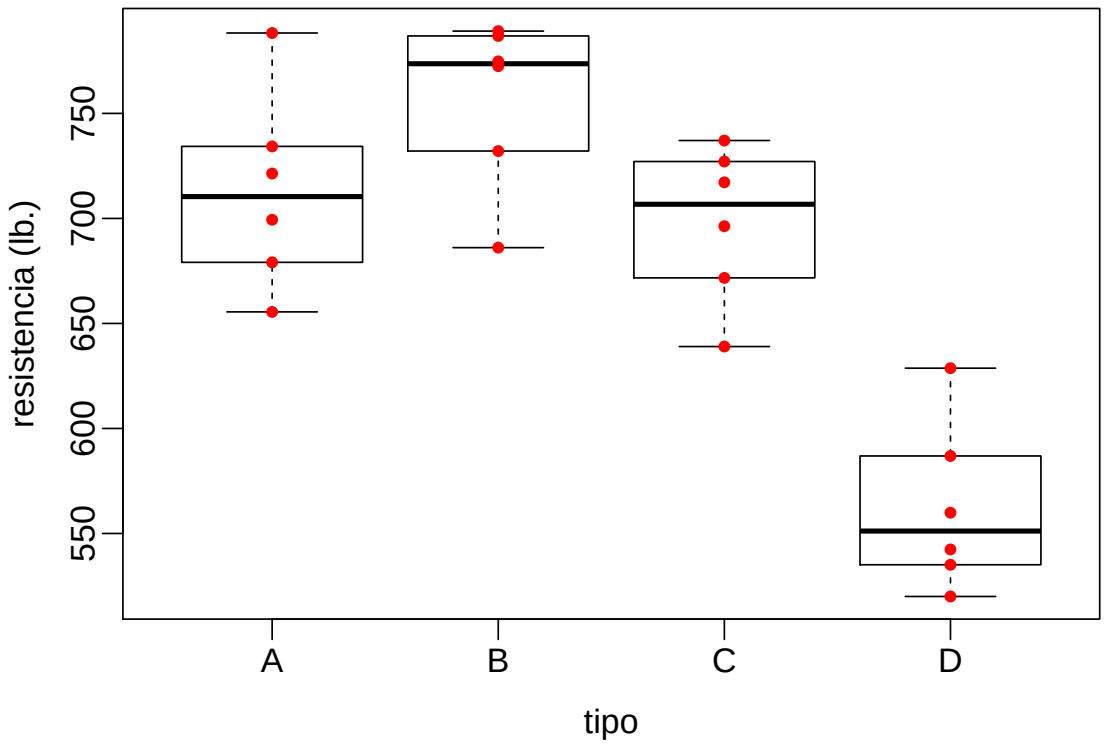
6 Análisis de la Varianza



−ebz

Contenido

Una *variable binaria* es aquella que toma solamente dos valores. E.g., sí, no; presente, ausente; mujer, hombre; encendido, apagado.
Las variables binarias acostumbran a denotarse numéricamente mediante *variables indicadoras* como $x = 0$ ó $x = 1$. A saber,

$$x = \begin{cases} 0 & : \text{ no, ausente, hombre, apagado} \\ 1 & : \text{ sí, presente, mujer, encendido} \end{cases}$$

Sea x una variable indicadora y considere el modelo de regresión lineal

$$y = \beta_0 + \beta_1 x + \epsilon$$

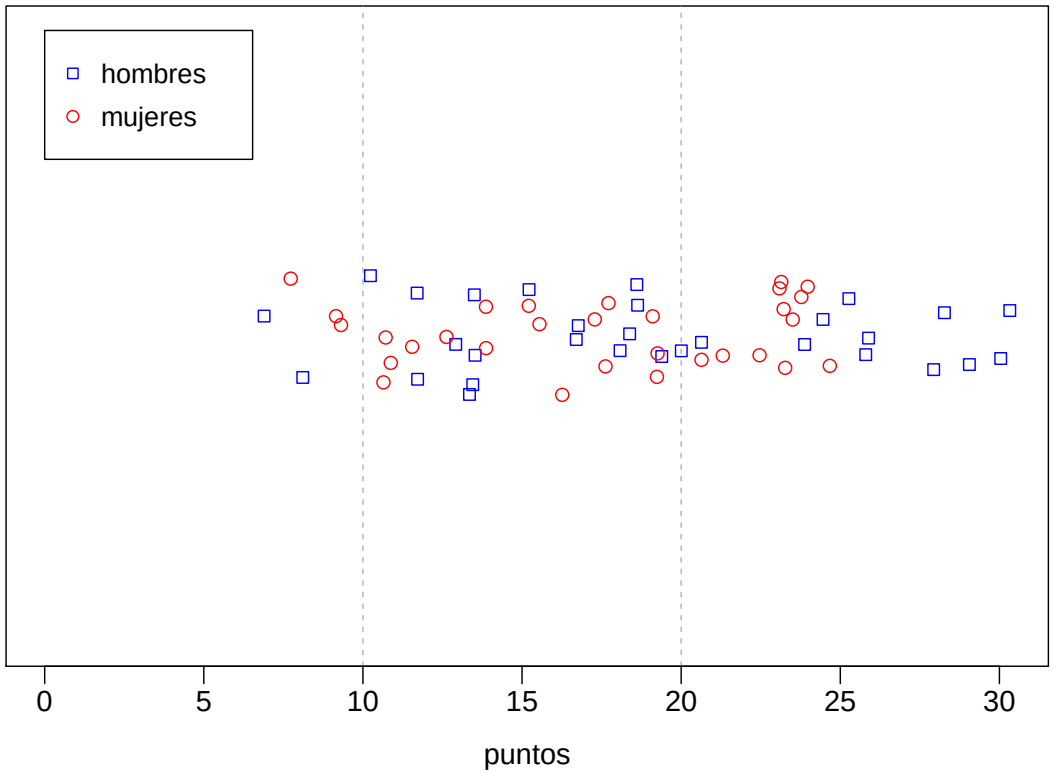
Entonces,

$$\begin{aligned} \bar{y}_{(x=0)} &= \beta_0 \\ \bar{y}_{(x=1)} &= \beta_0 + \beta_1 \end{aligned}$$

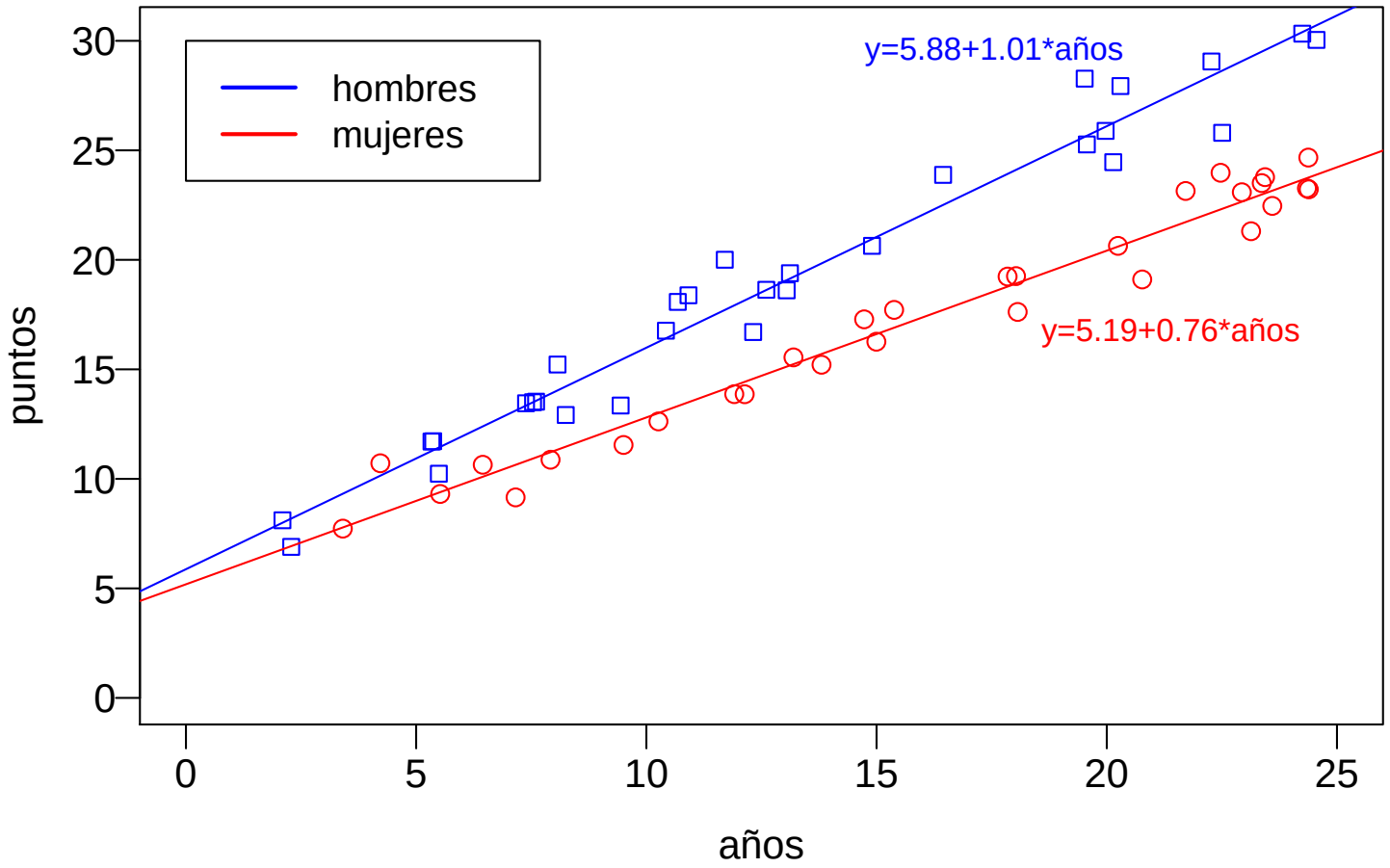
son respectivamente las *respuestas promedio* de los hombres, de los casos ausentes, apagados o negativos; o bien, de las mujeres, de los casos presentes, encendidos o positivos.
En este contexto, al parámetro β_1 se le llama el *efecto* de cambiar del estado 0 al estado 1.

Ejemplo: Salarios por género y antigüedad (simulado)

Una empresa tiene un *sistema de puntos*, que dependen básicamente de la antigüedad del empleado, y que están muy correlacionados con el salario. Se tomó una muestra aleatoria de 30 mujeres y 30 hombres y se observaron los puntos acumulados. ¿Hay diferencia de género en la asignación de puntos?



Ejemplo: Salarios por género y antigüedad (cont.)



Ejemplo: Salarios por género y antigüedad (cont.)

Una prueba de comparación de medias por género *no* rechazaría la hipótesis de igualdad de medias, con un estadístico $t = 1.017$ ($p = 0.31$).
Sin embargo, gráficamente pareciera que efectivamente hay diferencias en la asignación de puntos por género en un inicio y en el tiempo. ¿Son estas diferencias *significativas*? Más formalmente, ¿qué se puede decir de las siguientes hipótesis?

$$H_0^0 : \beta_0^M = \beta_0^H; \quad H_1^0 : \beta_1^M = \beta_1^H$$

Considere entonces el modelo de regresión

$$y = \beta_0 + \beta_1 z + \beta_2 t + \beta_3\, zt + \epsilon$$

donde y es la respuesta puntos, z la variable indicadora que denota el género, $z = 0$ si es mujer y $z = 1$ si es hombre, t el tiempo de antigüedad en la empresa, y ϵ la variación (error) aleatoria.

Note que $\beta_3 = 0$ implica que las pendientes no dependen del género. De otra forma, si $\beta_3 \neq 0$, la pendiente de la recta que corresponde al hombre es $\beta_2 + \beta_3$. Similarmente, los puntos asignados en un inicio a la mujer son β_0, mientras que si $\beta_1 \neq 0$ los puntos asignados al hombre en un inicio serán $\beta_0 + \beta_1$.

Ejemplo: Salarios por género y antigüedad (cont.)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.18630    0.53856   9.630 1.76e-13
z            0.68888    0.71298   0.966    0.338
t            0.76163    0.03105  24.530  < 2e-16
zt           0.24965    0.04478   5.574 7.40e-07

Residual standard error: 1.148 on 56 degrees of freedom
Multiple R-squared: 0.9664,     Adjusted R-squared: 0.9646
F-statistic: 537.6 on 3 and 56 DF,  p-value: < 2.2e-16
```

Estadísticamente $\beta_1 \approx 0$, por lo que concluiríamos que en un inicio no hay distinción por género pero en el tiempo se le acumulan más puntos a los hombres ($\beta_3 > 0$) que a las mujeres. En promedio las mujeres ganan 0.76 puntos por año, mientras que los hombres 0.76+0.25=1.01 puntos por año.

Ejemplo: Salarios por género y antigüedad (cont.)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.18630    0.53856   9.630 1.76e-13
z            0.68888    0.71298   0.966    0.338
t            0.76163    0.03105  24.530  < 2e-16
zt           0.24965    0.04478   5.574 7.40e-07

Residual standard error: 1.148 on 56 degrees of freedom
Multiple R-squared: 0.9664,     Adjusted R-squared: 0.9646
F-statistic: 537.6 on 3 and 56 DF,  p-value: < 2.2e-16
```

Estadísticamente $\beta_1 \approx 0$, por lo que concluiríamos que en un inicio no hay distinción por género pero en el tiempo se le acumulan más puntos a los hombres ($\beta_3 > 0$) que a las mujeres. En promedio las mujeres ganan 0.76 puntos por año, mientras que los hombres 0.76+0.25=1.01 puntos por año.

Principio de Herencia (de John A. Nelder)

Si en un modelo la interacción de cierto orden incluye, por ejemplo, al factor *A*, entonces el modelo debe incluir el efecto principal del factor *A*.

–ebz

Ejemplo: Salarios por género y antigüedad (cont.)

En el modelo anterior

$$y = \beta_0 + \beta_1 z + \beta_2 t + \beta_3 zt + \epsilon$$

donde z es una variable indicadora, la *significancia* del estimador $\hat{\beta}_1$ ($\beta_1 \neq 0$) indicaría la *desigualdad de las ordenadas al origen*. Similarmente, la *significancia* del estimador $\hat{\beta}_3$ ($\beta_3 \neq 0$) sugiere la *desigualdad de pendientes* de ambas rectas.

Variables categóricas

Una *variable categórica* es aquella cuyos valores corresponden a categorías. E.g., turno: matutino, vespertino y nocturno; experiencia: novato, medio y experto; proveedor zona: norte, sur, este, oeste; etc.
Numéricamente, las variables categóricas dan lugar una o más variables indicadoras.
Una manera de codificar la variable categórica x con $k+1$ categorías es construyendo k variables indicadoras $z_1, \ldots, z_k$ de la siguiente manera:

categorías x	z_1	z_2	$\cdots$	z_k
1	0	0	$\cdots$	0
2	1	0	$\cdots$	0
3	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$k+1$	0	0	$\cdots$	1

En ocasiones los modelos se refieren a las variables categóricas como *factores*.

Ejemplo: Consumo doméstico de electricidad[1]

Una compañía generadora de electricidad estudia el consumo doméstico como función del tamaño y del tipo de aire acondicionado en la casa-habitación. Sea y el consumo (kw-hr), x el tamaño de la casa (m^2) y C el tipo (4) del aire acondicionado: sin aire, unidades de ventana, bomba y central.

Tipo de aire acondicionado	z_1	z_2	z_3
Sin aire	0	0	0
Unidades de ventana	1	0	0
Unidad de bomba	0	1	0
Unidad central	0	0	1

Se podría construir un modelo lineal como

$$y = \alpha + \beta x + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \epsilon$$

El modelo anterior se conoce como un *modelo de covarianza* al incluir ambos tipos de variables: *variables categóricas* (tipo de aire acondicionado) y *regresores continuos* (superficie de la casa).

[1] Montgomery and Peck (1992)

Ejemplo: Consumo doméstico de electricidad (cont.)

Luego, dependiendo del tipo de aire acondicionado, los correspondientes modelos serían

Tipo de aire acondicionado	Modelo
Sin aire	$y = \alpha + \beta x + \epsilon$
Unidades de ventana	$y = (\alpha + \gamma_1) + \beta x + \epsilon$
Unidad de bomba	$y = (\alpha + \gamma_2) + \beta x + \epsilon$
Unidad central	$y = (\alpha + \gamma_3) + \beta x + \epsilon$

Cuando los efectos de pasar de un estado a otro de una de los factores no son similares se dice que las variables (regresores, factores) *interactúan*.

Ejemplo: Consumo doméstico de electricidad (cont.)

Luego, los correspondientes modelos serían:

Tipo de aire acondicionado	Modelo
Sin aire	$y = \alpha + \beta x + \epsilon$
Unidades de ventana	$y = (\alpha + \gamma_1) + \beta x + \epsilon$
Unidad de bomba	$y = (\alpha + \gamma_2) + \beta x + \epsilon$
Unidad central	$y = (\alpha + \gamma_3) + \beta x + \epsilon$

que implican un consumo fijo dependiendo del tipo de aire acondicionado.
Si a su vez el consumo por tipo de aire acondicionado depende del (*interacciona con*) tamaño de la casa, el modelo general queda como

$$y = \alpha + \beta x + \gamma_1 z_1 + \gamma_2 z_2 + \gamma_3 z_3 + \delta_1 x z_1 + \delta_2 x z_2 + \delta_3 x z_3 + \epsilon$$

que da lugar a los modelos

Tipo de aire acondicionado	Modelo
Sin aire	$y = \alpha + \beta x + \epsilon$
Unidades de ventana	$y = (\alpha + \gamma_1) + (\beta + \delta_1) x + \epsilon$
Unidad de bomba	$y = (\alpha + \gamma_2) + (\beta + \delta_2) x + \epsilon$
Unidad central	$y = (\alpha + \gamma_3) + (\beta + \delta_3) x + \epsilon$

–ebz

Modelo de Análisis de Varianza

Un *modelo de análisis de varianza* es un modelo lineal donde las variables *independientes* son todas categóricas.

Modelo de análisis de varianza de una vía – Igualdad de medias

El modelo más sencillo se le conoce como *modelo de análisis de varianza de una vía* y es aquel con una sola variable (categórica) independiente llamada *factor*, con *a niveles* o categorías. El modelo puede representarse como

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$$

con $j = 1, \ldots, n_i$; $i = 1, \ldots, a$. Note que la *respuesta media* para observaciones bajo el i-ésimo *tratamiento* es

$$\mathbb{E}[y_{ij}] = \mathbb{E}\left[\mu + \tau_i + \epsilon_{ij}\right] = \mu + \tau_i \equiv \mu_i$$

Una pregunta de interés es: *¿Son las respuestas similares bajo los distintos tratamientos?*. O bien, *¿Son los tratamientos iguales?* Formalmente, la pregunta la podemos plantear con la siguiente hipótesis compuesta:

$$H_0 : \mu_1 = \cdots = \mu_a \quad \textit{vs.} \quad Ha : \mu_i \neq \mu_j, \text{ para al menos un } i \neq j$$

Modelo de análisis de varianza de una vía – Igualdad de medias

Note que

$$y_{ij} = \bar{y}_{..} + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ij} - \bar{y}_{i.})$$

Se puede mostrar que

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2}_{\text{SCT}_{\text{Corregido}}} = \underbrace{\sum_{i=1}^{a} n_i (\bar{y}_{i.} - \bar{y}_{..})^2}_{\text{SC}_{\text{Tratamientos}}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}_{\text{SC}_{\text{Residuales}}}$$

y que bajo la *hipótesis nula* $\text{H}_0 : \mu_1 = \cdots \mu_a$

$$F_{\text{obs}} = \frac{\sum_{i=1}^{a} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 / (a-1)}{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 / (N-a)} \sim F_{a-1,N-a}$$

donde a es el número de niveles o categorías y $N = \sum_{i=1}^{a} n_i$, el número total de observaciones. Si es el caso que

$$F_{\text{obs}} > F_{\text{tablas}}(1-\alpha; a-1, N-a)$$

se rechaza la hipótesis nula (al nivel de significancia α), concluyendo que al menos las respuestas medias de 2 niveles son diferentes. El problema entonces es identificar qué categorías son las distintas.

–ebz

Modelo de análisis de varianza de una vía – Igualdad de medias

Modelo de análisis de varianza de una vía – Igualdad de medias

Los cálculos anteriores se presentan en una *tabla de análisis de varianza*

Fuente de variación	Grados de libertad (g.l.)	Suma de cuadrados (SC)	Cuadrados medios (CM)	Estadístico F	Valor p
Tratamientos	$a-1$	SC_{Trat}	$SC_{Trat}/(a-1)$	CM_{Trat}/CM_{Res}	·
Residuales	$N-a$	SC_{Res}	$SC_{Res}/(N-a)$		
Total $_{Corregido}$	$N-1$	SC_{Tot}			

Para ver qué medias son distintas, se calcula el intervalo de confianza para *todas* las diferencias de medias $\mu_i - \mu_\ell$, y se verifica si el intervalo incluye o no el cero. Un intervalo del $100(1-\alpha)$ % para $\mu_i - \mu_\ell$ está dado por

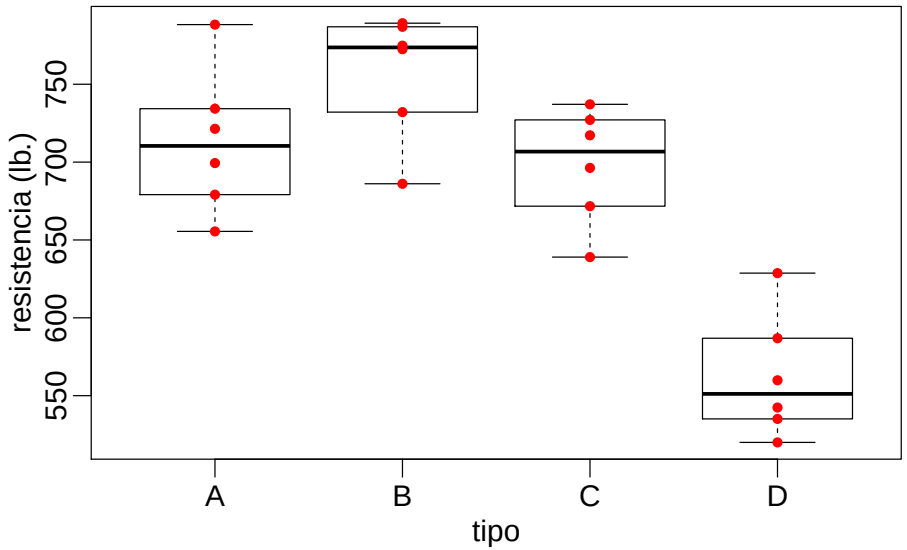
$$(\hat{\mu}_i - \hat{\mu}_\ell) \pm t_{1-\alpha/2;N-a}\, s\sqrt{\frac{1}{n_i}+\frac{1}{n_\ell}}, \qquad i,\ell = 1,\ldots,a;\ i \neq \ell$$

donde $\hat{\mu}_i = \bar{y}_{i\cdot}$, $\hat{\mu}_\ell = \bar{y}_{\ell\cdot}$ y $s^2 = CM_{Res} = SC_{Res}/(N-a)$.

Ejemplo: Resistencia de cajas de cartón[2]

Se compararon la resistencia a la compresión (lb.) de cuatro tipos distintos de cajas.

Tipo	Resistencia						$\bar{y}_{i\cdot}$	s_i
A	655.5	788.3	734.3	721.4	679.1	699.4	713.00	46.55
B	789.2	772.5	786.9	686.1	732.1	774.8	756.93	40.34
C	737.1	639.0	696.3	671.7	717.2	727.1	698.07	37.20
D	535.1	628.7	542.4	559.9	586.9	520.0	562.17	39.86
						$\bar{y}_{\cdot\cdot} =$	682.50	



[2] Devore (1995), Sec. 10.1

Ejemplo: Resistencia de cajas de cartón (cont.)

Salida de R

```
 Response: resistencia

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   713.00      16.79  42.462  < 2e-16
tipoB          43.93      23.75   1.850   0.0791
tipoC         -14.93      23.75  -0.629   0.5366
tipoD        -150.83      23.75  -6.352 3.37e-06

Residual standard error: 41.13 on 20 degrees of freedom
Multiple R-squared: 0.7898,     Adjusted R-squared: 0.7583
F-statistic: 25.06 on 3 and 20 DF,  p-value: 5.593e-07

Analysis of Variance Table
          Df Sum Sq Mean Sq F value    Pr(>F)
tipo       3 127158   42386  25.055 5.593e-07
Residuals 20  33834    1692
Total     23 160992
```

Salida de Minitab

```
                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level  N    Mean  StDev  -----+---------+---------+---------+----
A      6  713.00  46.55                       (----*----)
B      6  756.93  40.34                             (----*----)
C      6  698.07  37.20                        (----*----)
D      6  562.17  39.86  (----*----)
                          -----+---------+---------+---------+----
                             560       630       700       770
```

Modelo de Análisis de Varianza de Dos Vías

El *modelo de análisis de varianza de dos vías* es aquel con dos variables (categóricas) independientes o factores A y B, con a y b niveles respectivamente. El modelo puede representarse como

$$y_{ijk} = \mu + \tau_i + \eta_j + \tau\eta_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim \mathrm{N}(0, \sigma^2)$$

con $k = 1, \ldots, n_{ij}$; $i = 1, \ldots, a$; $j = 1, \ldots, b$, donde τ_i, η_j y $\tau\eta_{ij}$ denotan el *efecto* del factor A al nivel i, el efecto del factor B al nivel j y el efecto de la *interacción* del factor A con el factor B.

La *respuesta media* para observaciones bajo el i-ésimo nivel del factor A y el j-ésimo nivel del factor B es:

$$\mathbb{E}[y_{ijk}] = \mathbb{E}\left[\mu + \tau_i + \eta_j + \tau\eta_{ij} + \epsilon_{ij}\right] = \mu + \tau_i + \eta_j + \tau\eta_{ij} \equiv \mu_{ijk}$$

Modelo de Análisis de Varianza de Dos Vías

Nuevamente, note que las desviaciones respecto a la *gran media* pueden descomponerse como

$$(y_{ijk} - \bar{y}_{\cdots}) = (\bar{y}_{i\cdot} - \bar{y}_{\cdots}) + (\bar{y}_{\cdot j} - \bar{y}_{\cdots}) + (y_{ij\cdot} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdots}) + (y_{ijk} - \bar{y}_{ij\cdot})$$

dando lugar a la descomposición de suma de cuadrados

$$\underbrace{\sum_{ijk}(y_{ijk} - \bar{y}_{\cdots})^2}_{\text{SCT}_{\text{Corr}}} = \underbrace{\sum_{ij} n_{ij}(\bar{y}_{i\cdot} - \bar{y}_{\cdots})^2}_{\text{SC}_{\text{A}}} + \underbrace{\sum_{ij} n_{ij}(\bar{y}_{\cdot j} - \bar{y}_{\cdots})^2}_{\text{SC}_{\text{B}}} + \underbrace{\sum_{ijk}(y_{iju} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdots})^2}_{\text{SC}_{\text{AB}}} + \underbrace{\sum_{ijk}(y_{ijk} - \bar{y}_{ij\cdot})^2}_{\text{SC}_{\text{Res}}}$$

En un análisis de varianza de 2 vías, se pueden probar las siguientes hipótesis:

$$\begin{aligned}
&a) \text{ Igualdad de medias factor } A \quad &\Rightarrow \quad &H_0 : \tau_1 = \cdots \tau_a = 0 \\
&b) \text{ Igualdad de medias factor } B \quad &\Rightarrow \quad &H_0 : \eta_1 = \cdots \eta_b = 0 \\
&c) \text{ Interacciones } AB \text{ nulas} \quad &\Rightarrow \quad &H_0 : \tau\eta_{11} = \cdots \tau\eta_{ab} = 0
\end{aligned}$$

con los correspondientes estadísticos F dados en la siguiente tabla de análisis de varianza.

Modelo de Análisis de Varianza de Dos Vías

Tabla de Análisis de Varianza

Fuente	g.l.	SC	CM	F	p
Factor A	$\nu_A = a - 1$	SC_A	SC_A/ν_A	$F_A = \text{CM}_\text{A}/\text{CM}_\text{Res}$	·
Factor B	$\nu_B = b - 1$	SC_B	SC_B/ν_B	$F_B = \text{CM}_\text{B}/\text{CM}_\text{Res}$	·
Interacción AB	$\nu_{AB} = (a-1)(b-1)$	SC_AB	$\text{SC}_\text{AB}/\nu_{AB}$	$F_{AB} = \text{CM}_\text{AB}/\text{CM}_\text{Res}$	·
Residuales	$\nu_\text{Res} = \sum_{ij}(n_{ij} - 1)$	SC_Res	$\text{SC}_\text{Res}/\nu_\text{Res}$		
Total $_\text{Corregido}$	$\sum_{ij} n_{ij} - 1$	SC_Total			

En general, rechazamos las hipótesis nulas H_0 si

$$F_\text{Trat} > F_\text{tablas}(1 - \alpha; \nu_\text{Trat}, \nu_\text{Res})$$

Rechazando alguna de las hipótesis anteriores lo siguiente por averiguar es cuál efecto principal (τ_i o η_j) es distinto de cero, o cuál de las interacciones $\tau\eta_{ij}$no es nula. Esto es lo que se conoce en el análisis de varianza como *comparaciones múltiples*.

Ejemplo: Brillantez de fotografías[3]

Se compara la brillantez (**brillo**) de fotografías resultado de tres distintos procesos (**proc**: A, B y C) y de tres distintas compañías (**comp**: Agfa, Fuji, Kodak).

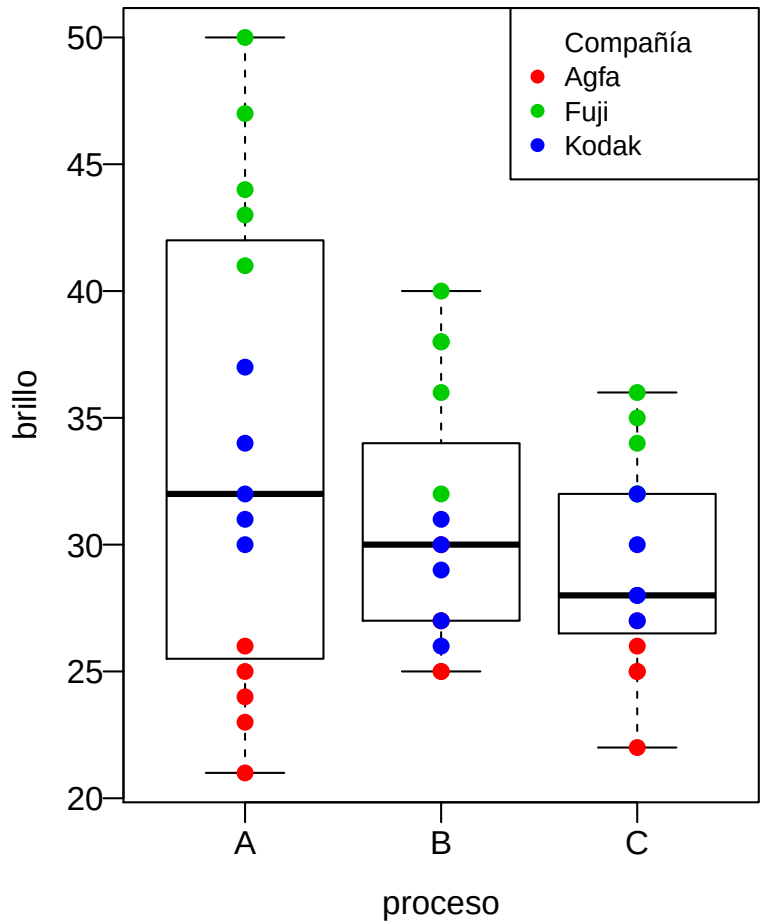
Brillantez

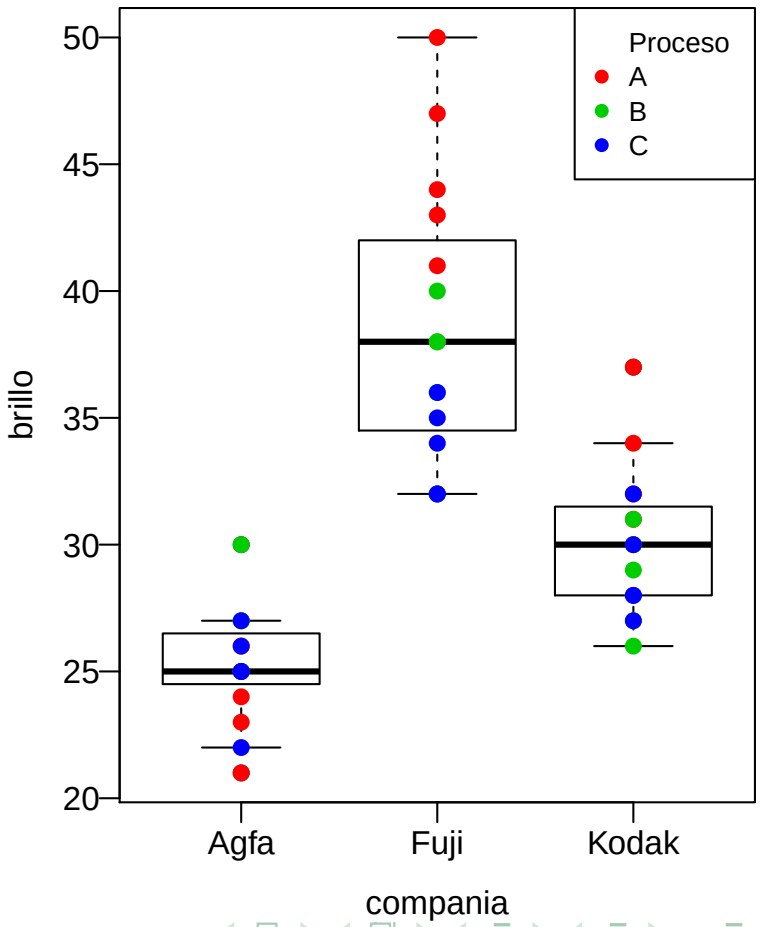
Agfa			Fuji			Kodak		
A	B	C	A	B	C	A	B	C
23	27	25	43	32	32	32	26	28
24	30	27	41	38	32	34	29	28
25	25	26	44	38	36	31	27	27
21	25	22	50	40	35	30	30	30
26	27	25	47	36	34	37	31	32

[3] Bernd Pape; http://lipas.uwasa.fi/ bepa/Riippu6.pdf, 27 Jul, 2018

Ejemplo: Brillantez de fotografías (cont.)

Brillantez por factor



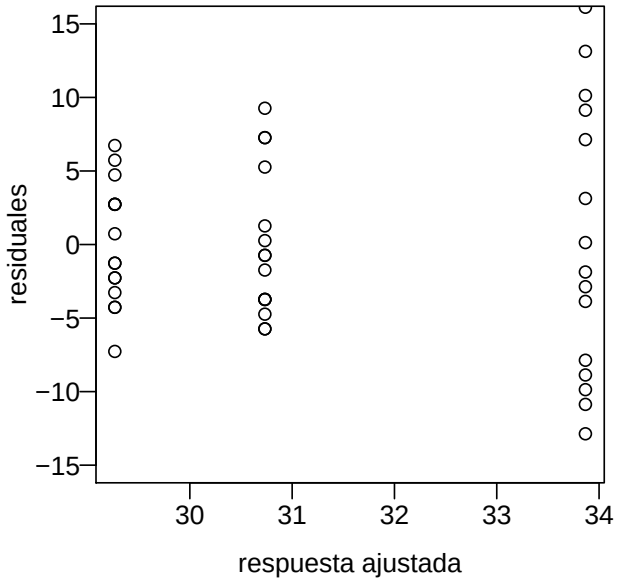


−ebz

Ejemplo: Brillantez de fotografías (cont.)

$$\text{brillo} = \mu + \text{proc} + \epsilon$$

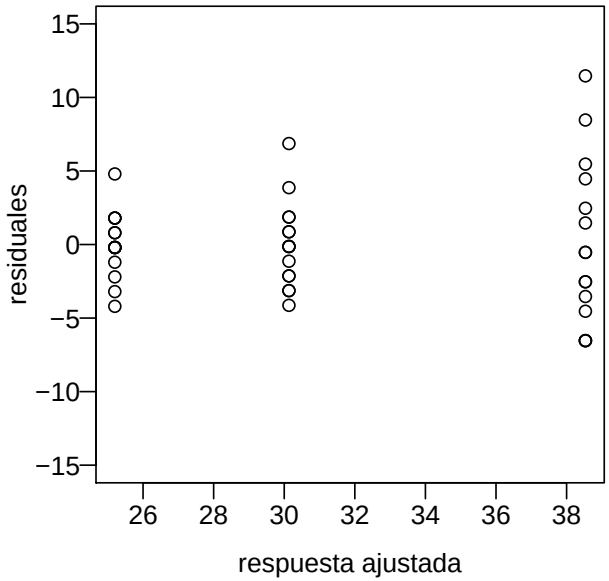
```
> anova(lm(brillo ~ proc, data=datos))

Response: brillo
          Df  Sum Sq Mean Sq F value Pr(>F)
proceso    2  165.64  82.822  1.9117 0.1605
Residuals 42 1819.60  43.324
```



Ejemplo: Brillantez de fotografías (cont.)

$$\text{brillo} = \mu + \text{comp} + \epsilon$$

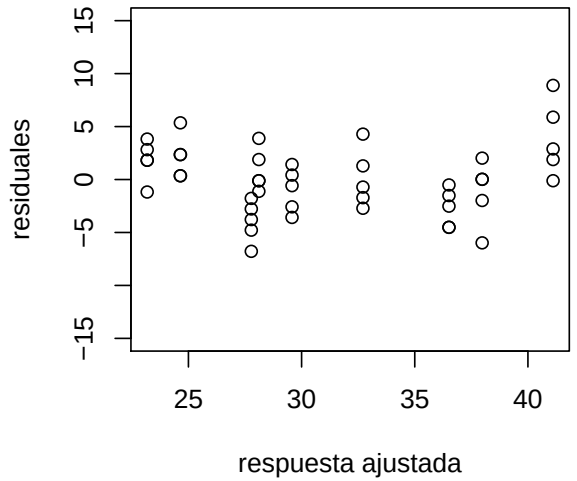
```
> anova(lm(brillo ~ comp, data=datos))

Response: brillo
          Df  Sum Sq Mean Sq F value    Pr(>F)
compania   2 1363.38  681.69   46.04 2.591e-11
Residuals 42  621.87   14.81
```



Ejemplo: Brillantez de fotografías (cont.)

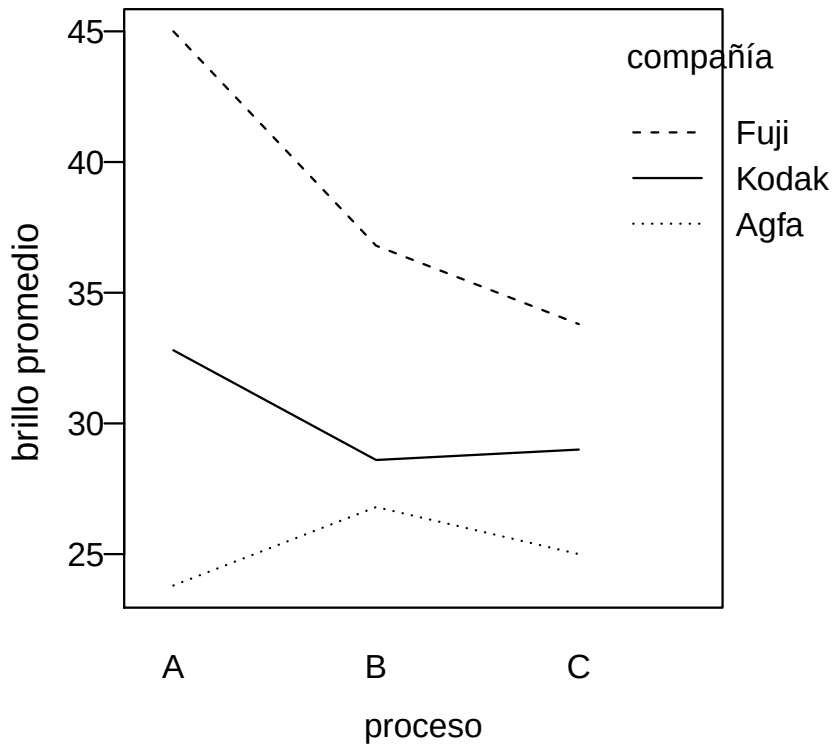
$$\text{brillo} = \mu + \text{proc} + \text{comp} + \epsilon$$

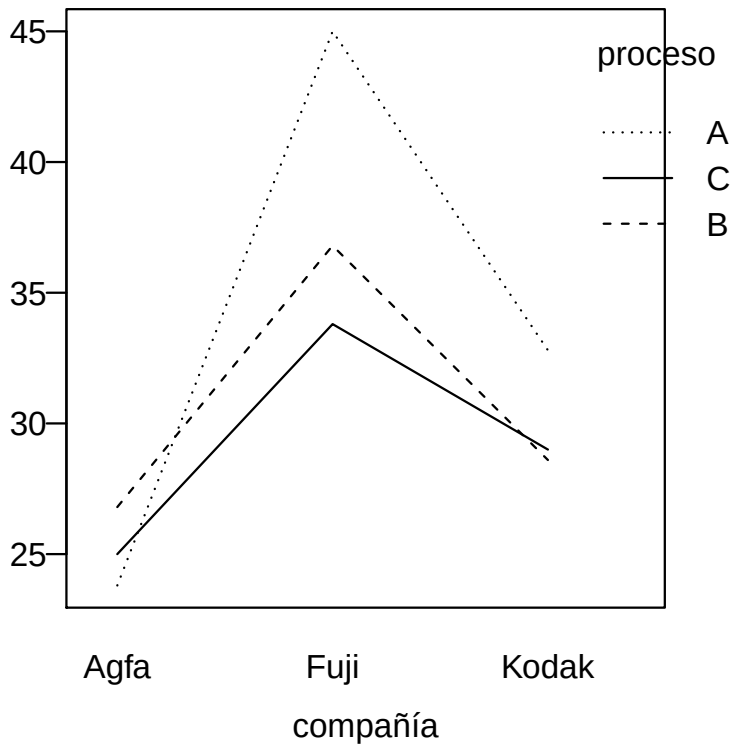
```
> anova(lm(brillo ~ proc + comp, data=datos))

Response: brillo
          Df  Sum Sq Mean Sq F value    Pr(>F)
proceso    2  165.64   82.82  7.2616   0.00204
compania   2 1363.38  681.69 59.7681 9.639e-13
Residuals 40  456.22   11.41
```



Ejemplo: Brillantez de fotografías (cont.)

Gráfica de interacciones





Ejemplo: Brillantez de fotografías (cont.)

$$\text{brillo} = \mu + \text{proc} + \text{comp} + \text{proc} : \text{comp} + \epsilon$$

```
Call:
lm(formula = brillo ~ proceso * compania, data = dat)

Residuals:
Min      1Q Median      3Q     Max
-4.8    -1.8     0.2     1.2     5.0

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)              23.800      1.078  22.077  < 2e-16
procesoB                  3.000      1.525   1.968  0.05684
procesoC                  1.200      1.525   0.787  0.43638
companiaFuji             21.200      1.525  13.905 4.79e-16
companiaKodak             9.000      1.525   5.903 9.36e-07
procesoB:companiaFuji   -11.200      2.156  -5.194 8.28e-06
procesoC:companiaFuji   -12.400      2.156  -5.751 1.50e-06
procesoB:companiaKodak   -7.200      2.156  -3.339  0.00196
procesoC:companiaKodak   -5.000      2.156  -2.319  0.02619

Residual standard error: 2.411 on 36 degrees of freedom
Multiple R-squared:  0.8946,Adjusted R-squared:  0.8712
F-statistic:  38.2 on 8 and 36 DF,  p-value: 2.491e-15
```

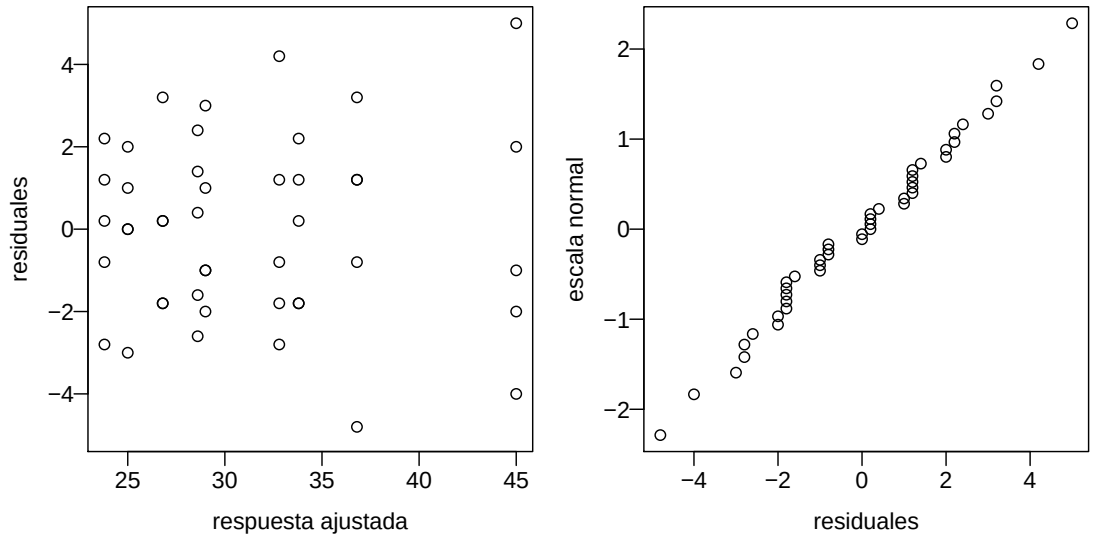
–ebz

Ejemplo: Brillantez de fotografías (cont.)

$$\text{brillo} = \mu + \text{proc} * \text{comp} + \epsilon$$

```
> anova(lm(brillo ~ proc * comp, data=datos))

Response: brillo
Df  Sum Sq Mean Sq F value    Pr(>F)
proceso           2  165.64   82.82  14.252 2.759e-05
compania          2 1363.38  681.69 117.308 < 2.2e-16
proceso:compania  4  247.02   61.76  10.627 8.633e-06
Residuals        36  209.20    5.81
```

Gráfica de residuales



–ebz