Supuestos Regresión Lineal: Autocorrelación

Clase 29

Nicolás Mejía M.
n.mejia10@uniandes.edu.co

Probabilidad y Estadística II
Departamento de Ingeniería Industrial
Universidad de Los Andes, Bogotá, Colombia

2020-19

Outline

Definición

Cuando planteamos el modelo de regresión lineal,

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_k X_k + \epsilon$$

hicimos el supuesto de que los errores aleatorios:

$$\epsilon \sim_{iid} N(0, \sigma^2)$$

es decir, las observaciones Y_i son independientes entre ellas. Es decir, el valor que toma un Y_i no se altera de ninguna forma por el valor de los otros.

Pero esto no es verdad siempre.

Definición

Existen diferentes formas en las que se puede tener una situación de dependencia entre las observaciones, algunas son:

- Dependencia por aglomerados: A veces nuestros datos pueden venir de estructuras de grupos en donde se tienen comportamientos comunes. Ej: los estudiantes suelen formar grupos de trabajo y estudian lo mismo, luego sus notas suelen ser bastante similares.
- Dependencia espacial: Cuando se tienen datos según su ubicación en un territorio. Ej: La imagen política de un candidato es alta en las ciudades en donde hayan ejercido algún empleo publico, y decrece a medida que se distancia de dichas ciudades.
- Dependencia temporal: Cuando se recolectan datos a través del tiempo. Ej: la población viva el día de hoy, depende de la poblacion viva que se tenia ayer.

Definición

Autocorrelación

Decimos que hay problemas de autocorrelación cuando nuestros datos Y no son independientes entre sí, es decir, existe algún tipo de correlación entre diferentes observaciones Y_i y Y_j

Nos enfocaremos únicamente en el caso de la autocorrelación temporal. Generalizando, realmente se estaría hablando de cuando se tienen observaciones cuyo orden de aparación en la base de datos tienen un patrón lógico, que por lo general es el tiempo en que se observa, aunque puede ser otro.

IMPORTANTE: Si sus datos no tienen una estructura de orden en su recolección, NO tiene ningún sentido lo que haremos a continuación. Esto no quiere decir que ya no haya autocorrelación, sino que esta no seria temporal.

Definición

De esta forma realmente se tiene una situación de la forma:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \cdots + \beta_k X_{tk} + \epsilon_t$$

donde ahora los errores aleatorios:

$$\epsilon_t \sim N(0, \sigma^2)$$

y

$$Cov(\epsilon_t, \epsilon_s) \neq 0$$

es decir, hay una estructura de covarianza a medida que se va avanzado en el tiempo.

En este caso decimos que hay AUTOCORRELACIÓN

Definición

Los supuestos originales eran:

$$E(\epsilon) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

$$Var(\epsilon) = \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_1, \epsilon_2) & Var(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\epsilon_1, \epsilon_n) & Cov(\epsilon_2, \epsilon_n) & \cdots & Var(\epsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & ^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

Definición

Si hay autocorrelación se tendría:

$$E(\epsilon) = \begin{bmatrix} E(\epsilon_1) \\ E(\epsilon_2) \\ \vdots \\ E(\epsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

$$Var(\epsilon) = \begin{bmatrix} Var(\epsilon_1) & Cov(\epsilon_1, \epsilon_2) & \cdots & Cov(\epsilon_1, \epsilon_n) \\ Cov(\epsilon_1, \epsilon_2) & Var(\epsilon_2) & \cdots & Cov(\epsilon_2, \epsilon_n) \\ \vdots & \vdots & \vdots & \vdots \\ Cov(\epsilon_1, \epsilon_n) & Cov(\epsilon_2, \epsilon_n) & \cdots & Var(\epsilon_n) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & \sigma_{1,2} & \cdots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma^2 & \cdots & \sigma_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{n,1} & \sigma_{n2} & \cdots & \sigma^2 \end{bmatrix} = \sigma^2\Omega \neq \sigma^2\mathbf{I}$$

Outline

Implicaciones

Varianza de coeficientes

$$Var(\hat{\beta}) = \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}$$

La expresión encontrada para la varianza deja de ser valida!

Prueba F

El estadístico $F = \frac{MSR}{MSE}$ ya no tiene la distribución F que habíamos determinado!. Para esto necesitaríamos conocer la estructura de covarianzas por observación, cosa que no tenemos.

Estimador de la Varianza

De forma análoga, es claro entonces que el mejor estimador de la varianza ya NO es el MSE debido a que no tiene en cuenta la estructura de covarianzas.

Implicaciones

Inferencia Estadística

Ahora bien, si la varianza de los coeficientes estimados no tiene sentido, entonces nuestros procedimientos de inferencia para combinaciones lineales tampoco:

$$\hat{\theta} = \sum_{j=0}^{k} c_j \hat{\beta}_j = \mathbf{c}^T \hat{\beta}_j \rightarrow \hat{Var}(\hat{\beta}) = \mathbf{c}^T MSE \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{c}$$

$$\text{Pruebas de hipótesis } t = \frac{\hat{\theta} - \theta_0}{\sqrt{\hat{Var}(\hat{\beta})}} \sim t_{gIE}$$

$$\text{Intervalos de confianza } IC_{1-\alpha/2}(\theta) = \hat{\theta} \pm t_{1-\alpha/2,gIE} \sqrt{\hat{Var}(\hat{\beta})}$$

Implicaciones

- Las pruebas de significancia individual y global no son creíbles: la varianza puede estar sub o sobrestimandose.
- Los intervalos de confianza tienen amplitudes erroneas! Ya no se puede asegurar su verdadero nivel de confianza.
- En resumen, se pierde la posibilidad de hacer las conclusiones con el nivel de significancia que se afirma.
- Los estimadores siguen siendo insesgados, pero dejan de ser eficientes!

Outline

Identificación

Para saber si hay problemas de Autocorrelación, hay estrategias gráficas y analíticas. La mayoria se basan en los residuales.

La autocorrelación está asociada a la covarianza entre los errores aleatorios

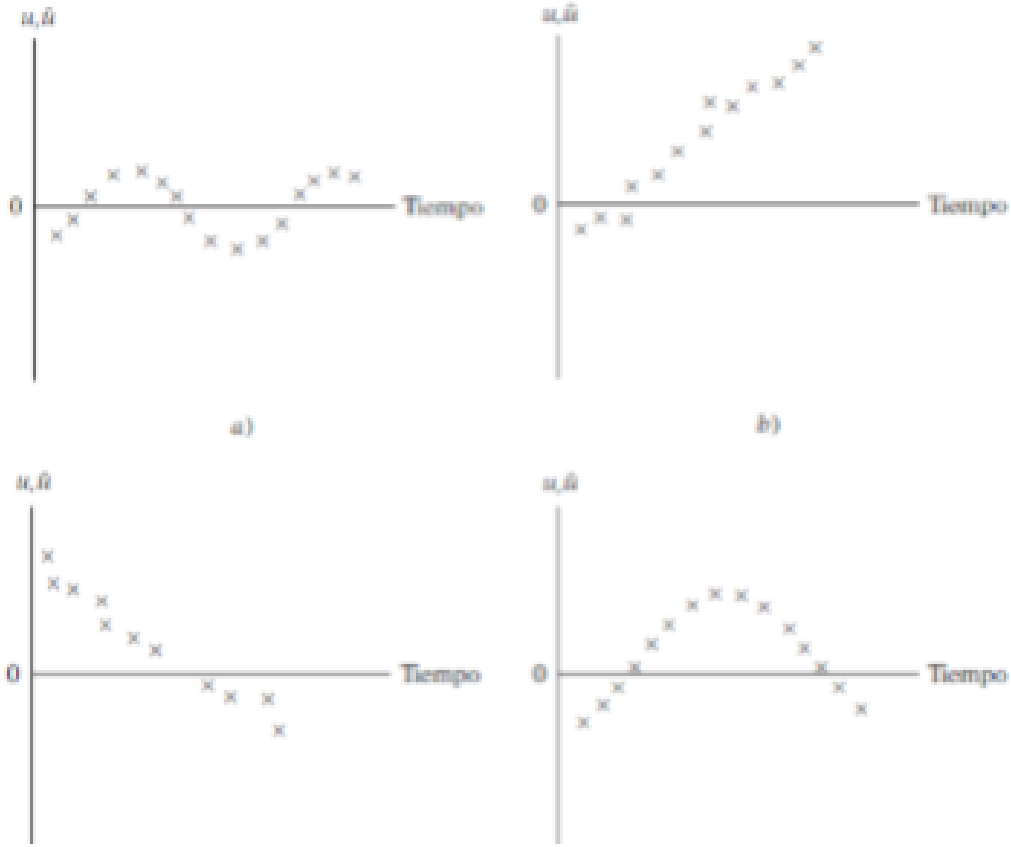
$$e_i = Y_i - \hat{Y}_i$$

Si NO hay autocorrelación, los e_t deberían ser completamente aleatorios en el tiempo, pero en caso de que si haya, estos mostrarán patrones

Identificación

Métodos Gráficos

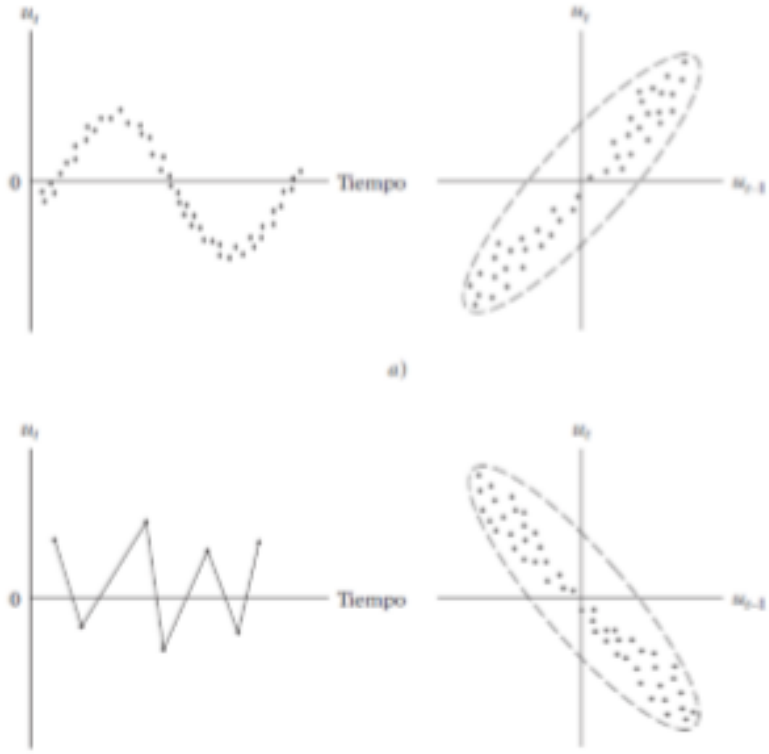
Si se gradican los e_t vs t, estos deben mantener un comportamiento aleatorio. De lo contrario, hay problemas de Autocorrelación



Identificación

Métodos Gráficos

Si se gradican los e_t vs e_{t-1}, estos deben mantener un comportamiento aleatorio. De lo contrario, hay problemas de Autocorrelación



a) Autocorrelación positiva y *b*) autocorrelación negativa.

Identificación

Métodos analíticos

Las pruebas estadísticas NO pueden probar la presencia de autocorrelación de forma general. Luego, se asumirá un modelo para los errores aleatorios:

Modelo AR(1)

Decimos que una serie de tiempo sigue un proceso autoregresivo de orden 1 o AR(1) si se satisface la siguiente expresión:

$$\epsilon_t = \rho\epsilon_{t-1} + a_t$$

donde a_t son otros errores aleatorios que NO están correlacionados, y ρ corresponde al coeficiente de correlación entre ϵ_ty ϵ_{t-1}.

Este es el modelo temporal mas simple que se tiene, y nuestras pruebas estadísticas serán para verificar y corregir sobre ese modelo

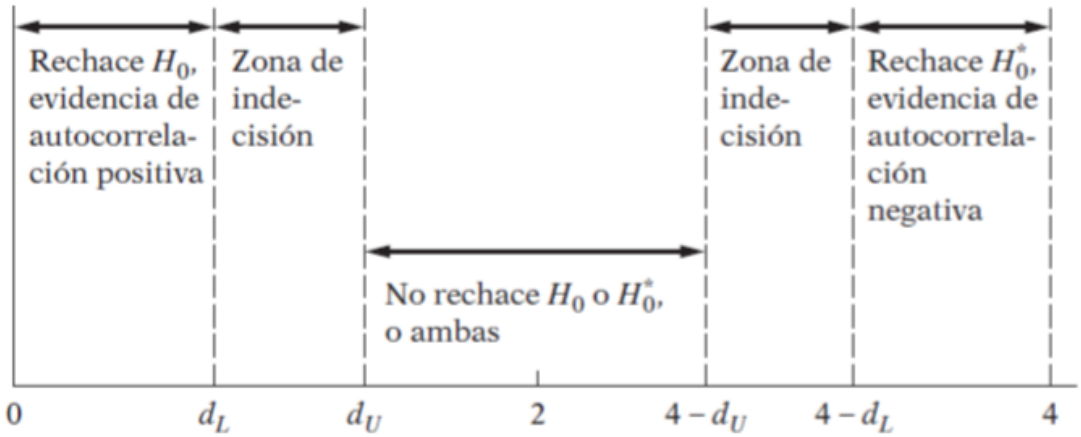
Identificación

Métodos analíticos: Prueba de Durbin-Watson

$$H_0 : \text{No hay Autocorrelación}$$

$$H_0 : \text{Si hay Autocorrelación}$$

$$DW = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \approx 2(1 - \hat{\rho})$$



Outline

Corrección

La metodología de corrección solo considera un AR(1), en caso de no serlo, lo mas seguro es que no funcione. Bajo esta formulación se tiene que:

$$Y_t = \beta_0 + \beta_1 X_t + \epsilon_t, \epsilon_t = \rho\epsilon_{t-1} + a_t$$

Ahora considere la resta $Y_t - \rho Y_{t-1}$

$$Y_t - \rho Y_{t-1} = \beta_0 - \rho\beta_0 + \beta_1 X_t - \rho\beta_1 X_{t-1} + \epsilon_t - \rho\epsilon_{t-1}$$

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \beta_1(X_t - \rho X_{t-1}) + (\epsilon_t - \rho\epsilon_{t-1})$$

Cambiando de nombres se tiene un nuevo modelo de regresión:

$$\tilde{Y}_t = \tilde{\beta}_0 + \beta_1 \tilde{X}_t + a_t$$

donde su error aleatorio no tiene problemas de autocorrelacion!

Correción

Para el procedimiento anterior, se asumió que se conocía el valor ρ. Este debe estimarse. Para esto estime los residuales e_t del modelo $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$

- Despejar el valor a partir del estadístico de Durbin-Watson $DW \approx 2(1 - \rho)$
- Utilizar el coeficiente de correlacion muestral entre e_t y e_{t-1}
- A partir de un modelo de regresión sin intercepto entre los residuales:

$$e_t = \rho e_{t-1} + a_t$$

Por lo general la estimación mas preferible es la ultima.

Correción

Recapitulando, la metodología de corrección es:

- Estime el valor de ρ
- Cree las nuevas variables:

$$\tilde{Y}_t = Y_t - \rho Y_{t-1}$$

$$\tilde{X}_t = X_t - \rho X_{t-1} \ \ \forall j = 1, \cdots, k$$

- Estime la regresión en términos de las nuevas variables:

$$\tilde{Y}_t = \tilde{\beta}_0 + \beta_1 \tilde{X}_t + a_t$$