Lab6

Sebastián Sánchez Sandí

2025-07-11

Preparación:

Primero cargamos las librerías que vamos a utilizar.

```
library(lattice)
```

Ahora cargamos los datos y les damos el formato correcto para comenzar a trabajar. Ademas creamos la variable de IMC.

```
load("100metros.Rdata")
str(base)
```

```
## 'data.frame':    66 obs. of  5 variables:
##  $ estatura: int  154 158 151 168 164 161 180 174 174 168 ...
##  $ peso    : num  60.9 55.3 53.6 53.8 53.3 59.6 61.5 80.4 72.4 51.3 ...
##  $ tiempo  : num  12.05 6.81 12.08 7.89 10.38 ...
##  $ calent  : Factor w/ 3 levels "A","B","C": 1 1 1 1 1 1 1 1 1 1 ...
##  $ salida  : Factor w/ 2 levels "-","+": 2 2 2 2 1 1 1 1 1 1 ...
```

```
base$imc = base$peso / base$estatura^2
```

Cuando se toma en cuenta solamente los factores, el experimento cuenta con 6 tratamientos.

```
table(base$calent, base$salida)
```

```
##
##      -  +
##   A 11 11
##   B 11 11
##   C 11 11
```

Para cada tratamiento contamos con 11 replicas.

Linealidad

Este apartado queda fuera del alcance del laboratorio.

Variabilidad de la respuesta

Primero obtenemos la varianza de la variable de respuesta bajo cada tratamiento.

```
v = with(data = base, tapply(tiempo, list(calent, salida), var))
v
```

```
##           -         +
## A  6.286529 13.562693
## B 13.628041 10.741678
## C  8.447528  9.812103
```

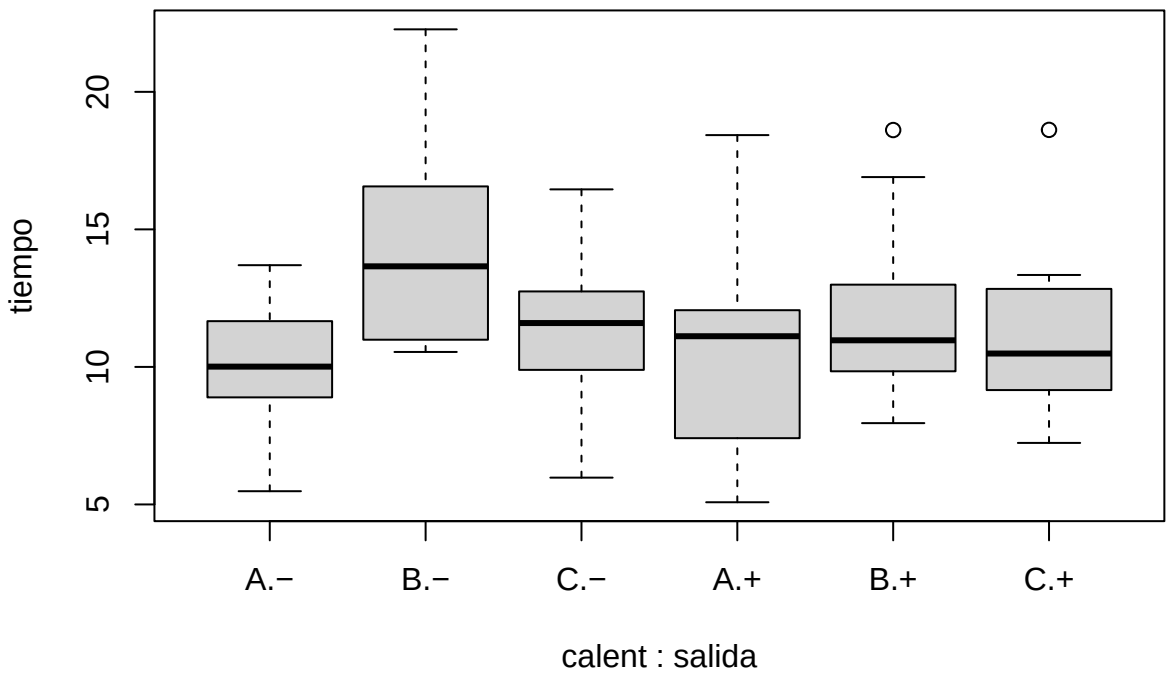
Calculamos la estimacion de la varianza dentro de los tratamientos como el cuadrado medio residual. Como el experimento es balanceado podemos realizar la estimacion como el promedio de las varianzas.

```
CMRes = mean(v)
CMRes
```

```
## [1] 10.4131
```

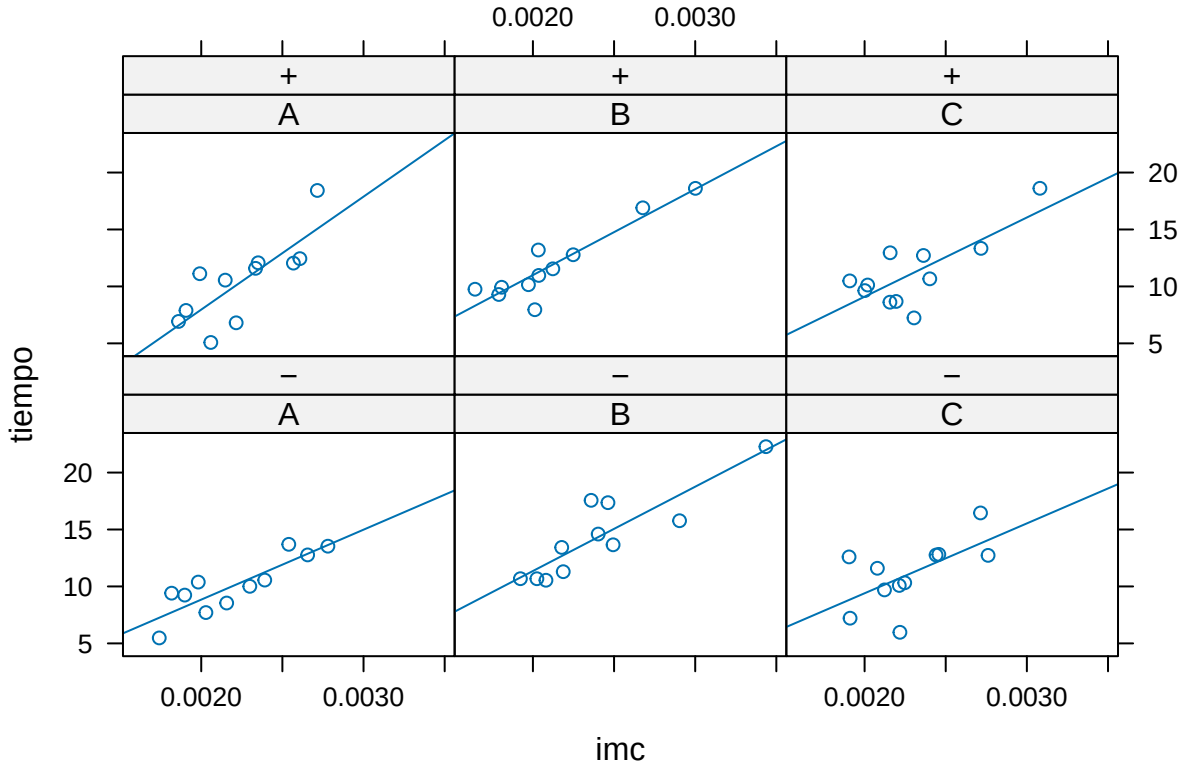
Se estima que la varianza de los tratamientos es de 10.41. Ahora para poder visualizar de forma grafica esta variabilidad construimos un grafico de cajas.

```
boxplot(tiempo~calent*salida, data = base)
```



Aca podemos visualizar la variabilidad de la variable de respuesta dentro de cada tratamiento sin tener en cuenta la covariable. Constuimos un grafico de puntos del tiempo contra el imc con una linea de regeresion para cada tratamiento para visualizar de forma grafica la varianza dentro de cada tratamiento.

```
xyplot(tiempo~imc|calent+salida, type = c("r", "p"), data = base)
```



Como podemos observar, dentro de cada tratamiento los residuales parecen ser pequeños. Los residuales son las distancias entre cada punto y su linea de regresion.

Inclusion de la covariable

Ahora construimos un modelo donde se tomen en cuenta la covariable con el fin de obtener los residuales. Este modelo puede ser con interaccion o sin interaccion.

```
mod1 = lm(tiempo~calent*salida+imc, data = base)
r1 = mod1$residuals
```

Como se definio anteriormente, los residuales representan las distancias entre cada punto y su recta correspondiente. Para poder calcular el cuadrado medio residual bajo este modelo de forma manual se necesita identificar la cantidad de parametros. Sea p igual al numero de parametros del modelo y sea n la cantidad de observaciones. Entonces el cuadrado medio residual se calcula como $CMRes = \frac{\sum r^2}{n-p}$. En este caso el modelo se representa como: $\mu_{ij} = \mu + \alpha_1 + \alpha_2 + \beta_1 + (\alpha\beta)_{11} + (\alpha\beta)_{21} + \delta_1 X$.

```r
n = 66
p = 7
CMResCov = sum(r1^2) / (n - p)
```

Como podemos ver el cuadrado medio residual se redujo de 10.41 a 3.75.

Prueba formal

Verificamos con finalidad didactica ciertas propiedades de la prueba ANOVA.

```r
mod2a = lm(tiempo~calent*salida, data = base)
mod2b = lm(tiempo~salida*calent, data = base)

print(anova(mod2a))
```

```
## Analysis of Variance Table
##
## Response: tiempo
##               Df Sum Sq Mean Sq F value  Pr(>F)
## calent         2  93.77  46.884  4.5024 0.01507 *
## salida         1   7.53   7.527  0.7229 0.39859
## calent:salida  2  25.62  12.811  1.2303 0.29947
## Residuals     60 624.79  10.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
print(anova(mod2b))
```

```
## Analysis of Variance Table
##
## Response: tiempo
##               Df Sum Sq Mean Sq F value  Pr(>F)
## salida         1   7.53   7.527  0.7229 0.39859
## calent         2  93.77  46.884  4.5024 0.01507 *
## salida:calent  2  25.62  12.811  1.2303 0.29947
## Residuals     60 624.79  10.413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como podemos ver, el orden en que se ingresen los factores es indiferente ya que al final vamos a obtener las mismas probabilidades. Sin embargo, cuando se ingresa la covariable el orden toma importancia. Como la probabilidad de cometer error tipo 1 es mayor a 0.05, entonces no se rechaza la hipotesis de no interaccion. Se espera que no haya interaccion entre los factores.

Como se detecto que no existe interaccion, se realiza la prueba de diferencia de medias con un modelo sin interaccion.

```r
mod3 = lm(tiempo~calent+salida, data = base)
print(anova(mod3))
```

```
## Analysis of Variance Table
##
## Response: tiempo
##           Df Sum Sq Mean Sq F value  Pr(>F)
## calent     2  93.77  46.884  4.4692 0.01537 *
## salida     1   7.53   7.527  0.7175 0.40022
## Residuals 62 650.41  10.490
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como la probabilidad de cometer error tipo 1 es menor a 0.05, entonces rechazamos la hipotesis nula. Se espera que al menos una media bajo un nivel de calentamiento sea distinta.

Ahora incluimos la covariable y probamos distintos ordenes para observar el cambio de probabilidades y porque es importante el orden.

```
mod4a = lm(tiempo~calent+salida+imc, data = base)
mod4b = lm(tiempo~imc+calent+salida, data = base)

print(anova(mod4a))
```

```
## Analysis of Variance Table
##
## Response: tiempo
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## calent     2  93.77   46.88  12.9096 2.112e-05 ***
## salida     1   7.53    7.53   2.0726    0.1551
## imc        1 428.87  428.87 118.0888 6.736e-16 ***
## Residuals 61 221.54    3.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
print(anova(mod4b))
```

```
## Analysis of Variance Table
##
## Response: tiempo
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## imc        1 444.35  444.35 122.3515 3.268e-16 ***
## calent     2  85.50   42.75  11.7705 4.753e-05 ***
## salida     1   0.32    0.32   0.0878     0.768
## Residuals 61 221.54    3.63
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La forma correcta de escribir el modelo es $\mu_{ij} = \beta_0 + \beta_1 X + \alpha_i + \gamma_j$ donde α_i es el efecto el i-esimo del nivel del factor de diseno y γ_j es el efecto del j-esimo nidel del otro factor.

Como vimos que si afecta el orden entonces lo acomodamos de la forma adecuada. Para continuar con el analisis necesitamos plantear dos modelos. Un modelo grande denominado Ω el cual incluye el factor de diseño y otro modelo pequeño denominado ω el cual excluye el factor de diseño.

```r
modGrande = lm(tiempo~imc+calent+salida, data = base)
modPeq = lm(tiempo~imc+salida, data = base)
```

Ahora, necesitamos la suma de cuadrados marginal la cual se define como la resta de la suma de cuadrados residual del modelo grande menos la del modelo pequeño. $SCResMarg = SCRes_{\omega} - SCRes_{\Omega}$. Esta representa la parte de la variabilidad que es explicada por calentamiento cuando entra despues de las otras dos variables.

```r
SCResGrande = anova(modGrande)[4, 2]
SCResPeq = anova(modPeq)[3, 2]
SCResMarg = SCResPeq - SCResGrande
SCResMarg
```

```
## [1] 85.51257
```

Entonces 85.51 unidades de la variabilidad son explicadas por el factor de diseño, en este caso calentamiento. Ahora construimos el estadistico F para realizar la prueba de hipotesis. Este se construye como $F = \frac{\frac{SCRes_{\omega} - SCRes_{\Omega}}{df_{\omega} - df_{\Omega}}}{\frac{SCRes_{\Omega}}{df_{\Omega}}}$

```r
dfGrande = anova(modGrande)[4, 1]
dfPeq = anova(modPeq)[3, 1]

f = (SCResMarg / (dfPeq - dfGrande)) / (SCResGrande / dfGrande)
f
```

```
## [1] 11.77287
```

El estadistico f es igual a 11.77 que podemos ver que coincide con el estadistico F en el anova cuando lo ralizamos en el orden correcto. Ahora buscamos la probabilidad de ver este estadistico en la distribucion f la cual deberia ser muy pequeña.

```r
pf(f, dfPeq - dfGrande, dfGrande, lower.tail = F)
```

```
## [1] 4.745278e-05
```

Como podemos ver es 0.00005 lo cual es diminuto. Recordando, la hipotesis nula es que los dos modelos explican lo mismo. Esto es equivalente a decir que los efectos de los niveles de calentamiento son iguales a 0, lo que es equivalente a decir que las medias del tiempo son iguales para todos los niveles de calentamiento. Por lo tanto, rechazamos la hipotesis nula. Se espera que los dos modelos no expliquen lo mismo, es decir que al menos un efecto de algun nivel de calentamiento sea distinto a 0.

La forma corta de realizar todo este analisis es el siguiente. Al utilizar el comando drop1 el orden en el que se escribio el modelo es indiferente.

```r
drop1(mod4b, test = "F")
```

```
## Single term deletions
##
## Model:
## tiempo ~ imc + calent + salida
```

```
##        Df Sum of Sq    RSS     AIC  F value    Pr(>F)
## <none>              221.54  89.922
## imc     1    428.87 650.41 159.004 118.0888 6.736e-16 ***
## calent  2     85.51 307.05 107.466  11.7729 4.745e-05 ***
## salida  1      0.32 221.86  88.017   0.0878     0.768
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede utilizar este comando para evaluar interaccion tambien.

```
drop1(mod1, test = "F")
```

```
## Single term deletions
##
## Model:
## tiempo ~ calent * salida + imc
##               Df Sum of Sq    RSS     AIC  F value   Pr(>F)
## <none>                     221.10  93.793
## imc            1    403.68 624.79 160.352 107.7194 6.27e-15 ***
## calent:salida  2      0.43 221.54  89.922   0.0578   0.9439
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Coincide con el analisis que realizamos anteriormente.