Resumen de lógica aplicada a los datos suministrados.

1. Limpieza: Se genero una validación previa a generar los modelos para cada base, se identificó que existían ciertos casos en los que habían datos faltantes o nulos, y otros casos en los que las variables eran totalmente ceros, por lo que para los modelos no fueron incluidas.
2. Análisis Exploratorio de Datos: Para este aparatado se valido tanto la logica de negocio como las correlaciones absolutas superiores al 70% para poder tener en cuenta aquellas variables que pudieran tener mayor relevancia.
3. Modelado: Se trabajo con 3 modelos diferentes
   1. En primer lugar, el modelo heckman de selección, este modelo se utiliza para analizar la relación entre una variable dependiente y una o más variables independientes, cuando la presencia de variables endógenas (aquellas que son influenciadas por la variable dependiente) puede afectar los resultados. El modelo heckman es una combinación de dos modelos: el modelo de selección y el modelo de regresión. El modelo de selección se utiliza para identificar las observaciones que participan en el modelo de regresión y el modelo de regresión se utiliza para analizar la relación entre la variable dependiente y las variables independientes.
   2. En segundo lugar, el modelo de selección que tiene como base la selección bivariada, que está formada por dos partes: la primera es el modelo de selección y la segunda es el modelo de relación.
   3. Por último, se contrastaron los resultados contra un modelo de efectos aleatorios.
4. Para el modelo de efectos aleatorios además, se incluyó una iteración sobre las variables preseleccionadas para intentar encontrar el mejor modelo posible tanto en significancia estadística de los estimadores como en explicación del modelo a través del R2.
5. Finalmente sobre los modelos realizados se generaron pruebas estadísticas para determinar que se cumplan los supuestos más relevantes:
   1. Normalidad de los residuos
   2. Homocedasticidad
   3. Independencia de los residuos

A continuación, se resumen los hallazgos para cada uno de los cultivos.

Cultivo de Café

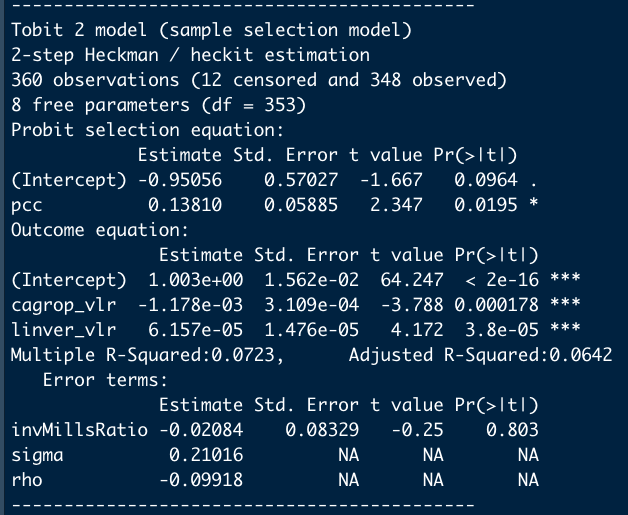
Se analizaron un total de 360 observaciones, de las cuales se contó con un total de 38 municipios en un periodo de tiempo de 9 años.

Para los modelos heckit (modelo de selección heckman) y el modelo de selección al modelar la probabilidad de acceso a crédito de forma independiente, se propone explicar este acceso a partir de la variable (pcc) que es la cantidad total de producto obtenido en el área cosechada de un municipio, si bien pueden haber factores externos que influyan sobre el acceso al crédito de cierto municipio, se trabajó con la información disponible en el dataset.

La anterior lógica solo modela la probabilidad de que un municipio pueda tener acceso a créditos, sin embargo, la variable dependiente principal a explicar es la productividad. Esta variable en el modelo heckit es explciada por:

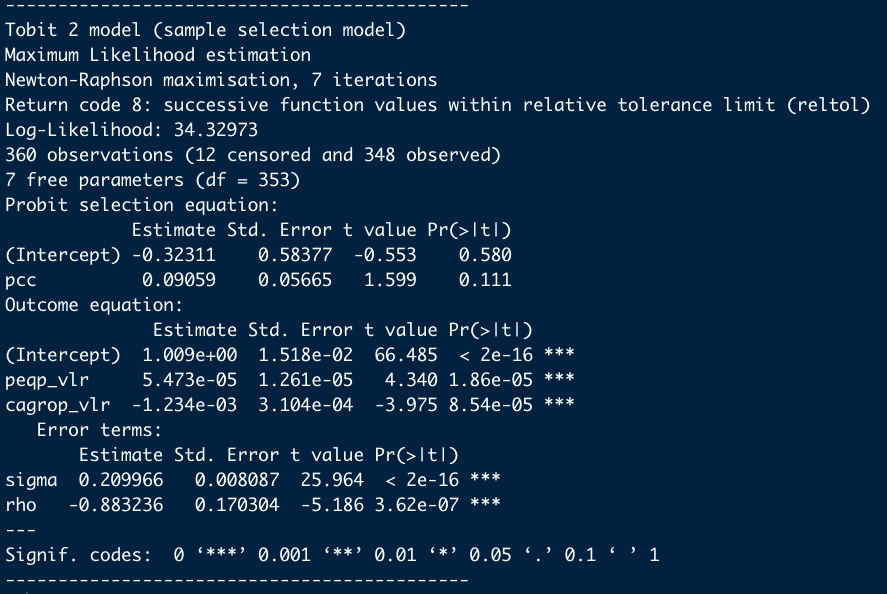
1. cagrop\_vlr: Cartera de crédito agropecuario.
2. linver\_vlr: Linea de crédito de inversión.

Para este modelo tanto para la probabilidad de acceso a crédito como las variables independientes para explicar la productividad terminan siendo significativas a mas del 95%.



Sin embargo, a pesar de que son significativas las variables, para este modelo el poder de explicación de estas variables sobre la productividad es del 7%.

En el modelo de selección de dos partes se obtuvieron los siguientes resultados



En este caso las variables significativas al 99.9% son:

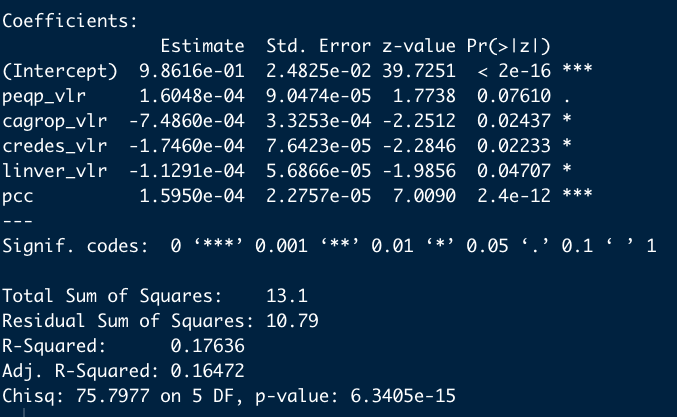
1. peqp\_vlr: Credito otorgado al pequeño productor
2. cagrop\_vlr: Cartera de crédito agropecuario.

Es de resaltar que este tipo de modelos no estiman directamente un R2 sino que se puede medir por el Log-Likelihood, el cual se calcula como la suma de los logaritmos de la función de verosimilitud evaluada en los datos observados. La función de verosimilitud describe la probabilidad de observar los datos dados el modelo y los parámetros estimados. Cuando el Log-Likelihood es alto, indica que el modelo se ajusta bien a los datos y que los parámetros estimados son coherentes con los datos observados.

En este caso, el Log-Likelihood para este modelo es de 34, lo que podría ser bajo en comparación a otros modelos, sin embargo, de las validaciones realizadas este es el mejor valor obtenido respetanto además la significancia de los estimadores.

Finalmente tal como se comentó al inicio, se generaron modelos de efectos aleatorios para comparar los resultados con respecto a los resultados de los dos modelos previos.

En resumen se testearon un total de 501 modelos distintos, una de las principales limitaciones fue que la librería plm tiene una función llamada del mismo modo que limita la cantidad de variables explicativas que puede llegar a tener un modelo, por lo que se limitó las pruebas a un máximo de 8 variables explicativas sobre la productividad, sin embargo, para este cultivo el mejor resultado estuvo en un 18% de R2.



Los estimadores estuvieron por encima del 90% de significancia.

Finalmente se incluye en el código las pruebas para estos modelos en los 3 principales supuestos, normalidad de los residuos, homocedasticidad e independencia de los residuos, de los cuales todos los modelos pasan estos test excepto, el modelo clásico, donde se identificó que los residuos no siguen una distribución normal.

Pruebas realizadas

1. La prueba de Shapiro-Wilk es una prueba estadística que se utiliza para determinar si un conjunto de datos sigue una distribución normal. La hipótesis nula de la prueba es que los datos provienen de una distribución normal, y el objetivo es determinar si existe suficiente evidencia para rechazar esa hipótesis. En R esta prueba se realiza a través de la función shapiro.test(). La sintaxis básica es la siguiente:
   1. X es el vector de datos que se desea analizar, en este caso los residuos del modelo, en este caso los residuos.

La interpretación de los resultados de esta prueba son los siguientes: si el p-value es menor que un nivel de significancia establecido previamente (por ejemplo, 0.05 o lo que es lo mismo a un nivel de confianza del 95%), se rechaza la hipótesis nula y se concluye que los datos no provienen de una distribución normal. Si el p-value es mayor que el nivel de significancia, no se puede rechazar la hipótesis nula y se concluye que los datos provienen probablemente de una distribución normal.

1. La prueba de Breusch-Pagan es una prueba estadística que se utiliza para determinar si existe heterocedasticidad en un modelo de regresión lineal. La heterocedasticidad se refiere a una situación en la que la varianza de los residuos no es constante a lo largo del rango de los predictores.

Para interpretar los resultados, se puede analizar los siguiente: si el p-value es menor que un nivel de significancia establecido (por ejemplo, 0.05 en este caso el 95% de significancia), se rechaza la hipótesis nula y se concluye que existe heterocedasticidad en el modelo. Si el p-value es mayor que el nivel de significancia, no se puede rechazar la hipótesis nula y se concluye que la varianza de los residuos es probablemente constante a lo largo del rango de los predictores.

1. La prueba de Box es una prueba estadística que se utiliza para detectar presencia de autocorrelación. La prueba se basa en el cálculo de la estadística de Box-Ljung y la comparación de dicha estadística con una distribución de chi-cuadrado con un grado de libertad específico.

La función Box.test() en R devuelve una estadística de prueba y un valor p, que indican la probabilidad de obtener una estadística de prueba tan extrema o más extrema que la obtenida si la hipótesis nula de ausencia de autocorrelación fuera cierta. Un valor p menor que el nivel de significancia establecido (por ejemplo, 0.05 o lo que es lo mismo al 95% de significancia) indica que se rechaza la hipótesis nula y se concluye que existe autocorrelación en los datos.

De forma resumida se incluyen los resultados de los modelos para los demás cultivos:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cultivo** | **Modelo** | **R2** | **Log-Likelihood** | **Estimadores significativos** | **Test 95%** |
| Café | heckit | 7% |  | cagrop\_vlr +l inver\_vlr | No pasa los tests |
| Café | selection |  | 34,32 | peqp\_vlr + cagrop\_vlr | No pasa los tests |
| Café | clasico | 18% |  | peqp\_vlr + cagrop\_vlr + credes\_vlr + linver\_vlr + pcc | Pasa Hetercedasticidad e independencia |
| Caña | heckit | 46% |  | linver\_vlr | No pasa los tests |
| Caña | selection |  | -1.613 | linver\_vlr + lnorm\_vlr | No pasa los tests |
| Caña | clasico | 29% |  | qtotal\_cred + lkdtrab\_vlr + lnorm\_vlr + pcc | No pasa los tests |
| Papa | heckit | 20% |  | peqp\_vlr + linver\_vlr + lnorm\_vlr | No pasa los test |
| Papa | selection |  | -861,1027 | peqp\_vlr + linver\_vlr + lnorm\_vlr | Pasa homocedasticidad e normalidad |
| Papa | clasico | 27% |  | qtotal\_cred + peqp\_vlr + csustit\_vlr + lkdtrab\_vlr | Pasa homocedasticidad e independencia |
| Plantano | heckit | 22% |  | no significativas | No pasa los test |
| Plantano | selection |  | -862,9155 | no significativas | No pasa los test |
| Plantano | clasico | 6% |  | peqp\_vlr + credes\_vlr | Pasa homocedasticidad e independencia |

De los resultados anteriores, se puede mencionar que para los datos recolectados para el cultivo de plátano, existe un bajo poder explicativo de la productividad por el acceso al crédito de los municipios analizados, en mejor modelo explica el 22% del comportamiento sin embargo, no son significativas las variables explicativas.

Por otro lado, para el cultivo de la papa, en el modelo de selección y del mismo modo para el modelo de caña se identificó un log-likelihood negativo, esto puede significar que si bien los estimadores sean significativos, el modelo no se ajusta bien a los datos y que los parámetros estimados no son coherentes con los datos observados.

Interpretación de modelos

1. Modelo Heckit para café:

Lo que se puede interpretar de este modelo es:

* 1. Para este modelo un aporte de cero cartera agropecuaria y línea de inversión representa un incremento superior al 1 puntos en el índice de productividad.
  2. Existe una relación negativa débil entre la productividad y la cartera de crédito agropecuario, por lo que el incremento de 1 millon en la cartera de crédito agropecuario representa un reducción de la productividad de 0.001178 de los cultivos de café.
  3. Por otro lado, existe una relación positiva débil entre la productividad y la línea de crédito de inversión, donde un incremento de 1 millon en la línea de crédito de inversión representa un incremento inferior a 1 sobre el índice de productividad.

1. Modelo Selection para café:

Lo que se puede interpretar de este modelo es:

* 1. Para este modelo un aporte de cero cartera agropecuaria y línea de inversión representa un incremento superior a 1 puntos en el índice de productividad.
  2. Existe una relación positiva débil entre la productividad y la línea de crédito de inversión, por lo que el incremento de 1 millón en la línea de crédito de inversión representa un incremento de la productividad menor a un punto de los cultivos de café.
  3. Por otro lado, existe una relación negativa débil entre la productividad y la cartera de crédito agropecuario, donde un incremento de 1 millon en la cartera de crédito agropecuario representa una reducción inferior a 1 punto sobre el índice de productividad.

1. Modelo Clásico para café:
   1. La no intervención de crédito para este modelo representa un incremento de 0.9 puntos en el índice de productividad.
   2. El incremento de un millón del crédito otorgado al pequeño productor, representa un incremento del 0.0001 en el índice de productividad.
   3. El incremento de un millón del crédito de cartera agropecuario, representa una reducción del 0.0007 del índice de productividad.
   4. El incremento de un millón de la cartera de redescuentos, representa una reducción de la productividad de 0.0001 puntos.
   5. El incremento de un millón de la línea de crédito de inversión, representa una reducción de la productividad del 0.0001 puntos.
   6. Finalmente, un incremento de una tonelada obtenida de siembra representa un incremento del 0.0001 puntos en la productividad.

Es de resaltar que los modelos obtenidos tienen un bajo poder explicativo sobre la variable independiente, lo que podría indicar que los datos disponibles pueden no ser suficientes para capturar de forma correcta la relación entre las variables.