Supuestos Regresión Lineal: Especificación

Clase 30

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,

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

junto al supuesto de que los errores aleatorios:

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

asumimos que el modelo está bien planteado, es decir, la variable Y está determinada por las variables X que se incluyen en el modelo, y que estas realizan su aporte de forma lineal, junto al efecto aditivo de un error aleatorio con distribución normal.

Esto por lo general no es verdad.

Definición

En este caso, decimos que existe un error de especificación

Error de Especificación

El error de especificación esta asociado a un error en el planteamiento de la forma funcional del efecto de las variables regresoras, así como el error aleatorio

Cuando decimos forma funcional, nos referimos al modelo que se plantea, y de como incluye los efectos de los distintos factores sistemáticos y no sistemáticos en la variable de interés Y.

El error de especificación es INEVITABLE, pero podemos intentar en lo posible disminuir su efecto.

Definición

Existen diferentes formas en las que se puede tener un error de especificación:

- Omisión de variables relevantes: Puede que una variable de importancia no se incluya en el modelo de regresión, ya sea por que no se tiene información de esta, o el analista no lo considero necesario.
- Inclusión de variables NO relevantes: Se incluyen en el modelo variables que son inútiles en el objetivo de explicar el comportamiento de Y
- Linealidad: Se plantean relaciones lineales entre las variables, cuando realmente se deben plantear relaciones más flexibles, o al contrario.
- Error estocástico: Se asume que el error aleatorio tiene una distribución normal y su efecto es aditivo, cuando es necesario utilizar otra distribución de probabilidad o un efecto multiplicativo.

Outline

Implicaciones

Omisión de variables relevantes

Suponga que el modelo de regresión real es $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ pero se estima el modelo $Y = \beta_0 + \beta_1 X_1 + \tilde{\epsilon}$, luego $\tilde{\epsilon} = \beta_2 X_2 + \epsilon$. Es decir, los errores de la regresión estimada estarían capturando todo el efecto de la variable omitida X_2. Dado esto, los errores podrían ser heterocedasticos o autocorrelacionados. Por otra parte, los estimadores β_j serian SESGADOS.

Inclusión de variables irrelevantes

Suponga que el modelo de regresión real es $Y = \beta_0 + \beta_1 X_1 + \tilde{\epsilon}$ pero se estima el modelo $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$. Como X_2 no es relevante, el incluirla no disminuye considerablemente el SSE, pero si disminuye indistintamente sus grados de libertad $glE = n - p$, luego $MSE = SSE/glE$ aumenta. Los estimadores serían insesgados, pero su varianza aumenta!

Implicaciones

No linealidad

Si el modelo real tiene una relación no lineal, utilizar relaciones lineales simplemente no tiene sentido. Los parámetros β de la regresión solo son una aproximación al efecto en cambios que produce X sobre Y, en particular, los estimadores β_j serían SESGADOS pues en principio no tienen a que converger!

Error aleatorio

Los errores aleatorios no siempre siguen una distribución normal, aunque realmente no es muy importante por virtud de uno de los teoremas del limite central. El problema es que se PIERDE POTENCIA en las pruebas estadísticas

Implicaciones

- Los estimadores NO son consistentes.
- Las pruebas de significancia individual y global no son creíbles.
- 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.
- Peor aún, las pruebas de validación de supuestos de heterocedasticidad y autocorrelación también dejan de funcionar!

Outline

Identificación

La principal herramienta para determinar si existen problemas de Especificación es por medio de análisis gráficos y analíticos de los residuales.

Ejemplo

Suponga que el modelo de regresión teórico está dado por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

pero usted estima el modelo $Y = \beta_0 + \beta_3 X_3 + \tilde{\epsilon}$ entonces se tendría que:

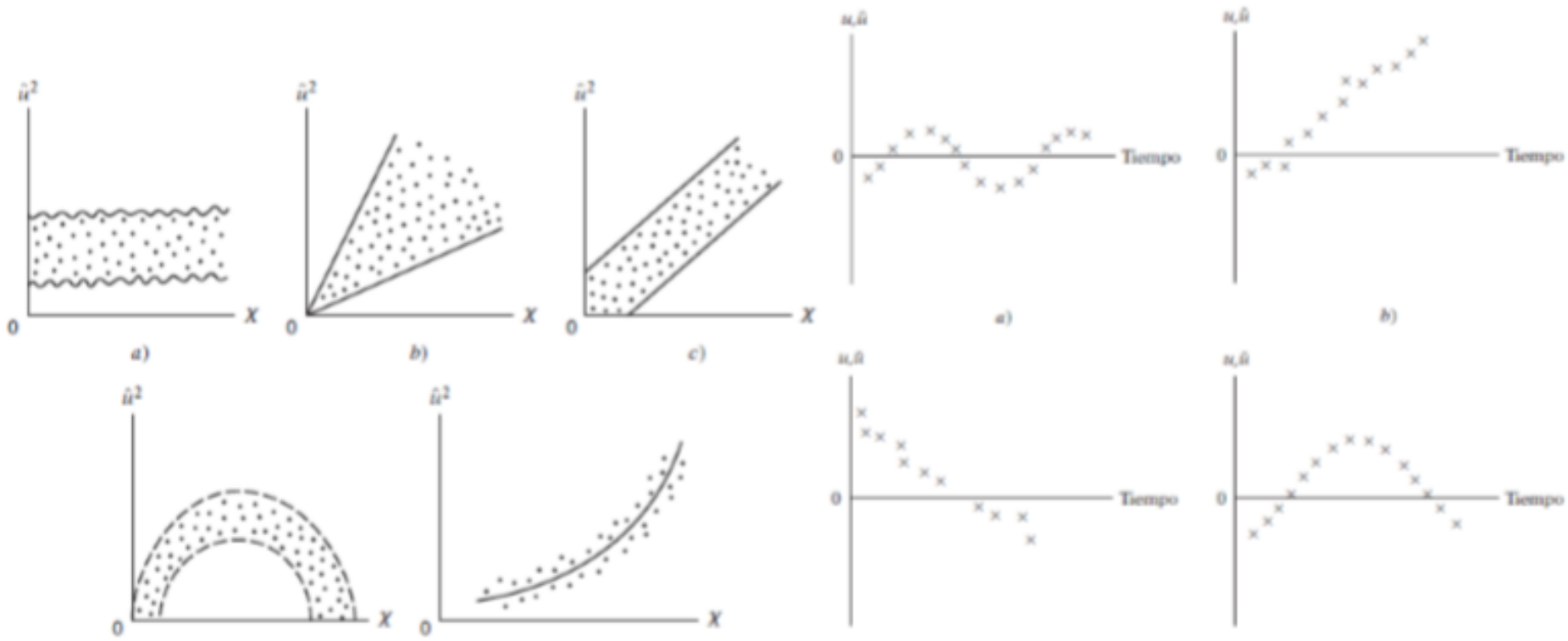
$$\tilde{\epsilon} = \beta_1 X_1 + \beta_2 X_2 - \beta_3 X_3 + \epsilon$$

es decir los errores no serían puramente aleatorios, sino que estarían en función de las variables X. Por lo tanto, si hay errores de especificación, los residuales e_i también tendrán tendencias en función de las variables.

Identificación

Métodos Gráficos

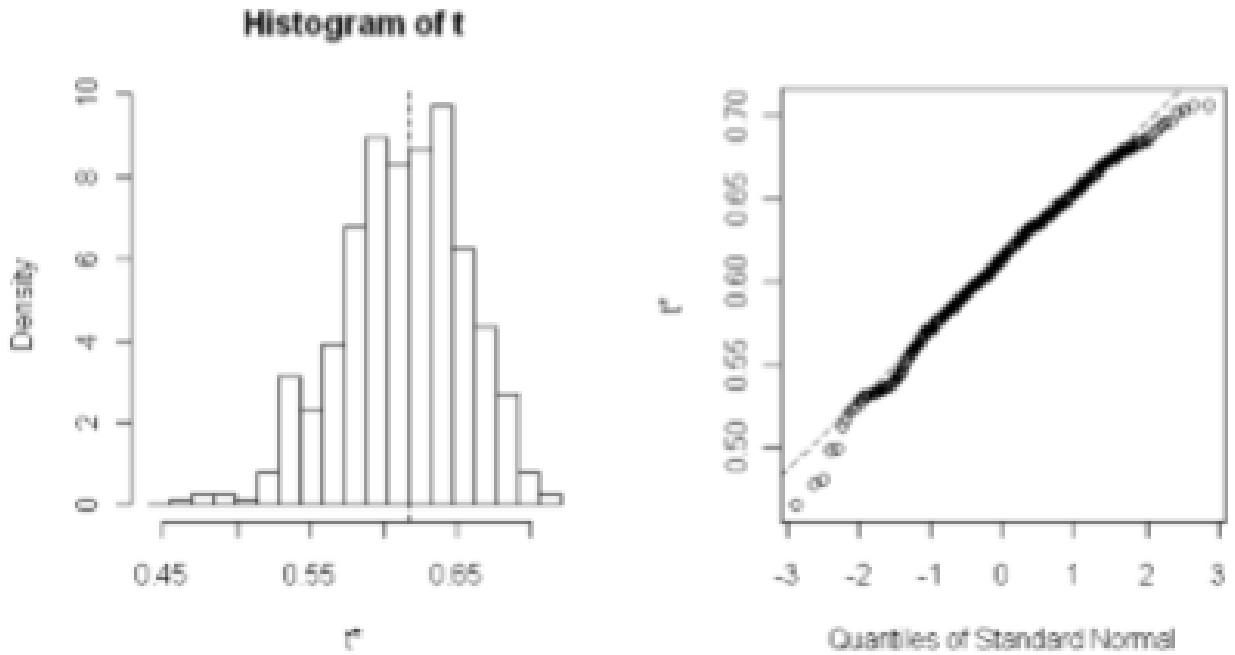
Si se grafican los e_i en función cualquier otra cantidad, estos deben mantener un comportamiento aleatorio. De lo contrario, hay problemas:



Identificación

Métodos Gráficos

Si se realiza un histograma de los e_i , estos deberían ser simétricos como la normal. El qqplot debería mostrar los residuales cerca a la linea:



Identificación

Métodos Anaíticos

Las pruebas estadísticas NO pueden probar la presencia de errores de Especificación de forma general. Realmente no existe prueba que permita determinarlo! Pero podemos basarnos en otras pruebas:

- Breush-Pagan: La prueba permite identificar si hay problemas de heterocedasticidad. Puede pasar que esta sea producto de variables omitidas en el modelo!
- Durbin-Watson: La prueba permite identificar si hay problemas de autocorrelación. Puede pasar que esta sea producto de variables omitidas en el modelo!
- Bondad de ajuste: Las pruebas de bondad de ajuste permiten ver si los residuales siguen la distribución de probabilidad adecuada

Identificación

Métodos Analíticos: Ramsey RESET test

Ramsey desarrolló la prueba *Regression Equation Specification Error Test (RESET)* para determinar si la forma funcional del modelo es adecuada, aunque es un test estadístico general para la especificación del modelo.

Suponga que estima el modelo $Y = \beta_0 + \beta_1 X_1 + \epsilon$. Si el modelo es verdadero, entonces su $\hat{Y}$ es el mejor estimador. Si estimamos una nueva regresión:

$$Y = \gamma_0 + \gamma_1 \hat{Y} + \gamma_2 \hat{Y}^2 + \gamma_3 \hat{Y}^3 + \nu = \beta_0 + \beta_1 X_1 + \gamma_2 \hat{Y}^2 + \gamma_3 \hat{Y}^3 + \nu$$

los términos NO lineales no serían signifcativos, pues $\hat{Y}$ ya es el mejor estimador. Pero si resultaran signifcativos, entonces el modelo original no es el mejor, y por tanto estaría mal especificado.

Identificación

Métodos Analíticos: Ramsey RESET test

$$H_0 : \text{No hay Error de Especificación}$$

$$H_1 : \text{Si hay Error de Especificación}$$

Es equivalente a probar:

$$H_0 : \gamma_2 = \gamma_3 = 0$$

$$H_1 : \text{Alguno diferente de } 0$$

$$Y = X^T\beta + \gamma_2\hat{Y}^2 + \gamma_3\hat{Y}_3 + \nu$$

donde $\hat{Y}$ es la estimacion del modelo $Y = X^T\beta + \epsilon$, siendo esta a su vez el modelo reducido. Esto se prueba con una F parcial

Outline

Correción

La metodología de corrección realmente no existe! Esta depende del contexto que se está.

Variables irrelevantes

Si el error de especificación está dado por tener variables de más,puede utilizar pruebas anidada como la F parcial o criterios de selección de modelos para encontrar un más adecuado.

Variables Omitidas

Si el error de especificación está dado porque faltan variables, considere la opción de agregar alguna otra variable. Si no se tiene información de la variable omitida, que es el caso interesante, evalúe que tan problemático es eso para su regresión

Correción

La metodología de corrección realmente no existe! Esta depende del contexto que se está.

Forma funcional

Si la forma funcional es problemática, pruebe relaciones NO lineales para sus variables X. En el peor de los caso, será necesario ajustar una regresión NO paramétrica.

Error Aleatorio

Si el error aleatorio está lejos de la normalidad, podría transformar su variable Y en búsqueda de dicha normalidad. Podría intentar con transformaciones como *Box-Cox*. Una alternativa es plantear un modelo lineal generalizado.

Correción

- El procedimiento de corrección no es tarea fácil y debe considerarse profundamente.
- Utilice siempre todas técnicas de identificación. No se quede solo con algunas.
- Nuevamente, la ESPECIFICACIÓN del modelo es algo imposible de lograr. Todos los modelos son incorrectos.
- Lo importante es que el modelo que usted plantee tenga sentido, y no pueda cuestionarse por cualquier perspectiva.
- Existen muchos otros modelos que no realizan supuestos tan fuertes. Es importante que siempre estén dispuestos a entender nuevos modelos.