Modelos de Regresión Aplicados

Laboratorio evaluado 2.

Este laboratorio evaluado se realiza en clases. Se puede hacer en grupos o individual. Tienen que entregarla el mismo 18 de noviembre antes de las 12:00 md. Si lo entregan posterior a las 12md, se les califica con base 90, y lo envían a gilbert.brenes.camacho@gmail.com. Pueden hacerlo en grupos de máximo 3 (idealmente, los mismos grupos del trabajo final).

Motivación de la tarea: Unos estadísticos afirman que la violación a la normalidad no afecta a los intervalos de confianza y las pruebas de hipótesis si la distribución de los errores es simétrica.

Objetivo de la tarea: Mostrar la validez del argumento de los estadísticos comparando la confianza de los intervalos de confianza con 4 distribuciones simétricas y una distribución asimétrica con variaciones en los parámetros.

Tome los datos que va a utilizar en el trabajo final. Estime un modelo lineal con su variable dependiente y dos predictoras. Utilice los valores de las estimaciones de β0, β1 y β2, y las distribuciones de las variables predictoras (o sea, puede escoger que sus variables se simulen con una normal, una uniforme, o cualquier otra distribución) para simular datos. O bien, puede tomar directamente las 2 variables predictoras de su base de datos. Use el tamaño de muestra de su trabajo final.

Suponga entonces que el modelo tiene la siguiente especificación:

Yi = β0 + β1\*X1i + β2\*X2i + εi

Esto quiere decir que tiene que simular la variable Yi.

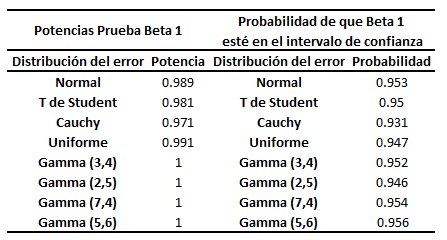
Haga 8 escenerios con base en la distribución del error. En otras palabras, simule el error con las siguientes 4 distribuciones continuas: normal, t de Student, Cauchy y uniforme (la media de los errores tiene que estar centrada en 0). Las funciones son rnorm(), rt, rcauchy y runif. Ustedes escogen el valor de los parámetros (investiguen en Wikipedia cuáles son los posibles valores que pueden escoger). Además, los otros 4 escenarios son con errores simulados con la distribución gamma; para ello, escoja a su gusto 4 combinaciones de los parámetros shape y rate (pero trate de que esos parámetros le generen distribuciones asimétricas).

Para que se cumpla el supuesto de que los errores están centrados en cero y que tengan variabilidad similar, entonces cuando crean el erorr con distribución gamma, métanlo dentro de la función scale() que lo estandariza, y lo multiplican por el error estándar de estimación que estimaron con el lm de sus datos. Ejemplo, scale(error)\*ee.

Con estos 8 escenarios, presente lo siguiente:

1. Usando repeticiones (ciclos), calcule la potencia de la prueba para β1 para los 8 escenarios
2. Usando repeticiones (ciclos), calcule la probabilidad de que el intervalo de confianza estimado para ^β1 contenga el verdadero valor del parámetro β1.

Presente los datos en forma ordenada en una especie de tabla comparativa; use 2 decimales para la potencia y para la probabilidad de que el intervalo estimado contenga al verdadero parámetro. Escriban en un párrafo si el argumento de los estadísticos es válido o no. Diga cómo afecta la asimetría en la distribución de los errores.

Esta evaluación se enmarca dentro del rubro “Trabajos en clase y quizes” que vale globalmente 10% de la nota total. El valor de cada evaluación va a depender de la cantidad de quizes y trabajos en clase que desarrollemos.

La potencia es la probabilidad de rechazar la hipótesis nula cuando la hipótesis nula es falsa. En este caso, se está utilizando la potencia como medida para calificar a las pruebas de hipótesis en su capacidad de determinar si se toma una decisión correcta o no . Se puede observar en los resultados que las distribuciones con errores basados en distribuciones no simétricas tienen potencias similares a aquellos con distribuciones simétricas. Se sabe que se está violando la normalidad, por lo que , de acuerdo a la teoría, no se le puede creer a las pruebas de hipótesis. Y, basado en los resultados, tampoco se puede saber si el que los errores tengan distribuciones simétricas permite asumir que la ausencia de normalidad no afecta a las pruebas de hipótesis. Sin embargo, el tener potencias altas, permite asumir que las pruebas de hipótesis están permitiendo rechazar cuando la hipótesis nula es falsa, incluso ante la ausencia de normalidad.

De acuerdo a los resultados obtenidos, se observa que las probabilidades de que Beta 1 esté en el intervalo de confianza, entre las distribuciones de errores simétricas y asimétricas, son muy similares. Esto no nos permite confirmar lo planteado por los estadísticos, que argumentan que la violación a la normalidad no afecta a los intervalos de confianza si la distribución de los errores es simétrica. El no poder llegar a conclusiones con respecto a lo planteado puede deberse a la naturaleza de los datos usados, pues, coincidentemente, los intervalos de confianza no paramétricos calculados a través de Bootstrap llegaban a las mismas conclusiones que los intervalos de confianza paramétricos. La principal diferencia es que los intervalos de confianza paramétricos eran más amplios, por lo que era esperable que las conclusiones entre ambos intervalos fueran similares (a pesar de que debido a la violación de supuestos no permite creerle a los intervalos de confianza).

En conclusión, para ambos casos y con este conjunto de datos específico, la violación al supuesto de normalidad no afecta las pruebas de hipótesis ni los intervalos de confianza, mas no se puede asumir que esto se deba a la distribución simétrica o no de los errores.