Análisis de Variable Instrumental en Homeless

Mateo, Sebastián, Genaro

Problema

En numerosos condados a lo largo de Estados Unidos, el incremento en los costos de vivienda ha sido una preocupación creciente, no solo por su impacto directo en la asequibilidad de la vivienda, sino también por sus posibles efectos en el aumento de la población sin hogar. Identificar las dinámicas precisas entre estos factores es crucial para desarrollar políticas efectivas de vivienda y programas de asistencia social. Sin embargo, la relación entre el costo de la vivienda y el número de personas sin hogar es compleja y puede estar confundida por variables omitidas, como el desempleo o las políticas locales, lo que dificulta obtener estimaciones causales claras.

Nuestra hipótesis plantea que el costo promedio de renta en un condado tiene un impacto significativo en la cantidad de personas sin hogar en ese mismo condado. Para explorar esta relación de manera más precisa y controlar la posible endogeneidad de la variable de costo de vivienda, utilizamos el costo promedio de renta de los condados colindantes en el periodo de medida anterior como variable instrumental. Esperamos demostrar que, al aislar la influencia de otros factores, un aumento en el costo de la renta contribuye directamente al incremento en el número de personas sin hogar, proporcionando así evidencia que podría ser utilizada para guiar intervenciones y políticas públicas efectivas.

Variables Instrumentales

En análisis estadístico, a menudo nos enfrentamos al problema de la endogeneidad, que ocurre cuando hay una correlación entre una variable explicativa X y el término de error U. Esta correlación puede surgir de una variable omitida que influye simultáneamente en X, y en la variable dependiente Y, generando estimaciones sesgadas y conclusiones erróneas en modelos de regresión lineal. Para abordar esta complicación, uno de los enfoques más efectivos es el uso de variables instrumentales. Este método nos permite aislar el efecto verdadero de X sobre Y, proporcionando una estimación más precisa de la relación causal.

Definición y supuestos

Una variable Z se considera instrumental si cumple con los siguientes supuestos:

- Relevancia: Z debe estar correlacionada con la variable explicativa X, pero no necesariamente con la variable dependiente Y (Que Z tenga efecto sobre Y). Esto se verifica a través de la condición:

$$cov(X_i, Z_i) \neq 0$$

- Exogeneidad: Z no debe estar correlacionada con el término de error U, garantizando que no está afectada por las variables omitidas que influyen en Y. Esto asegura que los efectos estimados no están sesgados por variables no observadas:

$$cov(U_i, Z_i) = 0$$

- Restricción de Exclusión: Z solo debe influir en Y a través de X y no debe tener ningún efecto directo o a través de otros canales no observados. Esta condición es crucial para asegurar que la relación causal que medimos es exclusivamente a través de X:

$$cov(Y_i, Z_i|X_i) = 0$$

Datos

```r
data <- read.csv("./metodos_lineales_data.csv")
#data[is.na(data$Previous_CoC_Neighbor_Median_Rent), ]
# los primeros periodos están en null por lo que para hacer la regresión filtramos
data <- na.omit(data, cols = "Previous_CoC_Neighbor_Median_Rent")
summary(lm(Overall_Homeless ~ Previous_CoC_Neighbor_Median_Rent, data = data))
```

```
Call:
lm(formula = Overall_Homeless ~ Previous_CoC_Neighbor_Median_Rent,
    data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -5909  -1078   -489    149  75360
```

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1412.600    268.700   -5.26  1.6e-07 ***
Previous_CoC_Neighbor_Median_Rent     4.012      0.354   11.33  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4780 on 3378 degrees of freedom
Multiple R-squared:  0.0366,    Adjusted R-squared:  0.0363
F-statistic:  128 on 1 and 3378 DF,  p-value: <2e-16
```

- `Year`: Año del registro. (2013 - 2022)
- `CoC_Number`: Identificador numérico del Continuum of Care (CoC).
- `CoC_Name`: Nombre descriptivo del CoC.
- `Overall_Homeless`: Cantidad total de personas sin hogar.
- `Chronic_Homeless`: Cantidad de personas sin hogar crónicamente.
- `Non-Chronic_Homeless`: Cantidad de personas sin hogar no crónicamente.
- `Actual_CoC_Median_Rent`: Renta media en el área del CoC.
- `Previous_CoC_Neighbor_Median_Rent`: Renta media en los CoCs vecinos.
- `State`: Estado en EE.UU. donde se encuentra el Continuum of Care (CoC).
- `CoC_AVG_Median_Household_Income`: Ingreso medio del hogar promedio en el área del CoC.
- `CoC_Population_Estimate`: Estimación de la población total en el área del CoC.
- `Poverty_Percentage`: Porcentaje de la población bajo el umbral de pobreza en el área del CoC.
- `CoC_Civilian_Labor_Force`: Tamaño de la fuerza laboral civil en el área del CoC.
- `Unemployment_Rate`: Tasa de desempleo en el área del CoC.
- `Party_Affiliation`: Afiliación política predominante en el área del CoC.
- `Vote_Percentage`: Porcentaje de votos obtenidos por el partido predominante en las últimas elecciones.
- `CoC_Poverty_Estimate_Age_0_17`: Estimación del número de menores de 18 años en situación de pobreza en el área del CoC.
- `Birth_Rate`: Tasa de natalidad en el área del CoC.
- `Death_Rate`: Tasa de mortalidad en el área del CoC.

- `International_Migration_Rate`: Tasa de migración internacional en el área del CoC.
- `Domestic_Migration_Rate`: Tasa de migración doméstica dentro del país en el área del CoC.

Agregamos una varibale dummy llamada `Party_Democrat`

```
data$Party_Democrat <- as.integer(data$Party_Affiliation == 'Democrat')
```

```
str(data)
```

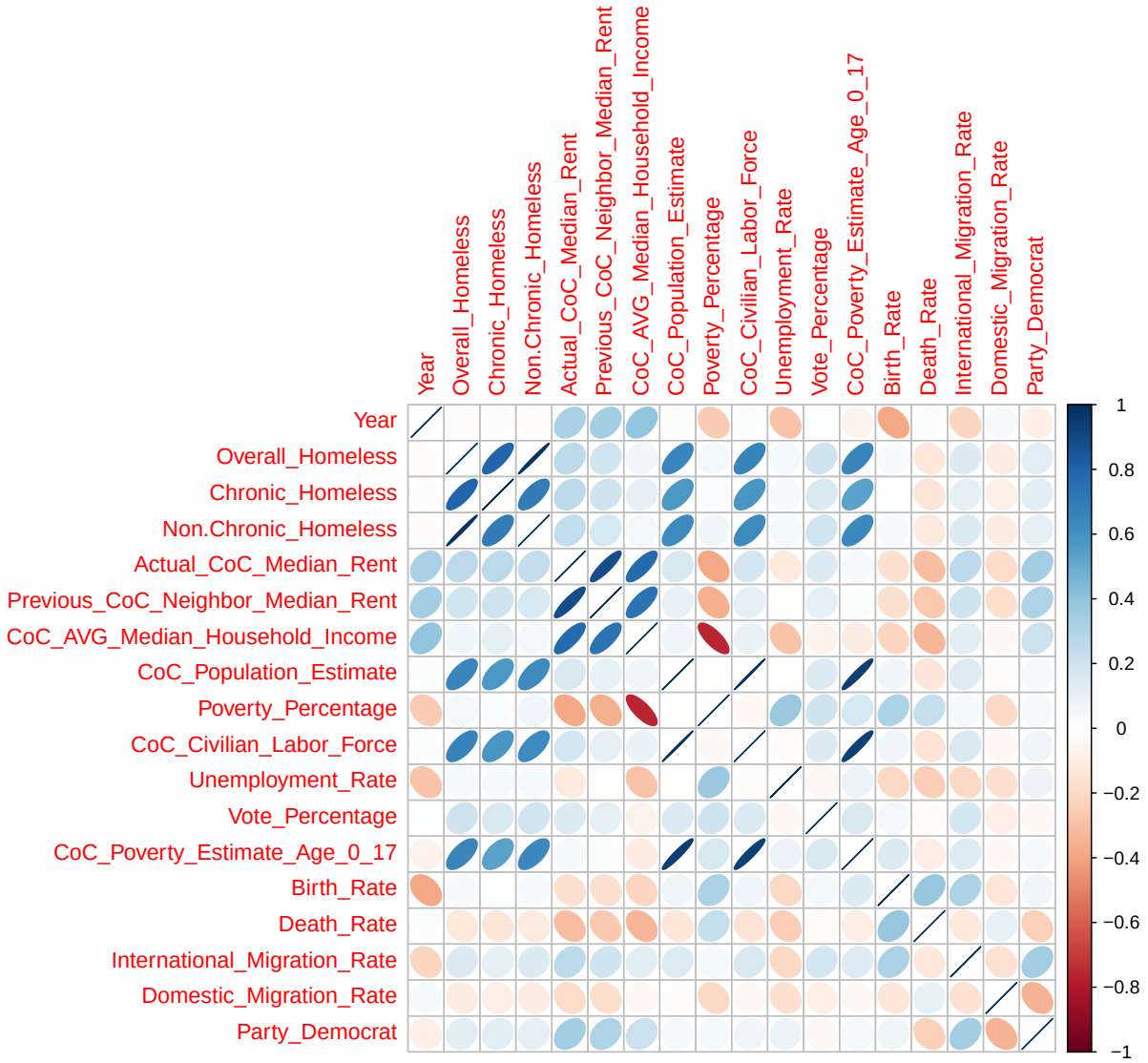
```
'data.frame':   3380 obs. of  22 variables:
 $ Year                              : int  2022 2022 2022 2022 2022 2022 2022 2022 2022 2022
 $ CoC_Number                        : chr  "AL-500" "AL-501" "AL-502" "AL-503" ...
 $ CoC_Name                          : chr  "Birmingham/Jefferson, St. Clair, Shelby Counties
 $ State                             : chr  "AL" "AL" "AL" "AL" ...
 $ Overall_Homeless                  : int  943 585 232 549 278 40 935 859 343 814 ...
 $ Chronic_Homeless                  : int  267 76 11 118 57 0 22 290 49 111 ...
 $ Non.Chronic_Homeless              : int  676 509 221 431 221 40 913 569 294 703 ...
 $ Actual_CoC_Median_Rent            : num  834 783 554 682 674 ...
 $ Previous_CoC_Neighbor_Median_Rent: num  728 636 551 534 539 ...
 $ CoC_AVG_Median_Household_Income   : num  61996 63010 52939 75282 44976 ...
 $ CoC_Population_Estimate           : int  665084 657929 271888 638704 246227 236690 1383394
 $ Poverty_Percentage                : num  15.9 16.4 15.7 11.1 18.5 ...
 $ CoC_Civilian_Labor_Force          : int  323290 295161 120142 308386 115243 103432 591702 1
 $ Unemployment_Rate                 : num  0.0258 0.0293 0.0262 0.0212 0.0313 ...
 $ Party_Affiliation                 : chr  "Democrat" "Republican" "Republican" "Republican"
 $ Vote_Percentage                   : num  0.516 0.627 0.746 0.614 0.628 ...
 $ CoC_Poverty_Estimate_Age_0_17     : int  31449 33537 12649 19794 16108 9540 69752 22117 180
 $ Birth_Rate                        : num  11.9 11.5 10.8 11 13.1 ...
 $ Death_Rate                        : num  13.2 13.1 15.2 10.9 12.2 ...
 $ International_Migration_Rate      : num  0.952 0.813 0.427 1.359 1.474 ...
 $ Domestic_Migration_Rate           : num  -6.72 8.37 10 15.78 -5.23 ...
 $ Party_Democrat                    : int  1 0 0 0 1 0 0 1 0 0 ...
```

```
summary(data)
```

```
      Year       CoC_Number          CoC_Name            State           Overall_Homeless Chr
 Min.   :2013   Length:3380        Length:3380        Length:3380        Min.   :   23    Min
 1st Qu.:2015   Class :character   Class :character   Class :character   1st Qu.:  310    1st
 Median :2018   Mode  :character   Mode  :character   Mode  :character   Median :  610    Med
 Mean   :2018                                                           Mean   : 1485    Mea
```

```
 3rd Qu.:2020                                                          3rd Qu.: 1351    3rd
 Max.   :2022                                                          Max.   :78676    Max
 Previous_CoC_Neighbor_Median_Rent CoC_AVG_Median_Household_Income CoC_Population_Estimate Po
 Min.   : 387                      Min.   : 30141                  Min.   :    5791        Mi
 1st Qu.: 556                      1st Qu.: 49215                  1st Qu.:  261847        1s
 Median : 658                      Median : 57575                  Median :  495740        Me
 Mean   : 722                      Mean   : 61497                  Mean   :  811826        Me
 3rd Qu.: 828                      3rd Qu.: 69529                  3rd Qu.:  843110        3r
 Max.   :2158                      Max.   :150502                  Max.   :10264268        Ma
 Party_Affiliation  Vote_Percentage CoC_Poverty_Estimate_Age_0_17   Birth_Rate      Death_Rat
 Length:3380        Min.   :0.420   Min.   :   491                Min.   : 1.36   Min.   : 1
 Class :character   1st Qu.:0.541   1st Qu.:  8631                1st Qu.: 9.69   1st Qu.: 7
 Mode  :character   Median :0.589   Median : 17174                Median :11.09   Median : 9
                    Mean   :0.600   Mean   : 33415                Mean   :10.60   Mean   : 8
                    3rd Qu.:0.646   3rd Qu.: 33459                3rd Qu.:12.45   3rd Qu.:10
                    Max.   :0.853   Max.   :624784                Max.   :21.77   Max.   :19
 Party_Democrat
 Min.   :0.000
 1st Qu.:0.000
 Median :1.000
 Mean   :0.521
 3rd Qu.:1.000
 Max.   :1.000
```

```r
correlation_matrix <- cor(data %>% select_if(is.numeric))
corrplot(correlation_matrix, method = "ellipse")
```



Análisis

1. Regresión simple

```
model <- lm(Overall_Homeless ~ Actual_CoC_Median_Rent, data = data)
summary(model)
```

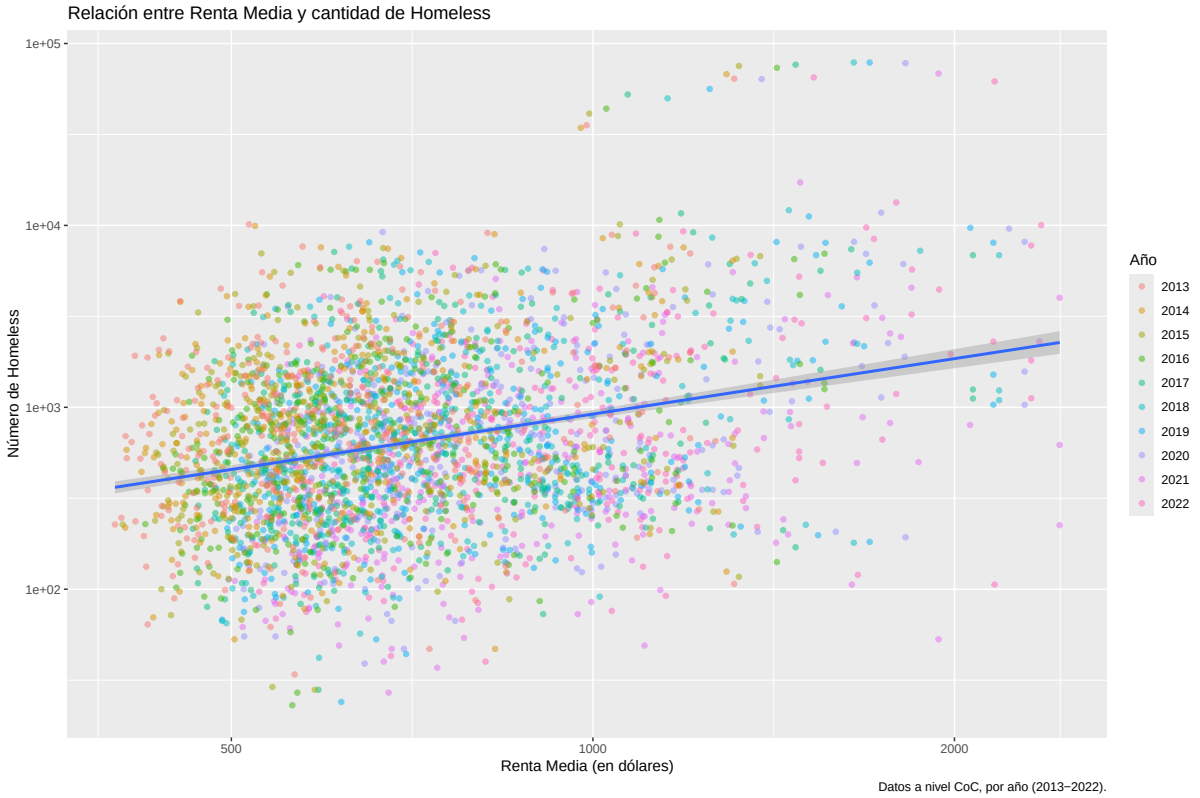
```
Call:
```

```
lm(formula = Overall_Homeless ~ Actual_CoC_Median_Rent, data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -8735  -1048   -337    305  73258

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)            -1902.651    229.779   -8.28   <2e-16 ***
Actual_CoC_Median_Rent     4.439      0.282   15.75   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4700 on 3378 degrees of freedom
Multiple R-squared:  0.0684,    Adjusted R-squared:  0.0681
F-statistic:  248 on 1 and 3378 DF,  p-value: <2e-16
```

En este modelo, el coeficiente de `Actual_CoC_Median_Rent` es significativamente positivo, sugiriendo que un incremento de $1 en la renta media del condado está asociado con un aumento de aproximadamente 4.439 personas sin hogar, lo que indica una relación directa entre el costo de vivienda y la falta de vivienda. El alto nivel de significancia estadística (`p < 2e-16`) confirma la robustez de esta relación. Sin embargo, el bajo R-cuadrado de 0.0684 implica que este modelo solo explica un 6.84% de la variabilidad en la falta de vivienda, señalando que otros factores no considerados en el modelo también son relevantes.

```
# Gráfica 1: Relación entre X:"Renta Media" y Y:"Cantidad Homeless"
# Representación visual de la relación entre la renta media por Continuum of Care (CoC)
# y el número de personas sin hogar (en miles) para cada año entre 2013 y 2022.

ggplot(data = data, aes(x = Actual_CoC_Median_Rent, y = Overall_Homeless)) +
    geom_point(aes(color=as.factor(Year)),alpha = 0.5) +  # Ajusta la transparencia
    geom_smooth(method = "lm", se = TRUE, aes(group=1)) +  # Añade intervalos de confianza
    scale_y_log10()+
    scale_x_log10()+
  labs(title = "Relación entre Renta Media y cantidad de Homeless",
       x = "Renta Media (en dólares)",
       y = "Número de Homeless",
       color = "Año",
       caption = "Datos a nivel CoC, por año (2013-2022).")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



- Puntos en color para representar diferentes años. Podemos observar que la renta es mayor en los años más recientes.
- Línea de tendencia sugiere una correlación positiva entre la renta media y la cantidad de personas sin hogar.
- Por la escala logaritmica, las diferencias visuales son más pronunciadas y la intensidad de la tendencia puede verse afectada, pero existe relación.

2. Regresión con todas las variables

```
# Multiple linear regression with several predictors
model2 <- lm(Overall_Homeless ~ Actual_CoC_Median_Rent +
               CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
               Poverty_Percentage + CoC_Civilian_Labor_Force +
               Unemployment_Rate + Vote_Percentage +
               CoC_Poverty_Estimate_Age_0_17 + Birth_Rate +
               Death_Rate + International_Migration_Rate + Domestic_Migration_Rate,
             data = data)
summary(model2)
```

```
Call:
lm(formula = Overall_Homeless ~ Actual_CoC_Median_Rent + CoC_AVG_Median_Household_Income +
    CoC_Population_Estimate + Poverty_Percentage + CoC_Civilian_Labor_Force +
    Unemployment_Rate + Vote_Percentage + CoC_Poverty_Estimate_Age_0_17 +
    Birth_Rate + Death_Rate + International_Migration_Rate +
    Domestic_Migration_Rate, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-21122   -711    223    851  58551

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.04e+03   8.87e+02   -2.30  0.02153 *
Actual_CoC_Median_Rent           5.45e+00   4.00e-01   13.64  < 2e-16 ***
CoC_AVG_Median_Household_Income -5.97e-02   9.05e-03   -6.60  4.6e-11 ***
CoC_Population_Estimate         -1.65e-02   9.47e-04  -17.42  < 2e-16 ***
Poverty_Percentage              -3.92e+01   2.75e+01   -1.43  0.15286
CoC_Civilian_Labor_Force         3.23e-02   1.76e-03   18.40  < 2e-16 ***
Unemployment_Rate                6.12e+03   3.35e+03    1.83  0.06761 .
Vote_Percentage                  3.01e+03   8.14e+02    3.70  0.00022 ***
CoC_Poverty_Estimate_Age_0_17    7.25e-02   4.28e-03   16.92  < 2e-16 ***
Birth_Rate                      -1.00e+02   2.31e+01   -4.35  1.4e-05 ***
Death_Rate                       8.58e+01   2.38e+01    3.60  0.00032 ***
International_Migration_Rate    -1.12e+02   3.21e+01   -3.47  0.00053 ***
Domestic_Migration_Rate          6.65e+00   6.64e+00    1.00  0.31655
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3310 on 3367 degrees of freedom
Multiple R-squared:  0.539, Adjusted R-squared:  0.538
F-statistic:  328 on 12 and 3367 DF,  p-value: <2e-16
```

Este modelo más complejo incorpora varias variables adicionales y muestra que el costo de renta sigue siendo un predictor significativo del número de personas sin hogar, con un coeficiente aumentado a 5.45, lo cual es estadísticamente significativo (`p < 2e-16`). El incremento en R-cuadrado a 0.539 indica que el modelo ahora explica aproximadamente el 53.9% de la variabilidad en la falta de vivienda, mostrando una mejora considerable respecto al modelo simple. Esto demuestra la importancia de considerar múltiples factores como ingresos del hogar, tamaño de la población y la fuerza laboral en el análisis de la falta de vivienda.

3. Regresión con variables significativas

```r
# Multiple linear regression with several predictors
model3 <- lm(Overall_Homeless ~ Actual_CoC_Median_Rent +
               CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
               CoC_Civilian_Labor_Force +
               Vote_Percentage +
               CoC_Poverty_Estimate_Age_0_17 + Birth_Rate +
               Death_Rate + International_Migration_Rate,
             data = data)
summary(model3)
```

```
Call:
lm(formula = Overall_Homeless ~ Actual_CoC_Median_Rent + CoC_AVG_Median_Household_Income +
    CoC_Population_Estimate + CoC_Civilian_Labor_Force + Vote_Percentage +
    CoC_Poverty_Estimate_Age_0_17 + Birth_Rate + Death_Rate +
    International_Migration_Rate, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-20819   -676    216    857  58723

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -2.02e+03   6.23e+02   -3.24  0.00119 **
Actual_CoC_Median_Rent           5.25e+00   3.51e-01   14.99  < 2e-16 ***
CoC_AVG_Median_Household_Income -5.33e-02   6.21e-03   -8.59  < 2e-16 ***
CoC_Population_Estimate         -1.62e-02   9.27e-04  -17.48  < 2e-16 ***
CoC_Civilian_Labor_Force         3.19e-02   1.71e-03   18.59  < 2e-16 ***
Vote_Percentage                  2.69e+03   7.99e+02    3.37  0.00077 ***
CoC_Poverty_Estimate_Age_0_17    7.13e-02   4.13e-03   17.27  < 2e-16 ***
Birth_Rate                      -1.15e+02   2.22e+01   -5.19  2.2e-07 ***
Death_Rate                       7.52e+01   2.24e+01    3.36  0.00079 ***
International_Migration_Rate    -1.22e+02   3.15e+01   -3.88  0.00011 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3310 on 3370 degrees of freedom
Multiple R-squared:  0.539, Adjusted R-squared:  0.537
F-statistic:  437 on 9 and 3370 DF,  p-value: <2e-16
```

Al ajustar el modelo solo con variables que mostraron significancia estadística, se mantiene un alto R-cuadrado de 0.539, lo que indica que las variables seleccionadas retienen su ca-

pacidad explicativa sin la carga de variables no significativas. El coeficiente para el costo de renta ajustada muestra un impacto de 5.25 personas sin hogar por cada aumento de $1 en la renta, subrayando la persistente influencia significativa del costo de vivienda (`p < 2e-16`). La precisión del modelo, reflejada por un error estándar residual de 3310, y su ajuste mejorado resaltan la relevancia de estas variables en estudios de falta de vivienda.

Variable Instrumental

4. Regresión simple

```
# First Stage: Regress Actual_CoC_Median_Rent on Previous_CoC_Neighbor_Median_Rent and other
first_stage_model <- lm(Actual_CoC_Median_Rent ~ Previous_CoC_Neighbor_Median_Rent,
                        data = data)
data$fitted_Actual_CoC_Median_Rent <- fitted(first_stage_model)

# Second Stage: Regress Overall_Homeless on the fitted values from the first stage and other
second_stage_model <- lm(Overall_Homeless ~ fitted_Actual_CoC_Median_Rent,
                         data = data)
summary(second_stage_model)
```

```
Call:
lm(formula = Overall_Homeless ~ fitted_Actual_CoC_Median_Rent,
    data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -5909  -1078   -489    149  75360

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -1308.599    259.973   -5.03  5.1e-07 ***
fitted_Actual_CoC_Median_Rent     3.661      0.323   11.33  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4780 on 3378 degrees of freedom
Multiple R-squared:  0.0366,    Adjusted R-squared:  0.0363
F-statistic:  128 on 1 and 3378 DF,  p-value: <2e-16
```

Este modelo utiliza la renta actual predicha, ajustada por la renta de condados vecinos, como una variable instrumental para aislar el efecto del costo de la vivienda sobre la falta de vivienda. El coeficiente de `fitted_Actual_CoC_Median_Rent` es de 3.661, lo que indica que un aumento de $1 en la renta media ajustada se asocia con un aumento de aproximadamente 3.661 en el número de personas sin hogar, con una significancia estadística extremadamente alta (`p < 2e-16`). Aunque el modelo tiene un R-cuadrado bajo (0.0366), esto sugiere que otras variables no incluidas podrían estar influyendo en la falta de vivienda. El modelo confirma la relevancia del costo de la vivienda pero también destaca la necesidad de considerar otros factores o de aplicar correcciones para los errores estándar para mejorar la estimación.

5. Regresión con todas las variables

```r
# First Stage: Regress Actual_CoC_Median_Rent on Previous_CoC_Neighbor_Median_Rent and other
first_stage_model2 <- lm(Actual_CoC_Median_Rent ~ Previous_CoC_Neighbor_Median_Rent +
                         CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
                         Poverty_Percentage + CoC_Civilian_Labor_Force +
                         Unemployment_Rate + Vote_Percentage +
                         CoC_Poverty_Estimate_Age_0_17 + Birth_Rate + Death_Rate +
                         International_Migration_Rate + Domestic_Migration_Rate,
                       data = data)
data$fitted_Actual_CoC_Median_Rent2 <- fitted(first_stage_model2)

# Second Stage: Regress Overall_Homeless on the fitted values from the first stage and other
second_stage_model2 <- lm(Overall_Homeless ~ fitted_Actual_CoC_Median_Rent2 +
                          CoC_AVG_Median_Household_Income + CoC_Population_Estimate + Poverty_
                          CoC_Civilian_Labor_Force + Unemployment_Rate + Vote_Percentage + CoC
                          Birth_Rate + Death_Rate + International_Migration_Rate + Domestic_Mi
summary(second_stage_model2)
```

```
Call:
lm(formula = Overall_Homeless ~ fitted_Actual_CoC_Median_Rent2 +
    CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
    Poverty_Percentage + CoC_Civilian_Labor_Force + Unemployment_Rate +
    Vote_Percentage + CoC_Poverty_Estimate_Age_0_17 + Birth_Rate +
    Death_Rate + International_Migration_Rate + Domestic_Migration_Rate,
    data = data)

Residuals:
   Min     1Q Median     3Q    Max
-21272   -648    176    825  61147
```

```
Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.97e+03   1.01e+03   -1.95  0.05170 .
fitted_Actual_CoC_Median_Rent2   5.53e+00   6.26e-01    8.82  < 2e-16 ***
CoC_AVG_Median_Household_Income -6.10e-02   1.25e-02   -4.89  1.0e-06 ***
CoC_Population_Estimate         -1.65e-02   9.62e-04  -17.14  < 2e-16 ***
Poverty_Percentage              -4.12e+01   3.06e+01   -1.35  0.17836
CoC_Civilian_Labor_Force         3.23e-02   1.78e-03   18.12  < 2e-16 ***
Unemployment_Rate                6.06e+03   3.42e+03    1.77  0.07616 .
Vote_Percentage                  2.98e+03   8.54e+02    3.49  0.00050 ***
CoC_Poverty_Estimate_Age_0_17    7.25e-02   4.36e-03   16.64  < 2e-16 ***
Birth_Rate                      -9.96e+01   2.40e+01   -4.15  3.4e-05 ***
Death_Rate                       8.55e+01   2.43e+01    3.51  0.00045 ***
International_Migration_Rate    -1.13e+02   3.36e+01   -3.36  0.00079 ***
Domestic_Migration_Rate          6.77e+00   6.79e+00    1.00  0.31875
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3360 on 3367 degrees of freedom
Multiple R-squared:  0.525, Adjusted R-squared:  0.523
F-statistic:  310 on 12 and 3367 DF,  p-value: <2e-16
```

En este enfoque, la renta media ajustada mediante variables instrumentales se combina con un amplio rango de controles socioeconómicos y demográficos. El coeficiente para `fitted_Actual_CoC_Median_Rent2` es 5.53, mostrando un fuerte impacto del costo de la vivienda sobre la falta de vivienda (`p < 2e-16`). El R-cuadrado de 0.525 indica que el modelo es capaz de explicar más de la mitad de la variabilidad en la falta de vivienda. Aunque los errores estándar podrían necesitar ajustes debido a la posible propagación del error de la primera etapa, la significancia estadística de la renta ajustada refuerza su importancia como factor clave. Los resultados también subrayan la complejidad del fenómeno, con múltiples variables contribuyendo significativamente.

6. Regresión con variables significativas

```r
# First Stage: Regress Actual_CoC_Median_Rent on Previous_CoC_Neighbor_Median_Rent and other
first_stage_model3 <- lm(Actual_CoC_Median_Rent ~ Previous_CoC_Neighbor_Median_Rent +
                           CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
                           CoC_Civilian_Labor_Force +
                           Vote_Percentage +
                           CoC_Poverty_Estimate_Age_0_17 + Birth_Rate + Death_Rate +
                           International_Migration_Rate,
                         data = data)
```

```r
data$fitted_Actual_CoC_Median_Rent3 <- fitted(first_stage_model3)

# Second Stage: Regress Overall_Homeless on the fitted values from the first stage and other
second_stage_model3 <- lm(Overall_Homeless ~ fitted_Actual_CoC_Median_Rent3 +
                          CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
                          CoC_Civilian_Labor_Force +
                          Vote_Percentage +
                          CoC_Poverty_Estimate_Age_0_17 + Birth_Rate + Death_Rate +
                          International_Migration_Rate, data = data)
summary(second_stage_model3)
```

```
Call:
lm(formula = Overall_Homeless ~ fitted_Actual_CoC_Median_Rent3 +
    CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
    CoC_Civilian_Labor_Force + Vote_Percentage + CoC_Poverty_Estimate_Age_0_17 +
    Birth_Rate + Death_Rate + International_Migration_Rate, data = data)

Residuals:
   Min     1Q Median     3Q    Max
-21247   -666    179    807  61292

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -1.96e+03   6.49e+02   -3.01  0.00260 **
fitted_Actual_CoC_Median_Rent3   5.41e+00   4.97e-01   10.87  < 2e-16 ***
CoC_AVG_Median_Household_Income -5.53e-02   7.79e-03   -7.10  1.5e-12 ***
CoC_Population_Estimate         -1.62e-02   9.41e-04  -17.22  < 2e-16 ***
CoC_Civilian_Labor_Force         3.19e-02   1.74e-03   18.31  < 2e-16 ***
Vote_Percentage                  2.59e+03   8.42e+02    3.08  0.00212 **
CoC_Poverty_Estimate_Age_0_17    7.12e-02   4.21e-03   16.90  < 2e-16 ***
Birth_Rate                      -1.14e+02   2.26e+01   -5.03  5.1e-07 ***
Death_Rate                       7.50e+01   2.27e+01    3.30  0.00098 ***
International_Migration_Rate    -1.25e+02   3.26e+01   -3.83  0.00013 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3360 on 3370 degrees of freedom
Multiple R-squared:  0.524, Adjusted R-squared:  0.523
F-statistic:  413 on 9 and 3370 DF,  p-value: <2e-16
```

Este modelo optimiza el análisis al centrarse en variables significativas junto con la renta ajus-

tada como variable instrumental. El coeficiente de `fitted_Actual_CoC_Median_Rent3` de 5.41 sugiere un vínculo robusto y significativo entre el costo de la vivienda y el número de personas sin hogar (`p < 2e-16`). El modelo mantiene un R-cuadrado importante de 0.524, demostrando que, a pesar de la simplificación, las variables seleccionadas explican una proporción considerable de la variabilidad en la falta de vivienda. La persistencia de la significancia estadística en las variables clave enfatiza su relevancia para entender y abordar la falta de vivienda. Aunque el análisis de errores estándar podría ser necesario para asegurar la precisión de las estimaciones, la consistencia en los resultados subraya la fiabilidad de las conclusiones.

GLMs

Modelo Poisson

Ya que 'Overall_Homeless' es un conteo de personas sin hogar, suena sensato modelar esta variable con una regresión Poisson. Para ello, ajustamos un Modelo Lineal Generalizado.

```
poisson_model <- glm(Overall_Homeless ~ Actual_CoC_Median_Rent +
                       Previous_CoC_Neighbor_Median_Rent +
                       CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
                       Poverty_Percentage + Unemployment_Rate + Party_Democrat +
                       International_Migration_Rate + Domestic_Migration_Rate,
                     family = poisson(link = "log"), data = data)
summary(poisson_model)
```

```
Call:
glm(formula = Overall_Homeless ~ Actual_CoC_Median_Rent + Previous_CoC_Neighbor_Median_Rent +
    CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
    Poverty_Percentage + Unemployment_Rate + Party_Democrat +
    International_Migration_Rate + Domestic_Migration_Rate, family = poisson(link = "log"),
    data = data)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       5.05e+00   5.57e-03   906.7   <2e-16 ***
Actual_CoC_Median_Rent            1.43e-03   2.96e-06   483.6   <2e-16 ***
Previous_CoC_Neighbor_Median_Rent -6.53e-05   4.06e-06   -16.1   <2e-16 ***
CoC_AVG_Median_Household_Income   -6.11e-06   6.77e-08   -90.4   <2e-16 ***
CoC_Population_Estimate           3.66e-07   1.45e-10  2516.8   <2e-16 ***
Poverty_Percentage                3.30e-02   2.15e-04   153.3   <2e-16 ***
Unemployment_Rate                 1.88e+00   2.26e-02    83.4   <2e-16 ***
Party_Democrat                    3.07e-01   1.27e-03   242.8   <2e-16 ***
```

```
International_Migration_Rate       6.29e-02   1.79e-04   351.6   <2e-16 ***
Domestic_Migration_Rate            1.57e-02   5.36e-05   292.2   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11110305  on 3379  degrees of freedom
Residual deviance:  2287147  on 3370  degrees of freedom
AIC: 2315325

Number of Fisher Scoring iterations: 5
```

- El coeficiente de la variable Actual_CoC_Median_Rent nos dice que existe una relación positiva con la variable objetivo. Por cada dolar de incremento, el logaritmo del número esperado de personas sin hogar aumenta 0.00143.
- Contrario a las suposiciones previas, el coeficiente negativo de Previous_CoC_Neighbor_Median_Rent indica que al aumentar la renta mediana de los vecinos en el periodo anterior, disminuye el número de personas sin hogar. También hay que hacer notar que el valor del estimador es bastante bajo.
- Podemos notar que las únicas variables que tienen una relación negativa con la objetivo es Previous_CoC_Neighbor_Median_Rent y CoC_AVG_Median_Household_Income.
- Todos los coeficientes son estadísticamente significativos ($p < 0.001$), e indican (la mayoría) que son distintos de cero en la población. Así, tienen un efecto sobre nuestra variable dependiente.
- El resutlado de Residual deviance nos interesa que sea cercana a cero. Muestra la desviación del modelo con todas las variables. Al ser demasiado grande nos genera dudas, pues la diferencia nos indica que el modelo sí puede explciar una parte de la variabilidad en los homeless, pero para nada es perfecta.
- De la misma manera, nuestro Criterio de Akaike está muy elevado. Esto nos hace levantar sospechas.

Cálculo de la Devianza

```r
# Cálculo de la Devianza
deviance_null <- poisson_model$null.deviance
deviance_model <- poisson_model$deviance
deviance_r2 <- 1 - (deviance_model / deviance_null)

# Imprimir resultados
cat("Deviance R²: ", deviance_r2, "\n")
```

```
Deviance R²:  0.794
```

- El valor de la Devianza la utilizamos para evaluar la calidad de ajuste de nuestro modelo de regresión.
- El valor sugiere que el modelo explica aprox el 79.51% de la variabilidad en los conteos. Suena bastante bien; ajusta de manera buena a los datos.
- Pero se nos olvidó comprobar una cosa.
- Calculamos la media y la varianza para buscar la llamada "sobredisposición".
- La sobredisposición es cuando la varianza de los datos es mayor que la media.

```
mean_response <- mean(data$Overall_Homeless)
variance_response <- var(data$Overall_Homeless)

cat("Media de Total de Homeless: ",mean_response)
```

```
Media de Total de Homeless:  1485
```

```
cat("\nVarianza de Total de Homeless: ",variance_response)
```

```
Varianza de Total de Homeless:  23713544
```

```
# Calcular la estadística de dispersión
dispersion_stat <- sum(resid(poisson_model, type = "pearson")^2) / poisson_model$df.residual
cat("\nEstadística de dispersión: ",dispersion_stat)
```

```
Estadística de dispersión:  783
```

```
# Test de sobredispersión
dispersion_test <- dispersiontest(poisson_model, trafo = 1)
cat("\n\nPrueba de sobredisposición:\n")
```

```
Prueba de sobredisposición:
```

```
dispersion_test
```

```
    Overdispersion test

data:  poisson_model
z = 18, p-value <2e-16
alternative hypothesis: true alpha is greater than 0
sample estimates:
alpha
  780
```

- Al tener una varianza abismalmente mayor que la media, entraríamos en el problema de contradecir la suposición principal de la utilización de un modelo Poisson: que la media y la varianza son iguales.
- El test de sobredispersión confirma nuestro error. El valor cercano a 1 indica que la varianza es igual a la media. Dado que la nuestra es de 779, claramente observamos que la media y la varianza no son iguales.
- Así, el modelo Poisson no es el adecuado para el análisis de estos datos, pues podemos subestimar los errores de los estimadores.
- Tal vez por eso sonaban sospechosas los resultados previos.

Modelo Binomial

- Este modelo también es utilizado para datos que representan conteos (entre otras cosas). Este modelo permite que la varianza sea mayor que la media, por lo que resulta ser un modelo más flexible.
- Además, es recomendado cuando los datos presentan la sobredisposición.
- La función de enlace que utiliza también es la logarítmica.

```r
nb_model <- glm.nb(Overall_Homeless ~ Actual_CoC_Median_Rent + Previous_CoC_Neighbor_Median_
                   CoC_AVG_Median_Household_Income + CoC_Population_Estimate +
                   Poverty_Percentage + Unemployment_Rate + Party_Democrat +
                   International_Migration_Rate + Domestic_Migration_Rate,
                   data = data)

summary(nb_model)
```

```
Call:
glm.nb(formula = Overall_Homeless ~ Actual_CoC_Median_Rent +
    Previous_CoC_Neighbor_Median_Rent + CoC_AVG_Median_Household_Income +
    CoC_Population_Estimate + Poverty_Percentage + Unemployment_Rate +
    Party_Democrat + International_Migration_Rate + Domestic_Migration_Rate,
```

```
    data = data, init.theta = 1.917536384, link = log)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       5.02e+00   1.46e-01   34.29  < 2e-16 ***
Actual_CoC_Median_Rent            8.78e-04   1.11e-04    7.90  2.7e-15 ***
Previous_CoC_Neighbor_Median_Rent 5.65e-04   1.23e-04    4.60  4.3e-06 ***
CoC_AVG_Median_Household_Income  -8.74e-06   1.96e-06   -4.45  8.6e-06 ***
CoC_Population_Estimate           6.07e-07   1.10e-08   54.96  < 2e-16 ***
Poverty_Percentage                2.56e-02   5.51e-03    4.65  3.2e-06 ***
Unemployment_Rate                 2.23e+00   6.71e-01    3.32    9e-04 ***
Party_Democrat                    2.89e-01   2.94e-02    9.82  < 2e-16 ***
International_Migration_Rate      5.45e-02   6.55e-03    8.32  < 2e-16 ***
Domestic_Migration_Rate           2.03e-02   1.44e-03   14.08  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(1.92) family taken to be 1)

    Null deviance: 10481.5  on 3379  degrees of freedom
Residual deviance:  3663.8  on 3370  degrees of freedom
AIC: 51894

Number of Fisher Scoring iterations: 1

              Theta:  1.9175
          Std. Err.:  0.0434

 2 x log-likelihood:  -51872.0780
```

Cálculo de la Devianza

```r
# Cálculo de la Devianza
deviance_null <- nb_model$null.deviance
deviance_model <- nb_model$deviance
deviance_r2 <- 1 - (deviance_model / deviance_null)

# Imprimir resultados
cat("Deviance R²: ", deviance_r2, "\n")
```

```
Deviance R²:  0.65
```

- De entrada, podemos observar que todas las betas de las variables resultan ser significativas, por lo que sí tienen un efecto en el conteo de homeless (o eso podemos creer).
- Para este modelo, podemos observar que la única variable que tiene un efecto negativo es CoC_AVG_Median_Household_Income.
- La Devianza Nula nos dice cuánta variabilidad existe sin ningún predictor. La Devianza Residual nos dice cuánta variabilidad existe después de incluir los predictores. Podemos notar que disminuye el valor de la Devianza Residual. Esta reducción nos puede decir que el modelo puede explicar bien parte de los datos. Este valor de AIC (51,894) es mucho menor al obtenido con la regresión Poisson (2,315,325). Esto nos indica que el modelo utilizando una Binomial Negativa es mejor que el modelo Poisson que corrimos anteriormente.