Lección 12

Marcos Bujosa

4 de noviembre de 2023

Índice

1. Precio de casas unifamiliares (tasación de inmobiliarias y características de las viviendas)

Guión: Houses3.inp

1. Objetivo
   - Identificar un problema de multicolinealidad de grado.
   - Estudiar alternativas para evitar dicho problema.
2. Para empezar Para realizar la práctica, primero cargue los datos de la base de datos de Gretl:
   Archivo -->Abrir datos -->Archivo de muestra y en la pestaña *Wooldrige* seleccione `hprice1`.
   o bien teclee en linea de comandos:

   ```
   open hprice1
   ```

3. Los datos
   Ésta es una aplicación con datos reales de casas unifamiliares en EEUU:
   - price = precio de venta (en miles de $).
   - assess = tasación de la inmobiliaria (en miles de $).
   - bdrms = número de dormitorios.
   - lotsize = tamaño de la parcela en pies al cuadrado.
   - sqrft = tamaño de la vivienda en pies al cuadrado.
   - colonial = 1 si es de estilo colonial. 0 en el resto de casos.
4. El modelo inicial Inicialmente intentaremos emplear todas las variables disponibles para explicar los precios:

$$\boldsymbol{PRICE} = \beta_1 \mathbb{1} + \beta_2 \boldsymbol{ASSESS} + \beta_3 \boldsymbol{BDRMS} + \beta_4 \boldsymbol{LOTSIZE} + \beta_5 \boldsymbol{SQRFT} + \beta_6 \boldsymbol{COLONIAL} + \boldsymbol{U}$$

1.1. Actividad 1

Piense si alguna de las variables disponibles debería salir del modelo, por no ser relevante en la determinación del precio de la vivienda.

1.2. Actividad 2

Piense cuáles son los signos esperados de los parámetros del modelo.

1.3. Actividad 3

Ajuste por MCO el modelo:

$$\boldsymbol{price} = \widehat{\beta_1}\mathbf{1} + \widehat{\beta_2}\boldsymbol{assess} + \widehat{\beta_3}\boldsymbol{bdrms} + \widehat{\beta_4}\boldsymbol{lotsize} + \widehat{\beta_5}\boldsymbol{sqrft} + \widehat{\beta_6}\boldsymbol{colonial} + \widehat{\boldsymbol{e}}$$

```
Modelo1     <- ols price 0 assess bdrms lotsize sqrft colonial
```

- ¿Qué variables son significativas en este modelo? ¿Coincide esto con su previsión?
- A la luz de esta primera regresión. ¿Las inmobiliarias infravaloran o sobrevaloran los inmuebles?

1.4. Actividad 4

¡Sorprendentemente parece que las características del inmueble no son significativas para explicar el precio!

¿Es conjuntamente significativo este modelo para "explicar" los precios? Observe el estadístico F.

1.5. Actividad 5

Si hay una estrecha relación lineal entre regresores, es difícil discriminar el papel particular de cada uno. Esta incertidumbre se refleja en la varianza de los estimadores.

Consecuentemente algunas variables, en teoría importantes, pueden resultar estadísticamente no significativas.

Las inmobiliarias emplean las características del inmueble para arrojar una tasación. Así, es muy probable una estrecha relación entre las características y la tasación.

- Observe los diagramas de dispersión de las variables de cada

casa con su tasación

Ver -->Gráficos múltiples y elija `assess` como variable del eje `y`, y el resto como variables del eje `X`. *o bien teclee en linea de comandos*:

```
Diagramas   <- scatters bdrms lotsize sqrft colonial ; assess
```

1.6. Actividad 6

Analice la correlación entre `assess` y el resto de regresores.

```
corr assess bdrms lotsize sqrft colonial
```

1.7. Actividad 7

- La colinealidad se debe a que la muestra no contiene información suficiente para estimar con una precisión satisfactoria todos los parámetros.
- Las soluciones directas a este problema consisten en
  - añadir información (en este caso no podemos)
  - simplificar el modelo
  - También es posible eliminar información redundante mediante regresiones auxiliares

Mejor que analizar relaciones lineales entre pares (correlaciones) es estudiar si las características de la casa explican en un modelo lineal la valoración de las inmobiliarias.

- Ajuste por MCO el siguiente modelo

$$\boldsymbol{assess} = \widehat{\beta}_1\mathbf{1} + \widehat{\beta}_2\boldsymbol{bdrms} + \widehat{\beta}_3\boldsymbol{lotsize} + \widehat{\beta}_4\boldsymbol{sqrft} + \widehat{\beta}_5\boldsymbol{colonial} + \widehat{\boldsymbol{e}}$$

```
ModeloAux   <- ols assess 0 bdrms lotsize sqrft colonial
```

- Observe el coeficiente de determinación para comprobar hasta que punto las características de las vivienda en su conjunto "explican" la tasación de la inmobiliaria.
- ¿Qué variables son más significativas para explicar la tasación?

1.8. Actividad 8 - Eliminando información redundante

1. Las inmobiliarias suelen tener conocimiento del estado de conservación de la vivienda, su antigüedad, su situación, orientación, calidades, estética, etc. Todas estas características habrán sido tenidas en cuenta en las tasaciones, pero están mezcladas con las otras características explícitamente incluidas en el modelo. Ello explica que un 30 % de la variabilidad de `assess` no es explicada por el resto de regresores, pero si el 70 % restante.

2. Los errores de ajuste por MCO son ortogonales a los regresores. Así, los errores de la última regresión corresponden a todos esos factores que han intervenido en la tasación, pero que son ortogonales a los regresores del modelo.

3. En el modelo principal, omitiendo `assess` pero incluyendo como regresor los errores de esta última regresión, incorporamos información sobre la tasación que es ortogonal al resto de regresores

Guarde los residuos de la ultima regresión: En la ventana de la regresión auxiliar, pinche en ***Guardar -->Residuos*** y y ponga como nombre `OtrosFactores`.

o bien teclee en linea de comandos:

```
OtrosFactTasac = $uhat
```

- Ajuste por MCO el modelo:

$$\boldsymbol{price} = \widehat{\beta_1}\mathbf{1} + \widehat{\beta_2}\boldsymbol{bdrms} + \widehat{\beta_3}\boldsymbol{lotsize} + \widehat{\beta_4}\boldsymbol{sqrft} + \widehat{\beta_5}\boldsymbol{colonial} + \widehat{\beta_6}\boldsymbol{otrosfactores} + \widehat{\boldsymbol{e}}$$

```
Modelo2     <- ols price 0 bdrms lotsize sqrft colonial OtrosFactTasac
```

- ¿Hay indicios de multicolinealidad?
- Compare el ajuste de este modelo con el del modelo inicial.
- Omita las variables no significativas para obtener un modelo final.

```
ModeloFinal <- omit --auto
```

Código completo de la práctica

```
open hprice1

Modelo1     <- ols price 0 assess bdrms lotsize sqrft colonial

Diagramas   <- scatters bdrms lotsize sqrft colonial ; assess

corr assess bdrms lotsize sqrft colonial

ModeloAux   <- ols assess 0 bdrms lotsize sqrft colonial

OtrosFactTasac = $uhat

Modelo2     <- ols price 0 bdrms lotsize sqrft colonial OtrosFactTasac

ModeloFinal <- omit --auto
```

2. Construcción de vivienda nueva

Guión: RamanathanEX5-1.inp

Usando `data4-3.gdt` del libro de Ramanathan, con el nº de viviendas iniciadas (en miles) en los EEUU (`housing`), población (`pop`), PIB en miles de millones de dólares con base en el año 1982 (`gnp`) y tipo de interés del crédito hipotecario (`intrate`):

```
open data4-3.gdt
```

- Estime el siguiente modelo `ModeloA` y añádalo a la tabla de modelos

$$housing_n = \alpha_1 + \alpha_2 \cdot intrate_n + \alpha_3 \cdot pop_n + otros\ factores_n$$

```
ModeloA <- ols housing 0 intrate pop
modeltab add
```

- Estime el siguiente modelo `ModeloB` y añádalo a la tabla de modelos

$$housing_n = \beta_1 + \beta_2 \cdot intrate_n + \beta_3 \cdot gnp_n + otros\ factores_n$$

```
ModeloB <- ols housing 0 intrate gnp
modeltab add
```

- Estime el siguiente modelo `ModeloC` y añádalo a la tabla de modelos

$$housing_n = \gamma_1 + \gamma_2 \cdot intrate_n + \gamma_3 \cdot pop_n + \gamma_4 \cdot gnp_n + otros\ factores_n$$

```
ModeloC <- ols housing 0 intrate pop gnp
modeltab add
```

- Compare los resultados.

```
modeltab show
```

- Realice un contraste de colinealidad entre regresores (`vif`).

```
vif
```

- Contraste la significatividad conjunta de `gnp` y =pop en `ModeloC`.

```
omit pop gnp
scalar estadF  = $test
scalar pvalorF = pvalue(F, 2, ModeloC.$df, estadF)
```

- Mire las correlaciones entre los tres regresores.

```
corr intrate pop gnp
```

Código completo de la práctica

```
open data4-3.gdt

ModeloA <- ols housing 0 intrate pop
modeltab add

ModeloB <- ols housing 0 intrate gnp
modeltab add

ModeloC <- ols housing 0 intrate pop gnp
modeltab add

modeltab show

vif

omit pop gnp
scalar estadF  = $test
scalar pvalorF = pvalue(F, 2, ModeloC.$df, estadF)

corr intrate pop gnp
```

3. Construcción de vivienda nueva (transformación a datos per-cápita)

Guión: RamanathanEX5-1v2.inp

Abra el conjunto de datos `data4-3.gdt` del libro de Ramanathan, con el nº de viviendas iniciadas (en miles) en los EEUU (`housing`), población (`pop`), PIB en miles de millones de dólares con base en el año 1982 (`gnp`) y tipo de interés del crédito hipotecario (`intrate`):

```
open data4-3.gdt
```

- Estime de nuevo el siguiente modelo (que tiene multicolinealidad)

$$housing_n = \gamma_1 + \gamma_2 \cdot intrate_n + \gamma_3 \cdot pop_n + \gamma_4 \cdot gnp_n + otros\ factores_n$$

```
ModeloMCol <- ols housing 0 intrate pop gnp
modeltab add
```

- Defina la construcción de viviendas per-cápita (`housingPC}=housing/pop`) y la correspondiente producción per-cápita (`gnpPC`); y estime el siguiente modelo

$$housingPC_n = \beta_1 + \beta_2 \cdot intrate_n + \beta_3 \cdot gnpPC_n + otros\ factores_n$$

```
series housingPC = housing/pop
series gnpPC     = gnp/pop
ModeloPC <- ols housingPC 0 intrate gnpPC
```

- Compare los resultados.

```
modeltab show
```

- Mire las correlaciones entre los nuevos regresores y compárelas con las del antiguo modelo.

```
corr gnp   intrate
corr gnpPC intrate
```

- Realice un contraste de colinealidad `vif` entre regresores en ambos modelos.
  (*la función* ***`vif`*** *se debe ejecutar justo después de estimar cada modelo*).

```
ols housingPC 0 intrate gnpPC
vif

ols housing   0 intrate gnp   pop
vif
```

Código completo de la práctica

```
open data4-3.gdt

ModeloMCol <- ols housing 0 intrate pop gnp
modeltab add

series housingPC = housing/pop
series gnpPC     = gnp/pop
ModeloPC <- ols housingPC 0 intrate gnpPC

modeltab show

corr gnp   intrate
corr gnpPC intrate

ols housingPC 0 intrate gnpPC
vif

ols housing   0 intrate gnp   pop
vif
```

4. Gastos de mantenimiento de los automóviles

Guión: RamanathanEX5-2.inp

(*La siguiente práctica reproduce el ejemplo 5.2 del libro de Ramanathan.*)

Abra el conjunto de datos `data3-7.gdt`, del libro de Ramanathan, sobre costes de mantenimiento de coches Toyota (`cost`), así como la edad de cada coche en semanas (`age`) y en número de millas recorridas (`miles`).

```
open data3-7.gdt
```

- Mire las correlaciones entre los dos regresores.

  ```
  corr age miles
  ```

- Estime el siguiente modelo `ModeloA` y añádalo a la tabla de modelos

$$cost_n = \alpha_1 + \alpha_2 \cdot age_n + otros\ factores_n$$

  ```
  ModeloA <- ols cost 0 age
  modeltab add
  ```

- Estime el siguiente modelo `ModeloB` y añádalo a la tabla de modelos

$$cost_n = \beta_1 + \beta_2 \cdot miles_n + otros\ factores_n$$

  ```
  ModeloB <- ols cost 0 miles
  modeltab add
  ```

- Estime el siguiente modelo `ModeloC` y añádalo a la tabla de modelos

$$cost_n = \gamma_1 + \gamma_2 \cdot age_n + \gamma_4 \cdot miles_n + otros\ factores_n$$

  ```
  ModeloC <- ols cost 0 age miles
  modeltab add
  ```

- Compare los resultados.

  Fíjese que aunque `age` y `miles` son siempre significativas, su esfecto no está claro dada la enorme variabilidad en los resultados.

  ```
  modeltab show
  ```

- Realice un contraste de colinealidad entre regresores.

  ```
  vif
  ```

- Contraste la significatividad conjunta de `miles` y `age` en `ModeloC`.

```
scalar estadF  = $Fstat
scalar pvalorF = pvalue(F, 2, ModeloC.$df, estadF)
```

- La serie de errores absolutos de predicción (APE) se calcula como

$$ape_n = 100 \left| \frac{\widehat{e}_n}{y_n} \right|$$

Calcule la serie de errores absolutos de predicción de cada modelo.

```
# Error porcentual absoluto (ape)
series apeA = 100*abs(ModeloA.$uhat)/cost
series apeB = 100*abs(ModeloB.$uhat)/cost
series apeC = 100*abs(ModeloC.$uhat)/cost
```

- Calcule los errores absolutos de predicción medios (MAPE) de cada modelo. ¿Qué modelo predice mejor?

Los errores absolutos medios de predicción (MAPE) son:

- $mapeA = 227,75136$
- $mapeB = 278,19832$
- $mapeC = 48,240362$

```
# Error porcentual absoluto medio
scalar mapeA = mean(apeA)
scalar mapeB = mean(apeB)
scalar mapeC = mean(apeC)
print mapeA mapeB mapeC
```

Código completo de la práctica

```
open data3-7.gdt

corr age miles

ModeloA <- ols cost 0 age
modeltab add

ModeloB <- ols cost 0 miles
modeltab add

ModeloC <- ols cost 0 age miles
modeltab add

modeltab show

vif

scalar estadF  = $Fstat
scalar pvalorF = pvalue(F, 2, ModeloC.$df, estadF)

# Error porcentual absoluto (ape)
series apeA = 100*abs(ModeloA.$uhat)/cost
series apeB = 100*abs(ModeloB.$uhat)/cost
series apeC = 100*abs(ModeloC.$uhat)/cost

# Error porcentual absoluto medio
scalar mapeA = mean(apeA)
scalar mapeB = mean(apeB)
scalar mapeC = mean(apeC)
print mapeA mapeB mapeC
```

5. Gastos de mantenimiento de los automóviles (ortogonalización de regresores mediante regresiones auxiliares)

Guión: RamanathanEX5-2v2.inp

Abra el conjunto de datos `data3-7.gdt`, del libro de Ramanathan, sobre costes de mantenimiento de coches Toyota (`cost`), así como la edad de cada coche en semanas (`age`) y en número de millas recorridas (`miles`).

```
open data3-7.gdt
```

- Estime el siguiente modelo `ModeloA`

$$cost_n = \alpha_1 + \alpha_2 \cdot age_n + otros\ factores_n$$

```
ModeloA      <- ols cost 0 age
modeltab add
```

- Realice una regresión auxiliar de `miles` sobre una constante y `age`. Guarde los residuos con el nombre `uhatMiles`.

```
AuxRegA      <- ols miles const age
series uhatMiles = $uhat
```

Piense qué interpretación tienen los residuos de esta regresión auxiliar.

`uhatMiles` es la parte de las millas recorridas que no se puede aproximar por la edad del coche (algo así como un indicador de la intensidad de uso del coche).

- Verifique que no hay correlación entre los residuos de esta regresión auxiliar (`uhatMiles`) y `age`. ¿Por qué?

```
corr uhatMiles age
```

- Estime el siguiente modelo `ModeloA_Raux`

$$cost_n = \alpha_1 + \alpha_2 \cdot age_n + \alpha_3 \cdot uhatMiles_n + otros\ factores_n$$

y compare los resultados con el `ModeloA`. Realice un test de colinealidad para el modelo `ModeloA_Raux`. ¿Hay indicios de colinealidad?

```
ModeloA_Raux <- ols cost 0 age uhatMiles
modeltab add
vif
```

- Estime el siguiente modelo `ModeloB`

$$cost_n = \beta_1 + \beta_2 \cdot miles_n + otros\ factores_n$$

```
ModeloB      <- ols cost 0 miles
modeltab add
```

- Realice una regresión auxiliar de `age` sobre una constante y `miles`. Guarde los residuos con el nombre `uhatAge`.

```
AuxRegB      <- ols age 0 miles
series uhatAge   = $uhat
```

- Estime el siguiente modelo `ModeloB_Raux`

$$cost_n = \beta_1 + \beta_2 \cdot miles_n + \beta_3 \cdot uhatAge_n + otros\ factores_n$$

y compare los resultados con el `ModeloB`. Realice un test de colinealidad para el modelo `ModeloB_Raux`. ¿Hay indicios de colinealidad?

```
ModeloB_Raux <- ols cost 0 miles uhatAge
modeltab add
vif
```

- Piense qué interpretación tiene el regresor `uhatAge` en esta regresión.

  El regresor `uhatAge` reflejará la parte del coste de mantenimiento debido al mero paso del tiempo (pues ya hemos descontado el uso del coche). Es una medida por el deterioro de coche por el simple paso del tiempo. Interpretar el efecto que tiene `uhatAge` sobre el coste de mantenimiento es más natural que el que tenía `uhatMiles`.

- Estime el siguiente modelo `ModeloC`

$$cost_n = \gamma_1 + \gamma_2 \cdot age_n + \gamma_4 \cdot miles_n + otros\ factores_n$$

```
ModeloC      <- ols cost 0 age miles
modeltab add
```

- Compare los resultados de los cinco modelos.

```
modeltab show
```

- Piense qué signos serian los esperados para los coeficientes y decida qué modelo prefiere.

  En el cuarto modelo `ModeloB_Raux` todos los regresores (incluido `uhatAge`) tiene una interpretación natural; además, en ese modelo los regresores son significativos y con los signos esperados.

  Nótese que como tanto el `ModeloC` como los dos con regresiones auxiliares emplean la misma información (las mismas variables), los tres tiene estadísticos de selección de modelos idénticos. El ajuste es igual de bueno en los tres casos, aunque la interpretación de resultados no es igual.

Código completo de la práctica

```
open data3-7.gdt

ModeloA      <- ols cost 0 age
modeltab add

AuxRegA      <- ols miles const age
series uhatMiles = $uhat

corr uhatMiles age

ModeloA_Raux <- ols cost 0 age uhatMiles
modeltab add
vif

ModeloB      <- ols cost 0 miles
modeltab add

AuxRegB      <- ols age 0 miles
series uhatAge   = $uhat

ModeloB_Raux <- ols cost 0 miles uhatAge
modeltab add
vif

ModeloC      <- ols cost 0 age miles
modeltab add

modeltab show
```

6. Pobreza y sus determinantes (omisión de variables)

Guión: RamanathanAp5-4.inp

Abra el conjunto de datos `data4-6.gdt`, del libro de Ramanathan, sobre el porcentaje de familias con una renta por debajo del nivel de pobreza (`povrate`) en los condados de California. Intentaremos explicar dicho porcentaje con las siguientes variables: `urb` es el porcentaje de población urbana; `famsize` es el número medio de personas por hogar, `unemp` es la tasa de desempleo, `highschl` es el porcentaje de población mayor de 25 que ha completado el "high school", `college` es el porcentaje de la población mayor de 25 que ha completado 4 o más años de "college", y `medinc` es la renta media familiar.

```
open data4-6.gdt
```

- Observe la correlación entre las variables explicativas

  ```
  corr urb famsize unemp highschl college medinc
  ```

- Estime el modelo "ModeloA" con todas las variables explicativas.

  ```
  ModeloA <- ols povrate const urb famsize unemp highschl college medinc
  ```

- Observe los elevados p-valores de algunas variables. Observe también el inesperado signo para `college`... ¿quizá es debido a un problema de multicolinealidad?
- Omita la variable menos significativa del modelo (la de mayor p-valor). Observe que prácticamente no hay cambios en el resultado.

  ```
  ModeloB <- omit unemp
  ```

- Omita la siguiente variable menos significativa. Observe que ahora todas las variables son significativas, pero se mantiene el signo equivocado para la variable "college".

  ```
  ModeloC <- omit urb
  vif
  ```

- Aunque la variable de renta familiar media (`medinc`) es la más significativa, cabe esperar que su efecto pueda ser aproximadamente explicado por las variables `highschl` y `college`. Pruebe a excluir dicha variable y compruebe que ahora el signo de `college` es adecuado. Esto sugiere la posibilidad de multicolinealidad.

  ```
  ModeloD <- omit medinc
  ```

- Realice una regresión auxiliar de `medinc` sobre `famsize`, `unemp`, `highschl` y `college` y compruebe que dichas variables explican el 85 % de la varianza de `medinc`. Quizá una solución es omitir dicha variable desde el principio, para poder captar adecuadamente el efecto de las demás variables.

  ```
  Aux     <- ols medinc 0 famsize unemp highschl college
  ```

- Estime un modelo con todas las variables explicativas excepto `medinc`, y siga un procedimiento de eliminación secuencial de variables no significativas. ¿Es satisfactorio el modelo final?

  ```
  ols povrate const urb famsize unemp highschl college
  Final   <- omit --auto
  modeltab show
  ```

Código completo de la práctica

```
open data4-6.gdt

corr urb famsize unemp highschl college medinc

ModeloA <- ols povrate const urb famsize unemp highschl college medinc

ModeloB <- omit unemp

ModeloC <- omit urb
vif

ModeloD <- omit medinc

Aux     <- ols medinc 0 famsize unemp highschl college

ols povrate const urb famsize unemp highschl college
Final   <- omit --auto
modeltab show
```