Ejemplo. ANOVA via bootstrap

Javier Santibáñez

30 de mayo de 2017

Resumen

Se describe como contrastar la hipótesis de significancia del modelo de regresión a través de un ejemplo para una regresión simple con un conjunto de datos simulados.

Introducción

Suponer que se quiere modelar un conjunto de observaciones $\mathbf{y}$ con un modelo RLM con matriz de diseño $\mathbf{X}$ (de dimensión $n \times (p+1)$, con p variables auxiliares)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

donde $\boldsymbol{\beta}$ es es el vector de parámetros del modelo y $\boldsymbol{\epsilon}$ es un vector de errores con distribución G con media $\mathbf{0}$ y matriz de covarianzas $\sigma^2\mathbf{I}$. Si añadimos el supuesto de normalidad en los errores, podemos contrastar la hipótesis de significancia del modelo:

$$H_0 : \beta_1 = \ldots = \beta_p = 0 \qquad \text{vs.} \qquad H_1 : \beta_i \neq 0, \text{para algún } i = 1, ..., p. \tag{2}$$

utilizando el análisis de varianza (ANOVA) que se basa en el estadístico

$$F = \frac{SCR/p}{SCE/(n-p-1)} \tag{3}$$

que bajo H_0 se distribuye $F_{p,n-p-1}$. Si la distribución de los errores no es normal, entonces no se garantiza que la significancia real de la prueba F sea la especificada.

Desarrollo

Cuando no es razonable asumir normalidad en los errores del modelo, la distribución del estadístico F en (3), bajo H_0, no es $F_{p,n-p-1}$, entonces para contrastar las hipótesis en (2) se debe aproximar tal distribución. Para tal efecto utilizamos *bootstrap*.

Supongamos que tenemos una muestra aleatoria $\mathbf{Z} = (Z_1, \ldots, Z_m)$ de una distribución H, completamente conocida, y que estamos interezados en obtener la distribución de un estadístico $T(\mathbf{Z})$. Una forma de prodecer es aplicar los resultados sobre transformaciones de vectores aleatorios. Otra forma es aproximar la distribución de $T(\mathbf{Z})$ con simulación como sigue:

- Generamos muestras aleatorias independientes de H, $\mathbf{z}_1, \ldots, \mathbf{z}_r$.

- Con cada muestra evaluamos el estadístico T, $t_1 = T(\mathbf{z}_1), \ldots, t_r = T(\mathbf{z}_r)$.
- $t_1, \ldots, t_r$ constituyen una muestra aleatoria de la distribución de $T(\mathbf{Z})$ y puede ser utilizada para aproximar su distribución.

En nuestro caso queremos aproximar la distribución del estadístico F bajo H_0. Bajo esta hipótesis el modelo es $\mathbf{Y} = \beta_0\mathbf{1} + \boldsymbol{\epsilon}$. Si conocieramos β_0 y la distribución H (observemos que bajo H_0 las X son irrelevantes), podríamos generar muestras de $\mathbf{Y}$ como sigue:

- Generar muestras de H, $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_r$,
- Hacer $\mathbf{y}_k = \beta_0\mathbf{1} + \boldsymbol{\epsilon}_k$, para $k = 1, \ldots, r$.
- $\mathbf{y}_1, \ldots, \mathbf{y}_r$, constituyen una muestra de $\mathbf{Y}$ bajo el modelo en H_0.

donde r es el número de repeticiones, que generalmente es grande.

Como no conocemos β_0 ni la distribución de los errores, debemos estimarlos. Es fácil probar que bajo H_0 el estimador de β_0 es

$$\bar{y}_n = \frac{1}{n}\sum_{j=1}^{n} y_j.$$

La distribución de los errores ϵ se puede aproximar con la distribución de los residuos escalados $\mathbf{S}$. Recordemos que los residuos $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$, tienen media $\mathbf{0}$ y matriz de covarianzas $\sigma^2(\mathbf{I} - \mathbf{H})$, por lo que los residuos escalados $R_i = E_i/\sqrt{1 - h_{ii}}$ tienen, cada uno, varianza σ^2 y se consideran mejores para aproximar a los errores. Recordar que $\mathbf{H}$ es la matriz sombrero, que en el modelo bajo H_0 es simplemente $\frac{1}{n}\mathbf{J}$; y h_{ii} es el i-ésimo elemento de la diagonal de $\mathbf{H}$, que bajo H_0 es $\frac{1}{n}$. Por lo tanto, los residuos escalados se calculan como

$$R_j = \frac{E_j}{\sqrt{1 - \frac{1}{n}}} = \frac{Y_j - \bar{Y}_n}{\sqrt{1 - \frac{1}{n}}}, \qquad j = 1, \ldots, n.$$

Una vez calculados los residuos escalados $\mathbf{R}$, podemos aproximar la distribución de los errores ϵ a partir de la distribución empirica de los R_j y simular de esta distribución. En breve, y omitiendo algunos detalles, para generar una nueva observación de ϵ seleccionamos al azar un R_j, $j = 1, \ldots, n$; y para generar una nueva observación de $\boldsymbol{\epsilon}$ seleccionamos una muestra alaeatoria con reeemplazo de tamaño n de $R_1, \ldots, R_n$.

Entonces, las muestras de Y bajo H_0 se obtienen como sigue:

- Se generan errores $\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_r$ como muestras aleatorias con reemplazo de tamaño n de $R_1, \ldots, R_n$.
- Se hace $\mathbf{y}_k = \bar{y}_n\mathbf{1} + \boldsymbol{\epsilon}_k$, $k = 1, \ldots, r$.
- Las observaciones $\mathbf{y}_1, \ldots, \mathbf{Y}_r$ asi generadas se pueden consisderar como una muestra (aproximada) de $\mathbf{Y}$ bajo H_0.

Una vez que se tienen las muestras $\mathbf{Y}$, se calcula el estadístico F para cada una de ellas, en esta caso es necesario emparejar las observaciones con las X correcpondientes. Los valores $F_1, \ldots, F_r$ constituyen una muestra del estadístico F bajo H_0.

Finalmente, para contrastar las hipótesis en (2), podemos calcular el cuantil $100(1 - \alpha)\%$ de los F_k simulados y comparar con el F_0, calculado con los datos originales. Se rechaza H_0 si el estadístico

F_0 es mayor que el cuantil calculado. Otra opción es aproximar el *p-value* como la proporción de F_k que son mayores que F_0. En este segundo caso se rechaza H_0 si el *p-value* calculado es menor que el nivel de significancia α específicado previamente.

Ejemplo

Vamos a ilustrar el algoritmo anterior con un ejemplo con un modelo RLS. Primero generamos la problación. Se fija la semilla con fines de reproducibilidad. Se debe notar que esta población cumple con todos los supuestos del modelo RLS, por lo que para probar la significancia del modelo se puede utilizar la tabla ANOVA, sin embargo utilizaremos *bootstrap*.

```
set.seed(3)
n <- 70 # tamaño de muestra
x <- rnorm(n, 75, 20) # variable explicativa
y <- 10 + 0.5*x + rnorm(n, 0, 25)  # variable respuesta
```

Ajustamos el modelo $y = \beta_0 + \beta_1 x + \epsilon$ y guardamos el estadístico F.

```
mod_base <- lm(y ~ x)
f_base <- anova(mod_base)[1, 'F value']
```

Ahora debemos aproximar la distribución del estadístico F bajo H_0, por lo que debemos simular obervaciones bajo la hipótesis nula. En el siguiente código se calculan:

- la estimación de β_0,
- los residuos (ordinarios) del modelo, y
- los residuos escalados del modelo.

```
b0h <- mean(y)  # estimador de beta_0
res <- y - b0h  # residuos
res_esc <- res/sqrt(1 - 1/n)  # residuos escalados
```

Simulamos $r = 1000$ conjuntos de datos con el procedimento descrito anteriormente. Para cada conjunto de datos, se ajusta el modelo $y = \beta_0 + \beta_1 x + \epsilon$ y guardamos el estadístico F de la taba ANOVA.

```
f_boot <- c()
for(j in 1:1000)
{
    y_sim <- b0h + sample(res_esc, n, T)
    f_boot[j] <- anova(lm(y_sim ~ x))[1, 'F value']
}
```

El vector `f_boot` contiene una muestra bajo H_0 que podemos utilizar para contrastar la hipótesis de significancia del modelo $y = \beta_0 + \beta_1 x + \epsilon$. Para ello calculamos el cuantil 0.95 de la muestra y lo comparamos con el estadístico F guardado en `f_base`.

```
quantile(f_boot, 0.95)
```

```
##      95%
## 4.382102
```

```
f_base
```

```
## [1] 11.05041
```

En este caso en particular, el estadístico `f_base` es mayor que el cuantil 0.95 de su distribución (aproximada con bootstrap) bajo H_0, por lo que rechazamos la hipótesis nula.

También es posible aproximar el *p-value* a partir de la muestra en `f_boot` como la proporción de observaciones en `f_boot` que son mayores a `f_base`.

```
sum(f_boot > f_base)/length(f_boot)
```

```
## [1] 0.003
```

Referencias

Efron B., Tibshirani R. (1994) *An introduction to the bootstrap*. CRC press.