

Sesión 14 (C) — La importancia del R^2 ajustado

Marcos Bujosa

Descripción de la práctica

Las dos prácticas anteriores mostraron que añadir un regresor nunca empeora el R^2 , y que el R^2 ajustado corrige ese defecto penalizando la complejidad. Podría pensarse, entonces, que basta con perseguir el $\bar{R}^2$ más alto para encontrar “el mejor modelo”. Veámoslo con un conjunto de datos (posiblemente ficticios) minúsculo: el peso y la edad de 8 niños. Ajustaremos tres modelos alternativos.

Objetivo

1. Ajustar tres modelos polinómicos anidados (grado 1, 2 y 3) sobre la misma variable explicativa, *edad*.
2. Comparar los modelos en función del ajuste medido por $\bar{R}^2$.
3. Extrapolar los modelos más allá del rango muestral para observar las predicciones.
4. Distinguir con precisión *variable explicativa* (la edad) de *regresores* (edad, edad al cuadrado, edad al cubo): tres regresores distintos al servicio de una sola variable explicativa.
5. Comparar la elección de modelo que sugiere $\bar{R}^2$ con la que sugiere extrapolar fuera de la muestra.

Actividad 1 - Los datos

Introducimos a mano los datos de peso (kg) y edad (años) de los 8 niños.

en línea de comandos:

```
nulldata 8 --preserve
setobs 1 1 --cross-section

matrix medad = {7;7;8;10;10;11;12;14}
matrix mpeso = {39;40;42;49;51;54;56;58}
series edad = medad
series peso = mpeso

setinfo peso -d "Peso (kg)" -n "Peso (kg)"
setinfo edad -d "Edad (años)" -n "Edad (años)"

print edad peso --byobs
```

Con solo 8 observaciones, este ejemplo no pretende decir nada serio sobre el crecimiento infantil: es, deliberadamente, un caso de laboratorio para poner a prueba nuestra confianza en R^2 y $\bar{R}^2$.

Actividad 2 - Modelo 1: la recta

en línea de comandos:

Licencia: Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

```
ols peso const edad
gnuplot peso edad --output="peso_edad_lineal.png"
```

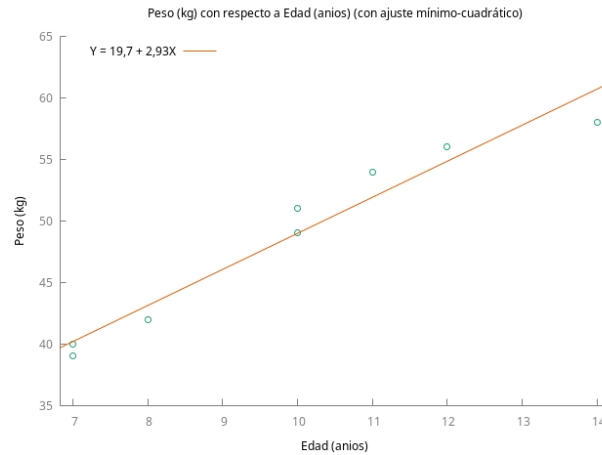


Figura 1: Modelo 1: peso edad (recta).

Anote R^2 y $\bar{R}^2$ de este primer modelo.

Actividad 3 - Modelo 2: la parábola

Añadimos edad al cuadrado como segundo regresor.

en línea de comandos:

```
series edad2 = edad^2
ols peso const edad edad2
gnuplot peso edad --fit=quadratic --output="peso_edad_cuadratico.png"
```

(Si su versión de Gretl no admite la opción `--fit=quadratic` en `gnuplot`, genere el gráfico desde el menú **Gráfico X-Y** y seleccione, en el diálogo, un ajuste cuadrático.)

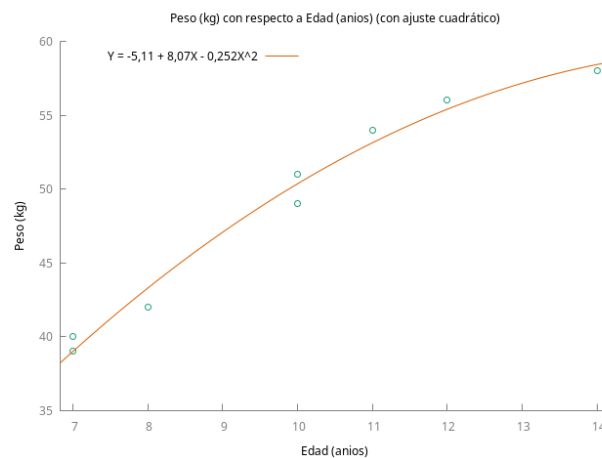


Figura 2: Modelo 2: peso edad+edad² (parábola).

Compare R^2 y $\bar{R}^2$ con los del Modelo 1. ¿Tiene sentido la comparación mirando ambos estadísticos? ¿O realmente sólo es lícito comparar el ajuste con uno de ellos? ¿Con cuál?

Actividad 4 - Modelo 3: la cúbica

Añadimos, además, edad al cubo.

en línea de comandos:

```
series edad3 = edad^3
ols peso const edad edad2 edad3
gnuplot peso edad --fit=cubic --output="peso_edad_cubico.png"
```

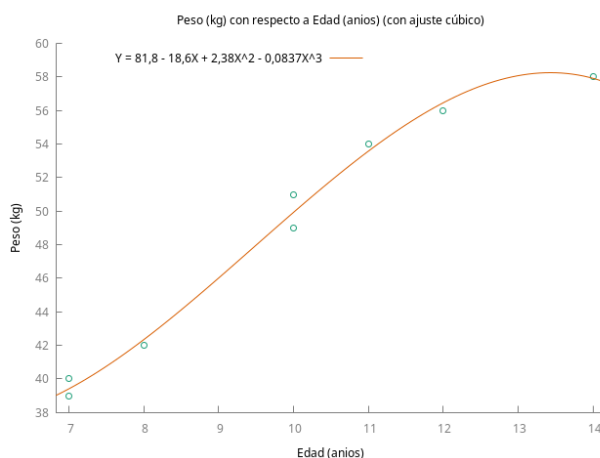


Figura 3: Modelo 3: peso edad+edad²+edad³ (cúbica). Fíjese en el tramo final de la curva.

Antes de continuar: mire con atención el tramo final de esta última curva, hacia las edades más altas de la muestra (13-14 años). ¿Ve algo que le llame la atención?

Actividad 5 - Los tres modelos, uno al lado del otro

en línea de comandos:

```
ols peso const edad --quiet
scalar Rbarsq1 = 1 - (1-$rsq)*($nobs-1)/($nobs-$ncoeff)
printf "Modelo 1 (recta):    R2 = %.4f    R2 ajustado = %.4f\n", $rsq, Rbarsq1

ols peso const edad edad2 --quiet
scalar Rbarsq2 = 1 - (1-$rsq)*($nobs-1)/($nobs-$ncoeff)
printf "Modelo 2 (parabola): R2 = %.4f    R2 ajustado = %.4f\n", $rsq, Rbarsq2

ols peso const edad edad2 edad3 --quiet
scalar Rbarsq3 = 1 - (1-$rsq)*($nobs-1)/($nobs-$ncoeff)
printf "Modelo 3 (cubica):   R2 = %.4f    R2 ajustado = %.4f\n", $rsq, Rbarsq3
```

```
Modelo 1 (recta):    R2 = 0,9490    R2 ajustado = 0,9405
Modelo 2 (parabola): R2 = 0,9840    R2 ajustado = 0,9776
Modelo 3 (cubica):  R2 = 0,9922    R2 ajustado = 0,9863
```

Tanto R^2 como $\bar{R}^2$ crecen sin excepción del Modelo 1 al Modelo 3. Si solo consideráramos esta tabla, la conclusión sería inmediata: “el Modelo 3 es el mejor”. La siguiente actividad pone a prueba esa conclusión.

Actividad 6 - Extrapolar los tres modelos

Vamos a predecir el peso a edades *fuera* del rango observado (7 a 14 años). Y lo haremos con los *tres* modelos a la vez, no solo con el “ganador”: si extrapoláramos únicamente el Modelo 3 podríamos llevarnos una impresión equivocada sobre los otros dos. Además, llegaremos hasta los 20 años, porque el problema que buscamos no aparece al mismo tiempo en todos.

en línea de comandos:

```
ols peso const edad edad2 edad3 --quiet
scalar b1 = $coeff(const)
scalar b2 = $coeff(edad)
scalar b3 = $coeff(edad2)
scalar b4 = $coeff(edad3)

ols peso const edad --quiet
scalar r1 = $coeff(const)
scalar r2 = $coeff(edad)

ols peso const edad edad2 --quiet
scalar p1 = $coeff(const)
scalar p2 = $coeff(edad)
scalar p3 = $coeff(edad2)

printf "edad      Modelo 1      Modelo 2      Modelo 3\n"
printf "      (recta)      (parabola)      (cubica)\n"
printf "-----\n"
loop e=13..20 --quiet
    printf "%3d    %9.2f    %9.2f    %9.2f\n", e, \
        r1 + r2*e, \
        p1 + p2*e + p3*e^2, \
        b1 + b2*e + b3*e^2 + b4*e^3
endloop

printf "\nEdad a la que cada modelo alcanza su maximo:\n"
printf "  Modelo 1 (recta)      : nunca, siempre crece\n"
printf "  Modelo 2 (parabola)   : %.1f anios\n", -p2/(2*p3)
printf "  Modelo 3 (cubica)     : %.1f anios\n", (-2*b3 - sqrt(4*b3^2 - 12*b2*b4))/(6*b4)
```

edad	Modelo 1 (recta)	Modelo 2 (parabola)	Modelo 3 (cubica)

13	57,78	57,17	58,08
14	60,71	58,43	57,92
15	63,64	59,19	55,49
16	66,57	59,44	50,29
17	69,50	59,19	41,82
18	72,43	58,43	29,57
19	75,36	57,17	13,04
20	78,29	55,41	-8,28

Edad a la que cada modelo alcanza su maximo:
 Modelo 1 (recta) : nunca, siempre crece
 Modelo 2 (parabola) : 16,0 anios
 Modelo 3 (cubica) : 13,4 anios

La tabla se lee de izquierda a derecha, y cuenta una historia muy ordenada.

El *Modelo 3* —el de mayor R^2 y mayor $\bar{R}^2$ de los tres— se desmorona enseguida: a los 15 años predice 55,49 kg, a los 17 baja a 41,82, y a los 20 da un peso *negativo*. Ningún estadístico de bondad de ajuste le habría advertido de esto.

Pero lo verdaderamente instructivo es la columna de en medio. El *Modelo 2*, la parábola, tiene el mismo defecto: su máximo está en los 16,0 años, y a partir de ahí también predice pérdida de peso (59,44 kg a los 16, 58,43 a los 18, 55,41 a los 20). Lo que ocurre es que lo hace *más tarde y más despacio*, de modo que si solo hubiéramos extrapolado a 15, 16 y 17 años —como es tentador hacer— la parábola habría parecido inocente.

Y el *Modelo 1*, la recta, es el único que no gira nunca. También es el de *menor* $\bar{R}^2$ de los tres.

Ordenando los tres puntos de giro:

modelo	máximo	$\bar{R}^2$
recta (grado 1)	nunca	el más bajo
parábola (grado 2)	16,0 años	intermedio
cúbica (grado 3)	13,4 años	el más alto

Cuanto mayor es el grado —y mayor el $\bar{R}^2$ —, *antes* llega el absurdo. Esa es la moraleja de la práctica, y conviene enunciarla sin suavizarla: aquí el orden de los modelos según $\bar{R}^2$ es exactamente el orden inverso de su credibilidad.

De hecho, para descartar el Modelo 3 ni siquiera hace falta salir de la muestra: su máximo (13,4 años) cae *dentro* del rango de edades observado, que llega hasta los 14. El descenso dentro de la muestra es pequeño —de 58,08 kg a los 13 a 57,92 a los 14, apenas 0,16 kg— y por eso en el gráfico de la Actividad 4 se aprecia como un aplanamiento más que como una caída; pero está ahí, y no tiene ningún fundamento.

Actividad 7 - Una variable explicativa, tres regresores

En los tres modelos de esta práctica hay una única *variable explicativa*: la edad. Sin embargo, el número de *regresores* crece de uno (Modelo 1: **edad**) a tres (Modelo 3: **edad**, **edad2**, **edad3**). Recordando la distinción de la lección 6: un regresor es cada vector que multiplica a un coeficiente $\hat{\beta}_j$ del modelo; una variable explicativa es el concepto económico o sustantivo que puede entrar en el modelo a través de uno o varios regresores a la vez —aquí, la propia edad, y sus potencias, todas ellas funciones de la misma variable explicativa original—.

Preguntas de interpretación para la clase

1. ¿Por qué R^2 y $\bar{R}^2$ crecen sin excepción del Modelo 1 al Modelo 3, si ya sabemos (lección 8, prácticas anteriores) que R^2 nunca puede bajar al añadir un regresor?
2. Si tuviera que elegir uno de los tres modelos únicamente en función de $\bar{R}^2$, ¿cuál elegiría? ¿Sigue eligiendo el mismo tras ver la Actividad 6?
3. ¿Qué le dice este ejercicio sobre la costumbre de comparar modelos *solo* mirando una tabla de estadísticos, sin generar ningún gráfico?
4. ¿Por qué edad, edad² y edad³ cuentan como tres regresores distintos, si los tres provienen de una única variable medida (la edad)?
5. ¿Diría que este problema (predicciones absurdas fuera del rango muestral) es exclusivo de los polinomios, o podría aparecer con cualquier modelo, lineal o no, si se extrapola lo suficiente fuera del rango de los datos con que fue estimado?

Para profundizar:

- Véanse las lecciones 8 y 10 ([lección 8](#), [lección 10](#)) para la advertencia general sobre qué certifica (y qué no) un buen ajuste, y para la distinción entre regresor y variable explicativa.
- Wooldridge, J. M. (2020). *Introductory Econometrics*, cap. 6, sección 6.2 (formas funcionales polinómicas y sus límites).

Código completo de la práctica

```
# -----
# Copyright (C) 2026 Marcos Bujosa
# Licencia: GNU General Public License v3.0 o posterior
# Este código es software libre y puede ser redistribuido y/o modificado bajo los términos de la GPL.
# Ver el archivo LICENSE del repositorio para más detalles.
# -----

# Los dos primeros comandos son necesarios para que Gretl guarde los resultados de la práctica en el
# ↪ directorio de trabajo
# al ejecutar lo siguiente desde un terminal (use los nombres y ruta que correspondan)
#
#   DIRECTORIO="Nombre_Directorio_trabajo" gretlcli -b ruta/nombre_fichero_de_la_practica.inp
#
# Si esto no le funciona en su sistema, comente las siguientes dos líneas y sitúese en el directorio de
# ↪ trabajo de gretl
# que corresponda (configure dicho directorio de trabajo desde la ventana principal de Gretl).

string directory = getenv("DIRECTORIO")
set workdir "@directory"

nulldata 8 --preserve
setobs 1 1 --cross-section

matrix medad = {7;7;8;10;10;11;12;14}
matrix mpeso = {39;40;42;49;51;54;56;58}
series edad = medad
series peso = mpeso

setinfo peso -d "Peso (kg)" -n "Peso (kg)"
setinfo edad -d "Edad (años)" -n "Edad (años)"

print edad peso --byobs

ols peso const edad
gnuplot peso edad --output="peso_edad_lineal.png"

series edad2 = edad^2
ols peso const edad edad2
gnuplot peso edad --fit=quadratic --output="peso_edad_cuadratico.png"

series edad3 = edad^3
ols peso const edad edad2 edad3
gnuplot peso edad --fit=cubic --output="peso_edad_cubico.png"

outfile --quiet comparacionTresModelosR2.txt
  ols peso const edad --quiet
  scalar Rbarsq1 = 1 - (1-$rsq)*($nobs-1)/($nobs-$ncoeff)
  printf "Modelo 1 (recta):   R2 = %.4f   R2 ajustado = %.4f\n", $rsq, Rbarsq1

  ols peso const edad edad2 --quiet
  scalar Rbarsq2 = 1 - (1-$rsq)*($nobs-1)/($nobs-$ncoeff)
  printf "Modelo 2 (parabola): R2 = %.4f   R2 ajustado = %.4f\n", $rsq, Rbarsq2

  ols peso const edad edad2 edad3 --quiet
  scalar Rbarsq3 = 1 - (1-$rsq)*($nobs-1)/($nobs-$ncoeff)
  printf "Modelo 3 (cubica):   R2 = %.4f   R2 ajustado = %.4f\n", $rsq, Rbarsq3
end outfile

outfile --quiet extrapolacionTresModelos.txt
  ols peso const edad edad2 edad3 --quiet
  scalar b1 = $coeff(const)
  scalar b2 = $coeff(edad)
  scalar b3 = $coeff(edad2)
  scalar b4 = $coeff(edad3)
```

```

ols peso const edad --quiet
scalar r1 = $coeff(const)
scalar r2 = $coeff(edad)

ols peso const edad edad2 --quiet
scalar p1 = $coeff(const)
scalar p2 = $coeff(edad)
scalar p3 = $coeff(edad2)

printf "edad      Modelo 1      Modelo 2      Modelo 3\n"
printf "      (recta)      (parabola)      (cubica)\n"
printf "-----\n"
loop e=13..20 --quiet
    printf "%3d    %9.2f    %9.2f    %9.2f\n", e, \
        r1 + r2*e, \
        p1 + p2*e + p3*e^2, \
        b1 + b2*e + b3*e^2 + b4*e^3
endloop

printf "\nEdad a la que cada modelo alcanza su maximo:\n"
printf "  Modelo 1 (recta)      : nunca, siempre crece\n"
printf "  Modelo 2 (parabola) : %.1f años\n", -p2/(2*p3)
printf "  Modelo 3 (cubica)      : %.1f años\n", (-2*b3 - sqrt(4*b3^2 - 12*b2*b4))/(6*b4)
end outfile

```

Enlace al guión: [S14-Prct-C-peso-edad.inp](#)

Respuestas

1. ¿Por qué crecen R^2 y $\bar{R}^2$ sin excepción aquí, si $\bar{R}^2$ en general puede bajar? La propiedad de la lección 8 (“ R^2 nunca baja”) es universal: se cumple siempre. La de $\bar{R}^2$ (“puede bajar”) es solo una posibilidad, no una certeza: baja únicamente cuando el regresor añadido aporta muy poco a la reducción de SRC. En este ejemplo concreto, con solo 8 observaciones y una relación no lineal genuina entre peso y edad, cada potencia adicional de la edad reduce sustancialmente los residuos, lo suficiente como para que también $\bar{R}^2$ suba. Que $\bar{R}^2$ suba aquí no contradice que en otros casos (como pudo ocurrir con **radial** en la práctica anterior) pueda bajar: ambos comportamientos son compatibles con la misma fórmula.
2. ¿Qué modelo elegiría solo con $\bar{R}^2$, y después de la Actividad 6? Mirando únicamente la tabla de la Actividad 5, la elección “mecánica” sería el Modelo 3, por tener el $\bar{R}^2$ más alto. Sin embargo, tras la Actividad 6, esa elección se revela problemática: un modelo que predice pérdida de peso con la edad, para niños en pleno crecimiento, no es creíble más allá de una tautología estadística. Ahora bien, la Actividad 6 muestra algo más incómodo que eso: el Modelo 2 *no se salva*. Su parábola también gira, a los 16,0 años, y también predice pérdida de peso a partir de ahí; solo lo hace más tarde que la cúbica, lo justo para pasar inadvertido si uno extrapola poco. El único de los tres sin punto de giro es el Modelo 1, que es precisamente el de menor $\bar{R}^2$. Dicho de otro modo: en este ejemplo, el orden de los modelos según $\bar{R}^2$ es el orden inverso de su credibilidad, y la elección razonable —la recta— es la que el criterio mecánico coloca en último lugar.
3. Lección sobre comparar solo con estadísticos, sin gráficos. Este ejercicio ilustra, con un caso extremo pero claro, por qué ningún estadístico de resumen —ni R^2 , ni $\bar{R}^2$, ni ningún otro número aislado— puede sustituir a la inspección visual del ajuste. Los números pueden mejorar de forma monótona mientras el modelo se vuelve, al mismo tiempo, cada vez menos razonable; solo el gráfico (o, como aquí, poner a prueba el modelo fuera de la muestra) revela ese deterioro cualitativo que ningún estadístico agregado captura por sí solo.
4. ¿Por qué edad, edad² y edad³ son tres regresores distintos de una sola variable? Porque un regresor, en el sentido técnico fijado en la lección 6, es cada vector concreto que multiplica a un coeficiente en la combinación lineal ajustada; edad, edad² y edad³ son tres vectores distintos de $\mathbb{R}^8$ (no son proporcionales entre sí), así que cuentan como tres regresores independientes a efectos del cálculo de $\hat{\beta}$, aunque los tres se calculen a partir de una sola magnitud medida (la edad) y respondan, en conjunto, a una única pregunta sustantiva: “¿cómo varía el peso con la edad?”.
5. ¿Es esto exclusivo de los polinomios? No: cualquier modelo, lineal o no en sus parámetros, puede producir predicciones absurdas si se extrapola suficientemente lejos del rango de datos con el que fue estimado. Los polinomios de grado alto son especialmente propensos a este problema porque su comportamiento fuera del rango muestral puede cambiar de forma muy brusca (curvaturas que se disparan hacia arriba o hacia abajo), pero el principio general —“un modelo ajustado a un rango de datos no garantiza nada fuera de ese rango”— es válido para cualquier forma funcional, lineal o logarítmica incluida. Tampoco es exclusivo de modelos poco razonables; un modelo adecuado a un rango muestral, puede funcionar de manera insatisfactoria fuera del rango en el que se estimó. Por otra parte, la extrapolación a valores irreales tampoco tiene sentido (por ejemplo, piense si tiene sentido interpretar el valor estimado para la constante en el modelo del precio de la vivienda de la Práctica A de hoy... sería lo mismo que preguntarse qué precio tiene una vivienda cuya superficie es cero).