estadistico_contraste

Jesús F García Gavilán

2026-01-31

Librerias:

```
library(dplyr)
library(knitr)
library(kableExtra)
```

Modelo:

El modelo analiza cómo diversas variables (`edad`, `sexo`, `tabaco`, `actividad física`, etc.) explican los niveles de `HDL` utilizando la base de datos `met_CQI_rl`.

En esta celda definimos nuestra hipótesis de investigación. Queremos entender cómo la `edad` y otras 15 variables (clínicas, demográficas y de estilo de vida) se relacionan con los niveles de colesterol `HDL`

- `lm()`: Es la función para modelos lineales (linear models).
- `hdl ~ ...`: El símbolo ~ separa la variable dependiente (`HDL`) de las independientes.
- `as.factor()`: Crucial para tratar variables categóricas (como el `sexo` o el `tabaco`) como grupos y no como valores numéricos.
- `data = met_CQI_rl`: Especificamos nuestra fuente de datos limpia.

```
rlm <- lm(hdl ~ edad0 + as.factor(sexo) + as.factor(tabaco0) + ps1 + ps2 +
            as.factor(grup_int) + energiat + alcoholg +
            imc1 + idcluster + escolar1 + getota_1 + as.factor(hipercol0) +
            as.factor(hta0) + as.factor(tra_col0) + as.factor(trathta0),
            data = met_CQI_rl)
```

Resumen estadístico

Aquí solicitamos a R que procese toda la información del modelo. No solo buscamos los coeficientes individuales, sino las métricas globales de calidad: el error residual, el R-cuadrado y, sobre todo, el estadístico F

- `summary(rlm)`: Genera un objeto que contiene toda la "inteligencia" del modelo.
- `Residuals`: Nos da una idea de la dispersión de los errores.
- `Adjusted R-squared`: Nos indica que el % de la variabilidad del HDL se explica por nuestro modelo.

```r
resumen <- summary(rlm)
resumen
```

```
##
## Call:
## lm(formula = hdl ~ edad0 + as.factor(sexo) + as.factor(tabaco0) +
##     ps1 + ps2 + as.factor(grup_int) + energiat + alcoholg + imc1 +
##     idcluster + escolar1 + getota_1 + as.factor(hipercol0) +
##     as.factor(hta0) + as.factor(tra_col0) + as.factor(trathta0),
##     data = met_CQI_rl)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.097  -7.741  -1.017   6.214  53.085
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.8443991  6.9870453   0.693  0.48819
## edad0                   0.0936644  0.0471792   1.985  0.04727 *
## as.factor(sexo)1       10.1525293  0.7823905  12.976  < 2e-16 ***
## as.factor(tabaco0)2     0.2675800  1.7082507   0.157  0.87555
## as.factor(tabaco0)3     1.3465117  1.5813028   0.852  0.39460
## as.factor(tabaco0)4     1.5187738  0.9620788   1.579  0.11460
## as.factor(tabaco0)5     1.4083376  0.9057566   1.555  0.12016
## ps1                    47.1183964  8.1348204   5.792 8.23e-09 ***
## ps2                    64.8023371  8.1789505   7.923 4.10e-15 ***
## as.factor(grup_int)2   -0.0761149  0.6559203  -0.116  0.90763
## as.factor(grup_int)3   -1.2297234  0.6733331  -1.826  0.06797 .
## energiat               -0.0004473  0.0004846  -0.923  0.35613
## alcoholg                0.1243415  0.0205467   6.052 1.75e-09 ***
## imc1                   -0.2501485  0.0796821  -3.139  0.00172 **
## idcluster              -0.0002208  0.0001229  -1.796  0.07263 .
## escolar1                0.9741236  0.2987839   3.260  0.00113 **
## getota_1                0.0019015  0.0012038   1.580  0.11439
## as.factor(hipercol0)1   1.2354470  0.7827131   1.578  0.11465
## as.factor(hta0)1        1.3861517  1.1244818   1.233  0.21785
## as.factor(tra_col0)1    0.0749714  0.6493700   0.115  0.90810
## as.factor(trathta0)1   -0.2972413  0.8863470  -0.335  0.73740
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.31 on 1737 degrees of freedom
##   (107 observations deleted due to missingness)
## Multiple R-squared:  0.1743, Adjusted R-squared:  0.1648
## F-statistic: 18.34 on 20 and 1737 DF,  p-value: < 2.2e-16
```

Extracción de los componentes del estadístico F

Para poder explicar el "corazón" de la regresión, aislamos numéricamente el estadístico de contraste. Esto nos permite cuantificar la relación señal/ruido de forma independiente para luego visualizarla.

- `resumen$fstatistic`: Accedemos directamente a los tres valores del test F

- `pf(...)`: Es la función de distribución de probabilidad. Calcula el área bajo la curva a la derecha de nuestro valor F para obtener el p-valor global.
- `lower.tail = F`: Le indicamos a R que queremos la probabilidad de la cola superior, que es la que define la significación estadística.

```
f_valor <- resumen$fstatistic[1]
df_num  <- resumen$fstatistic[2]
df_den  <- resumen$fstatistic[3]
p_global <- pf(f_valor, df_num, df_den, lower.tail = F)
```

```
resumen$fstatistic %>%
  as.list() %>%
  as.data.frame() %>%
  rename(f_valor = value, df_num = numdf, df_den = dendf) %>%
  mutate(p_global = pf(f_valor, df_num, df_den, lower.tail = F),
         across(everything(), as.character)) %>%
  tidyr::pivot_longer(cols = everything(), names_to = "Metrica", values_to = "Valor") %>%
  mutate(Metrica = case_when(Metrica == "f_valor" ~ "Estadístico F",
                             Metrica == "df_num" ~ "GL Numerador (Modelo)",
                             Metrica == "df_den" ~ "GL Denominador (Error)",
                             Metrica == "p_global" ~ "P-valor Global"),
         Valor = case_when(Metrica == "P-valor Global" ~ format.pval(as.numeric(Valor), eps = 0.001),
                           Metrica == "Estadístico F" ~ as.character(round(as.numeric(Valor), 2)),
                           T ~ Valor)) %>%
  kable() %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

Metrica	Valor
Estadístico F	18.34
GL Numerador (Modelo)	20
GL Denominador (Error)	1737
P-valor Global	< 0.001

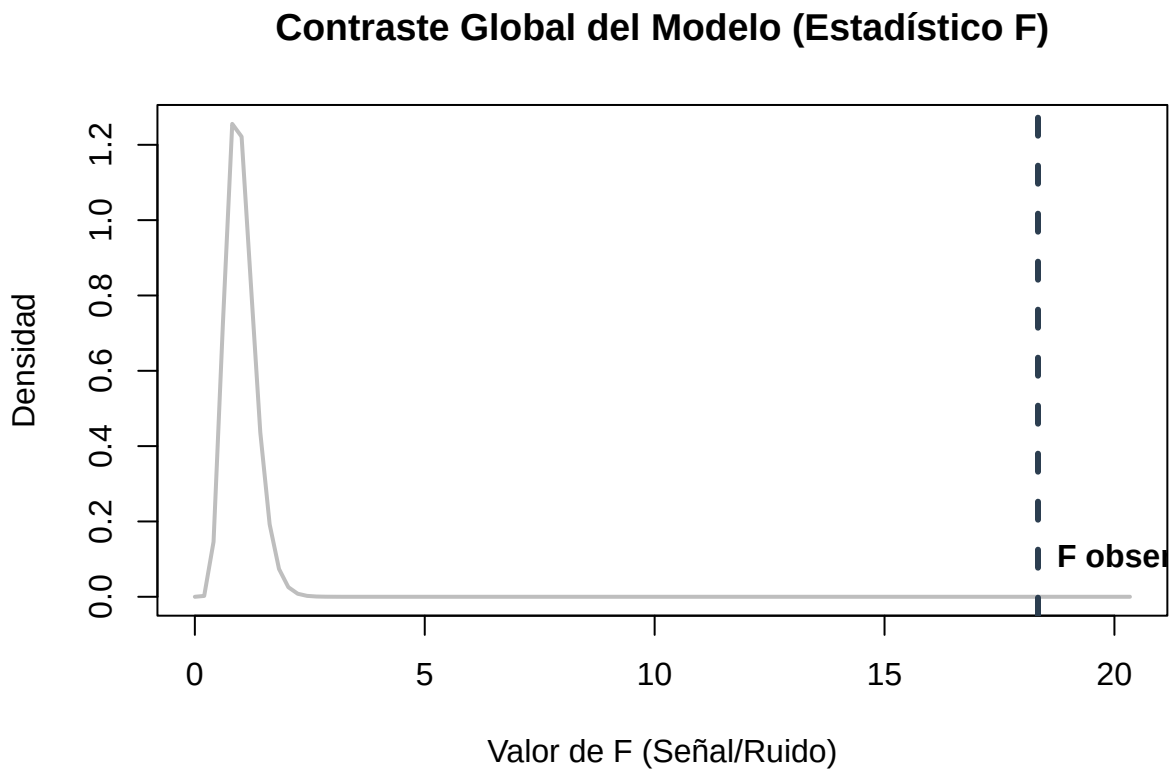
Visualización para la presentación

Esta es la parte más potente para una presentación. Traducimos números abstractos en una imagen. Mostramos cómo se vería el mundo si el azar fuera el único responsable (la curva) y dónde cae nuestra realidad (la línea del estadístico F).

- `curve(df(x, ...))`: Dibuja la densidad de la distribución F teórica bajo la hipótesis nula (H0).
- `abline(v = f_valor)`: Traza una línea vertical en el valor F.
- Interpretación visual: Al estar nuestra línea tan alejada de la masa principal de la curva (que está cerca de 1), demostramos visualmente que la probabilidad de que nuestros resultados sean casualidad es prácticamente inexistente

```
curve(df(x, df_num, df_den), from = 0, to = f_valor + 2,
      lwd = 2, col = "grey", main = "Contraste Global del Modelo (Estadístico F)",
      xlab = "Valor de F (Señal/Ruido)", ylab = "Densidad")
```

```r
# Sombreado de la zona de rechazo (crítica)
abline(v = f_valor, col = "#2c3e50", lwd = 3, lty = 2)
text(f_valor, 0.1, paste("F observado =", round(f_valor, 2)), pos = 4, font = 2)
```




```
## Señal detectada: SÍ
```

```
## Valor p global: < 2.22e-16
```

Como ven en el gráfico, si nuestro modelo fuera puro ruido, el valor F debería estar cerca de la zona gris. Pero nuestro valor está tan a la derecha que ni siquiera entra en el gráfico estándar; eso es una señal clínica real