ numerosos métodos, informales y formales, de probar si se presenta heteroscedasticidad.



Métodos informales

- **Naturaleza del problema**
  - Si ya se tienen expectativas sobre investigaciones anteriores que han demostrado tener varianzas desiguales.
- **Método gráfico**
  - Si no se tienen expectativas *a priori*, se puede hacer MCO suponiendo que ∄ heteroscedasticidad y hacer un examen de los residuos al cuadrado y graficarlos para ver si presentan algún comportamiento sistemático.

Métodos formales: Prueba de Park

- Park busca formalizar el método gráfico. Sugiere que σ_i^2 es una función de la variable X_i. Específicamente, $\sigma_i^2 \approx \epsilon_i^2 = \sigma^2 X_i^{\beta} e^{v_i}$

$$\begin{aligned} \ln\epsilon_i^2 &= \ln\sigma^2 + \beta\ln X_i + v_i \\ &= \alpha + \beta\ln X_i + v_i \end{aligned}$$

- Es un procedimiento de dos etapas:

1. Realizar la regresión por MCO, asumiendo $\nexists$ heteroscedasticidad. Obtener los residuos, ϵ_i^2.
2. Realizar una segunda regresión, ahora con la forma funcional propuesta por Park.
   - Si β resulta estadísticamente significativa, sugeriría heteroscedasticidad en los datos.

- Pudiera existir el problema de que v_i no cumpla los supuestos de MCO.

Prueba de Breusch-Pagan-Godfrey (BPG)

- Consideremos un modelo con *k* variables.

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

- Asuma que la varianza del error σ_i^2 es una función de las variables Z no estocásticas (una o más de las X pueden servir como Z.

$$\sigma_i^2 = \alpha_1 + \alpha_2 Z_{2i} + \cdots + \alpha_m Z_{mi}$$

- σ_i^2 es función lineal de las Z. Si $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$, $\sigma_i^2 = \alpha_1$ $(constante)$.
- Para probar si σ_i^2 es homoscedástica, se debe probar la H_0 de que $\alpha_2 = \alpha_3 = \cdots = \alpha_m = 0$ "Homoscedasticidad".

Procedimiento de la prueba BPG

1. Estimar la regresión con *k* variables mediante MCO y obtener residuos.
2. Obtener $\tilde{\sigma}^2 = \sum \hat{\epsilon}_i^2 / n$
3. Construir variables p_i, definidas como

$$p_i = \frac{\hat{\epsilon}_i^2}{\tilde{\sigma}^2}$$

1. Hacer la regresión de los p_i sobre las *Z*.
2. Obtener la SCE y defina $\Theta = \frac{1}{2}(\text{SCE})$

$$\Theta \sim \chi^2_{m-1}$$

Autocorrelación

Correlación serial

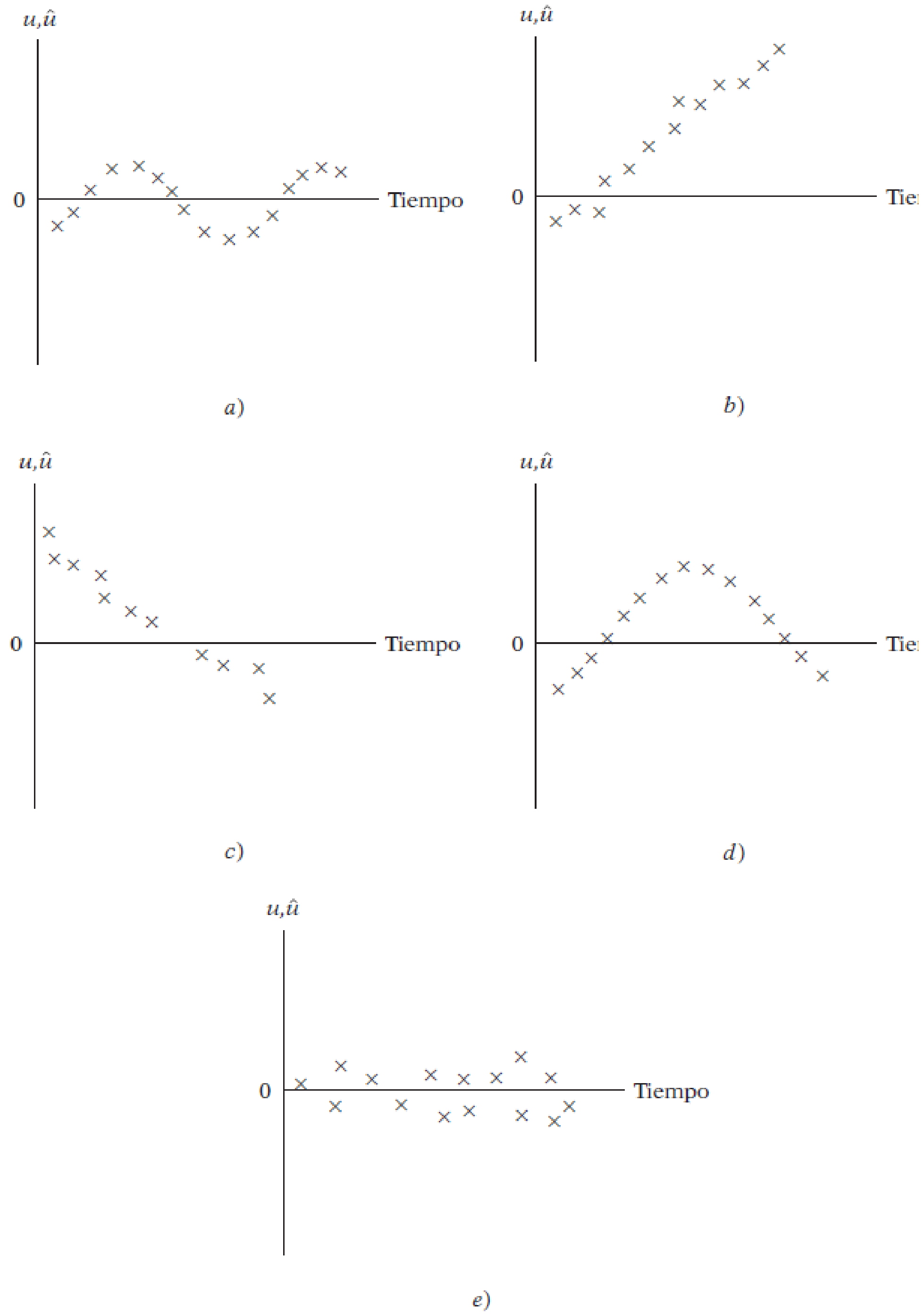
Autocorrelación

Suponemos que los errores tienen "memoria" y que están correlacionados en el tiempo.

Si existiera correlación en los errores de datos de corte transversal, se le llama autocorrelación espacial.

Con series de tiempo es muy probable que observaciones consecutivas muestren intercorrelaciones.

Especialmente si el intervalo entre mediciones es muy corto (segundos, minutos, horas, días,…).



… autocorrelación

- El modelo clásico de regresión lineal asume que ∄ autocorrelación en los errores. Esto es,
$$cov(\epsilon_i, \epsilon_j | x_i, x_j) = E(\epsilon_i, \epsilon_j) = 0, con\ i \neq j$$

- Esto quiere decir que ningún residuo es influenciado por otro. *"Los agentes no aprenden"*.

Razones de encontrar autocorrelación

- Inercia o pasividad
  - Muchas series financieras o económicas presentan ciclos en su comportamiento.
  - P. ej. Una recesión.
- Sesgo de especificación: variables excluídas
  - Suponga que el modelo correcto es $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_{4t} X_4 + \epsilon_t$,
    - donde Y_t: cantidad demandada de boletos de cine,
    - X_2: precio del boleto de cine,
    - X_3: ingreso del consumidor,
    - X_4: precio de la línea de boliche y
    - t: tiempo.
  - Pero se estima $Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + v_t$
    - El error estaría dado por $v_t = \beta_{4t} X_4 + \epsilon_t$ y presentaría un patrón sistemático.

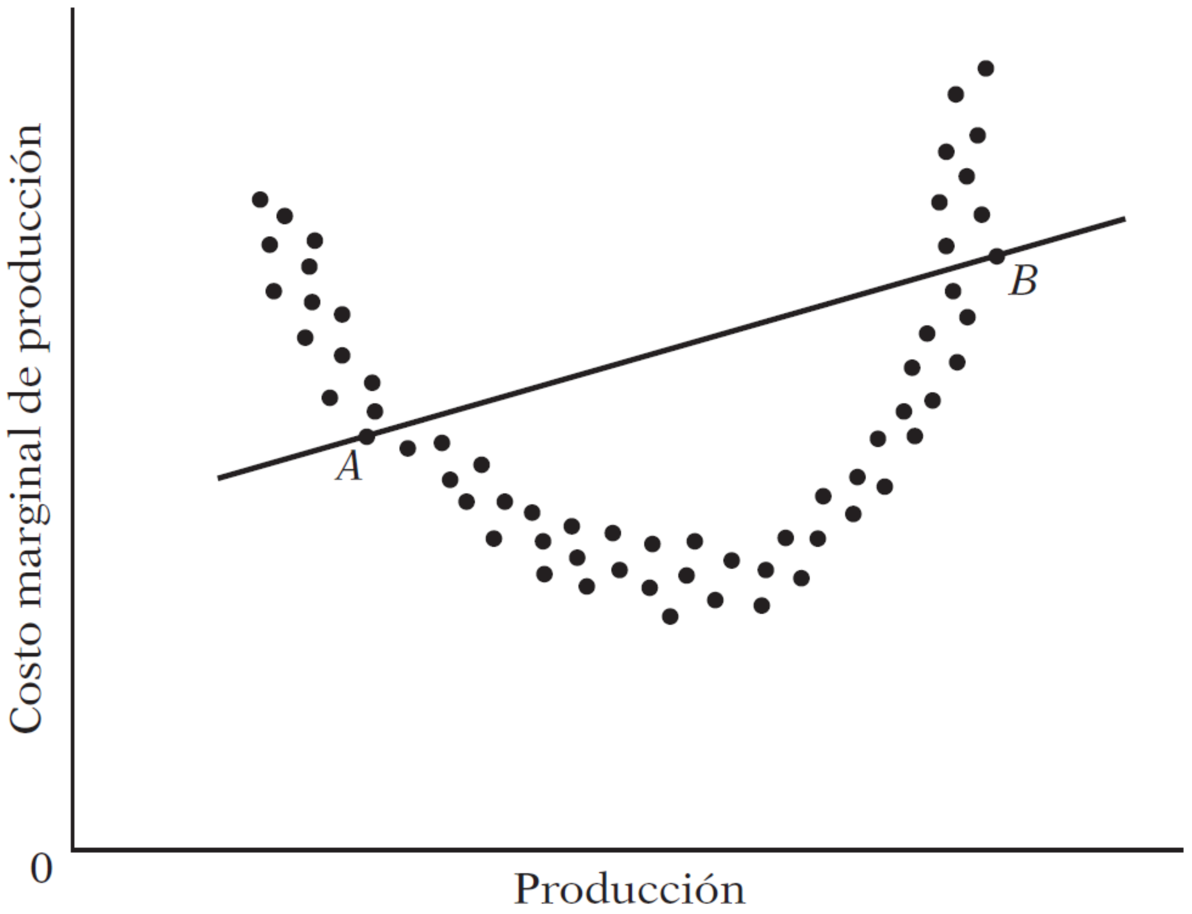
Sesgo de especificación: forma funcional incorrecta

- Suponga el modelo correcto para estudiar la relación costo-producción está dado por

$$CMg_i = \beta_1 + \beta_2 prod_i + \beta_3 prod_i^2 + \epsilon_i$$

- Pero se ajusta

$$CMg_i = \alpha_1 + \alpha_2 prod_i + v_i$$



Rezagos

- En series de tiempo, es muy común que el comportamiento de una variable el día de hoy, dependa de su comportamiento en periodos anteriores.
  - Cuánto gasta una familia marzo puede explicarse en parte por cuánto gastó en febrero.

$$Consumo_t = \beta_1 + \beta_2 ingreso_t + \beta_3 consumo_{t-1} + \epsilon_t$$

- A este tipo de regresiones se les conoce como **autorregresión**.
  - “Los consumidores no cambian fácilmente sus patrones de consumo”.

Manipulación de datos

Interpolación o extrapolación

- P. ej., al querer interpolar datos trimestrales, al promediar las cifras se suavizan las fluctuaciones mensuales.

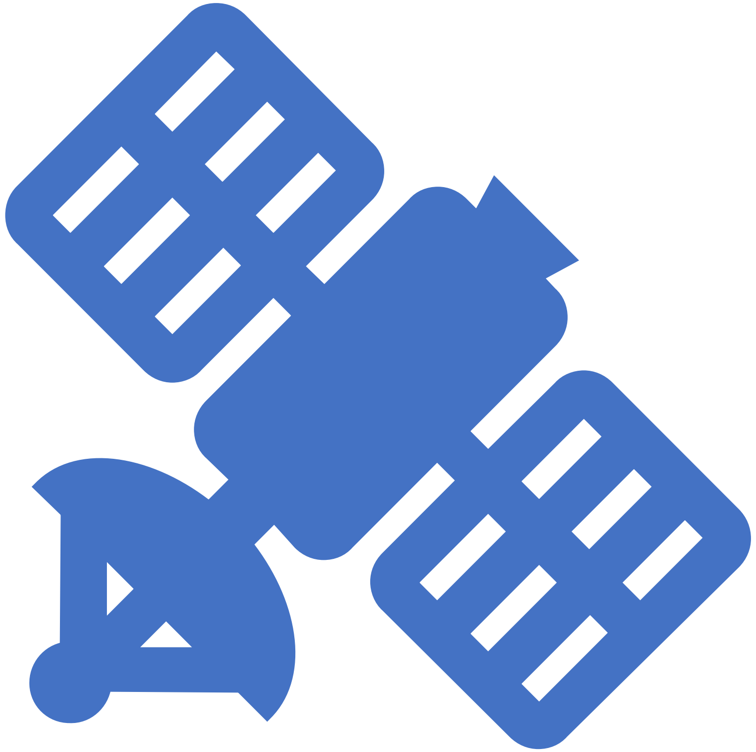
Transformación de datos

- Si se le aplica alguna transformación a los datos, puede corregir o provocar autocorrelación.

No estacionariedad

- Serie de tiempo estacionaria: su media, varianza y covarianza son invariantes en el tiempo.

Detección de autocorrelación



Método gráfico

- Gráfica secuencial de tiempo
  - Graficar los residuos vs. el tiempo

- Gráfica de los residuos estandarizados vs. el tiempo
  - Residuos estandarizados:

$$\frac{residuos}{error\ estándar\ de\ la\ regresión} = \frac{\hat{\epsilon}_t}{\sqrt{\hat{\sigma}^2}} = \frac{\hat{\epsilon}_t}{\hat{\sigma}}$$

Prueba de las “rachas”

Prueba de las rachas	También conocida como la “prueba de Geary”.
Es una prueba no paramétrica:	No se hacen supuestos sobre la distribución de las observaciones.
Racha:	Sucesión ininterrumpida de un símbolo o atributo.
Longitud de la racha:	Número de elementos que contiene.

Prueba de las rachas

$$N: número\ total\ de\ observaciones = N_1 + N_2$$

$$N_1: número\ total\ de\ símbolos\ +$$

$$N_2: número\ total\ de\ símbolos\ -$$

$$R = número\ de\ rachas$$

- Según la H_0, de que los resultados sucesivos son independientes y $N_1 > 10$ ^ $N_2 > 10$

$Media$: $$E(R) = \frac{2N_1N_2}{N} + 1$$

$Varianza$: $$\sigma_R^2 = \frac{2N_1N_2(2N_1N_2 - N)}{(N)^2(N-1)}$$

$Si\ se\ sostiene\ H_0 \rightarrow$ $\mathrm{Prob}[E(R) - 1.96\sigma_R \leq R \leq E(R) + 1.96\sigma_R] = 0.95$

- Regla de decisión: $Rechazar\ H_0\ si\ R\ está\ fuera\ del\ intervalo$.

Prueba de Durbin-Watson

- El estadístico, *d*, de Durbin-Watson se define:

$$d = \frac{\sum_{t=2}^{n}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=1}^{n}(\hat{\epsilon}_t)^2}$$

- Supuestos:
  1. El modelo incluye el intercepto.
  2. Las variables X son no estocásticas.
  3. Las perturbaciones, ϵ_t se generan mediante un esquema autorregresivo de primer orden: $\epsilon_t = \rho\epsilon_{t-1} + v_t$.
  4. ϵ_t está normalmente distribuido.
  5. El modelo no incluye valores rezagados de Y.
  6. No hay observaciones faltantes en los datos.

… prueba de Durbin-Watson

$$\hat{\rho} = \frac{\sum \hat{\epsilon}_t \hat{\epsilon}_{t-1}}{\sum \hat{\epsilon}_t^2}$$

- $\hat{\rho}$ es el estimador de ρ, y es el coeficiente de autocorrelación muestral de primer orden.

$$d \approx 2(1 - \hat{\rho})$$

- Como $-1 \leq \rho \leq 1$,

$$0 \leq d \leq 4$$

- ¿Cuándo diría este estadístico que $\nexists$ autocorrelación de primer orden?