

**Ejercicio:** Para los datos de los árboles de cerezo negro (hoja de datos `trees` de  $R$ ), añada uno o varios términos de orden superior para resolver el problema presente en los residuos. Haga las pruebas de hipótesis y los análisis de residuos que considere pertinentes.

La prueba  $F$  únicamente permite comparar modelos anidados. Para la comparación de modelos no anidados para el mismo conjunto de datos no existen pruebas de hipótesis, pero sí diversos criterios, los cuales se basan en la suma de cuadrados de los errores ajustada de diversas maneras por la complejidad del modelo.

Dos criterios de este tipo son

- El coeficiente de determinación múltiple  $R^2$  y su versión ajustada por complejidad.
- El *Criterio de Información de Akaike (AIC)*

La definición más común del AIC es:

$$AIC = -2(\text{máximo de la verosimilitud}) + 2(\text{número de parámetros})$$

Para un modelo de regresión con  $n$  observaciones,  $p$  parámetros y errores con distribución normal con varianza desconocida,

$$AIC = n \log(SSE/n) + 2p$$

El “mejor” modelo será aquel que tenga un AIC más pequeño.

Para poner en práctica todo lo discutido hasta ahora en relación a modelos de regresión analizaremos el conjunto de datos contenido en la hoja de datos `swiss` en *R*. Estos datos constan de medidas de fertilidad e indicadores socio-económicos para cada una de las 47 provincias francófonas de Suiza alrededor de 1888.

Los indicadores socioeconómicos incluidos en el estudio son:

- **Agriculture**: Porcentaje de la población dedicado a la agricultura.
- **Examination**: Porcentaje de reclutas que obtuvieron el máximo puntaje en los exámenes del ejército.
- **Education**: Porcentaje de la población que ha recibido educación más allá de la escuela primaria.
- **Catholic**: Porcentaje de católicos.
- **Infant.Mortality**: Mortalidad infantil (en porcentaje).

La variable de respuesta es un índice ajustado de fertilidad, **Fertility**

Para analizar gráficamente un conjunto de datos como éste, es útil el comando `pairs`, que produce gráficos causa-efecto para cada par de variables.

En lo que sigue usaremos tres nuevos comandos:

- `add1`: Dado un modelo y un conjunto de variables extra, determina el cambio en la suma de cuadrados residual y en el AIC al añadir cada una de las variables extra al modelo original.

```
> swiss.mod0_lm(Fertility ~ 1,data=swiss)
> add1(swiss.mod0, ~Agriculture + Examination + Education
+ + Catholic + Infant.Mortality)
Single term additions
```

Model:

```
Fertility ~ 1
```

	Df	Sum of Sq	RSS	AIC
<none>			7178.0	238.3
Agriculture	1	894.8	6283.1	234.1
Examination	1	2994.4	4183.6	215.0
Education	1	3162.7	4015.2	213.0
Catholic	1	1543.3	5634.7	229.0
Infant.Mortality	1	1245.5	5932.4	231.4

- `drop1`: Es equivalente a `add1`, pero eliminando variables.

```
> swiss.mod12345_lm(Fertility ~ Agriculture + Examination
+ + Education + Catholic + Infant.Mortality,data=swiss)
> drop1(swiss.mod12345, .~ Agriculture + Examination
+ +Catholic + Infant.Mortality)
Single term deletions
```

Model:

```
Fertility ~ Agriculture + Examination + Education
+ Catholic + Infant.Mortality
```

	Df	Sum of Sq	RSS	AIC
<none>			2105.04	190.69
Agriculture	1	307.72	2412.76	195.10
Examination	1	53.03	2158.07	189.86
Catholic	1	447.71	2552.75	197.75
Infant.Mortality	1	408.75	2513.79	197.03

Este comando es prácticamente equivalente a eliminar variables del modelo usando la prueba *t*.

- `step`: realiza un proceso automático de selección basado en el AIC, ya sea agregando variables (`direction = "forward"`), eliminándolas (`direction = "backward"`) o ambas (`direction = "both"`) La opción por defecto es añadir y eliminar variables.

```
> step(swiss.mod0, .~ Education + Catholic + Examination)
```

Para añadir términos de orden superior (cuadráticos, cúbicos, etc.) es necesario usar la notación  $I(x^k)$

```
> trees.mod2 <- lm(Volume~Girth + I(Girth^2)+Height,
+ data=trees)
> trees.mod3 <- lm(Volume~Girth + I(Girth^2)
+ + I(Girth^3) + Height,data=trees)
> summary(trees.mod3)
```

Call:

```
lm(formula = Volume ~ Girth + I(Girth^2) + I(Girth^3) + Height,
    data = trees)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.4552	-1.7100	0.1735	1.9336	4.1896

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.838176	32.113905	0.088	0.930253
Girth	-6.029064	7.617803	-0.791	0.435846
I(Girth^2)	0.498601	0.550646	0.905	0.373525
I(Girth^3)	-0.005434	0.012964	-0.419	0.678545
Height	0.391061	0.096206	4.065	0.000395 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.666 on 26 degrees of freedom

Multiple R-Squared: 0.9772, Adjusted R-squared: 0.9737

F-statistic: 278.7 on 4 and 26 degrees of freedom, p-value: 0