Regresión Avanzada
Universidad Austral

PhD. Débora Chan

Junio - Julio de 2023

Organización

Predictores categóricos en el modelo lineal

Variables Dummy

Cuando se introduce una variable categórica como predictor, un nivel se considera el de referencia (normalmente codificado como 0) y el resto de niveles se comparan con él. En el caso de que el predictor categórico tenga más de dos niveles, se generan lo que se conoce como variables dummy, que son variables creadas para cada uno de los niveles del predictor categórico y que pueden tomar el valor de 0 o 1.

Cada vez que se emplee el modelo para predecir un valor, solo una variable dummy por predictor adquiere el valor 1 (la que coincida con el valor que adquiere el predictor en ese caso) mientras que el resto se consideran 0.

Interpretación de los coeficientes

Valores relativos

El valor del coeficiente parcial de regresión β_i de cada variable dummy indica el porcentaje promedio en el que influye dicho nivel sobre la variable dependiente Y en comparación con el nivel de referencia de dicho predictor.

Analicemos un caso práctico

Supongamos que la variable respuesta a predecir es el volumen de un fruto de damasco y como predictoras tenemos la longitud de la hoja y la variedad del fruto con tres niveles A, B y C. La ecuación completa sería:

$$volumen = \beta_0 + \beta_1 long\_hoja + \beta_B + \beta_C$$

Si el fruto es de variedad A, las dos dummies son 0, si es de variedad B la β_2 es distinta de cero y las otras dos son cero.

Una predictora categórica con dos niveles

Usamos los datos Salaries de la biblioteca carData de R.

```
library(carData); library(dplyr)
mod1< − lm(salary ∼ sex, data = Salaries)
summary(mod1)
```

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	101002.4103	4809.3860	21.00	0.0000	Resid-
sexMale	14088.0087	5064.5792	2.78	0.0057	

ual standard error: 30030 on 395 degrees of freedom

Multiple R-squared: 0.01921, Adjusted R-squared: 0.01673

F-statistic: 7.738 on 1 and 395 DF, p-value: 0.005667

Una predictora categórica con dos niveles

Si queremos cambiar el nivel de base contra el que se contraste

```
# veamos cómo crea automáticamente la dummy
contrasts(Salaries$sex)

#S i se desea recodificar el campo Sexo
Salaries <- Salaries %>%
mutate(sex = relevel(sex, ref = "Male"))
```

```
# veamos cómo crea automáticamente la dummy
       Male
Female    0
  Male    1
```

Una predictora categórica con dos niveles

```
# veamos cómo queda el modelo cuando cambiamos el nivel de base
mod2 <- lm(salary ~ sex, data = Salaries)
summary(mod2)
```

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   115090.4190  1587.3782   72.50   0.0000
sexFemale     -14088.0087  5064.5792   -2.78   0.0057
Residual standard error: 30030 on 395 degrees of freedom
Multiple R-squared: 0.01921, Adjusted R-squared: 0.01673
F-statistic: 7.738 on 1 and 395 DF, p-value: 0.005667
```

Una predictora categórica con más de dos niveles

Una variable categórica con n niveles, estas son transformadas a n-1 variables. La variable rank de la base Salaries tiene 3 niveles.

```
mod3 <- lm(salary ~ rank , data = Salaries)
summary(mod3)
```

```
                 Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   80775.9851   2887.3126    27.98    0.0000
rankAssocProf 13100.4524   4130.8504     3.17    0.0016
rankProf      45996.1239   3230.5398    14.24    0.0000
Residual standard error: 23630 on 394 degrees of freedom
Multiple R-squared: 0.3943, Adjusted R-squared: 0.3912
F-statistic: 128.2 on 2 and 394 DF, p-value:< 2.2e − 16
```

Una predictora categórica con más de dos niveles

Cómo es la matriz de contraste del factor rank?

```
contr <- model.matrix( rank, data = Salaries)
head(contr[, -1])
```

	rankAssocProf	rankProf
1	0.00	1.00
2	0.00	1.00
3	0.00	0.00
4	0.00	1.00
5	0.00	1.00
6	1.00	0.00

Dos predictoras categóricas

A partir de la salida, identificar la cantidad de niveles de cada variable categórica y el nivel base.

```
mod4 <- lm(salary~ rank + discipline , data = Salaries)
summary(mod4)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	71944.3261	3135.1664	22.95	0.0000
rankAssocProf	13761.5432	3960.6611	3.47	0.0006
rankProf	47843.8386	3111.5521	15.38	0.0000
disciplineB	13760.9570	2296.0309	5.99	0.0000

Residual standard error: 22650 on 393 degrees of freedom
Multiple R-squared: 0.445, Adjusted R-squared: 0.4407
F-statistic: 105 on 3 and 393 DF, p-value: $< 2.2e-16$

Una predictora categórica y otra continua

```
mod5 <- lm(salary ~ yrs.since.phd + discipline , data = Salaries)
summary(mod5)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	80158.3090	3328.3497	24.08	0.0000
yrs.since.phd	1118.5486	105.7674	10.58	0.0000
disciplineB	15784.2299	2733.2579	5.77	0.0000

Residual standard error: 26470 on 394 degrees of freedom
Multiple R-squared: 0.2401, Adjusted R-squared: 0.2362
F-statistic: 62.24 on 2 and 394 DF, p-value: $< 2.2e-16$

Predictoras categóricas y continuas con interacción

Cómo incluir la interacción

- Uno de los supuestos del modelo de regresión lineal múltiple es la de aditividad, de los efectos de las variables predictoras sobre la variable respuesta Y.
- Esto es que variaciones en el predictor X_i son independientes del valor que asuman los restantes predictores.
- Se conoce como efecto de interacción cuando el efecto de un predictor varía dependiendo del valor que adquiera otro predictor del modelo.
- Si esto ocurre, el modelo mejorará su poder explicativo al incluir dicha interacción. Cuando incorporamos un término de interacción, debemos incluir también los predictores individuales que forman la interacción aunque per se no resulten significativos.

Interpretación de modelos (predictores categóricos/ interacción)

Guía de Interpretación

- El análisis de regresión múltiple cuando incorpora una variable categórica para evaluar los efectos causales de dicha variable transformada en dummy, los coeficientes de regresión calculados se interpretan como las diferencias de efecto que ejerce cada una de las categorías dummy respecto de la categoría de referencia.
- En el modelo de regresión se compara el efecto de cada uno de los valores de respuesta o atributos aisladamente con respecto a la categoría de referencia.
- Para interpretar los efectos de las variables que participan en un término de interacción se deben combinar el efecto individual de la variable y el de la interacción.

Predictoras categóricas y continuas con interacción

```
antro$Sexo=factor(antro$Sexo)
mod6 <- lm(Peso~ Estatura*Sexo + Edad_meses , data = antro)
summary(mod6)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-526.1336	13.0620	-40.28	0.0000
Estatura	0.6159	0.0146	42.12	0.0000
Sexo2	88.3520	12.0642	7.32	0.0000
Edad_meses	0.3968	0.0683	5.81	0.0000
Estatura:Sexo2	-0.0721	0.0086	-8.38	0.0000

Residual standard error: 64.13 on 3770 degrees of freedom

Multiple R-squared: 0.8701, Adjusted R-squared: 0.8699

F-statistic: 6310 on 4 and 3770 DF, p-value: $< 2.2e-16$

Organización

ANOVA

El ANOVA es un caso especial de modelo lineal, donde los predictores son variables categóricas.

Problema Inicial: Comparar los Valores medios de varias subpoblaciones La mejor respuesta fue desarrollada por Fisher entre los años 1920 y 1930, conisdera tres o más poblaciones independientes con distribuciones normales de igual varianza. El análisis que desarrollaremos a continuación y se denomina Análisis de la varianza (ADEVA) o, en inglés, *analysis of variance* (ANOVA).

Ejemplo Té

El té es la bebida más usual en el mundo entero después del agua, actualmente se ha difundido mucho el consumo del té verde, dado que se ha encontrado que contiene vitamina B. Recientes avances en métodos de

Ejemplo Té

Los Datos

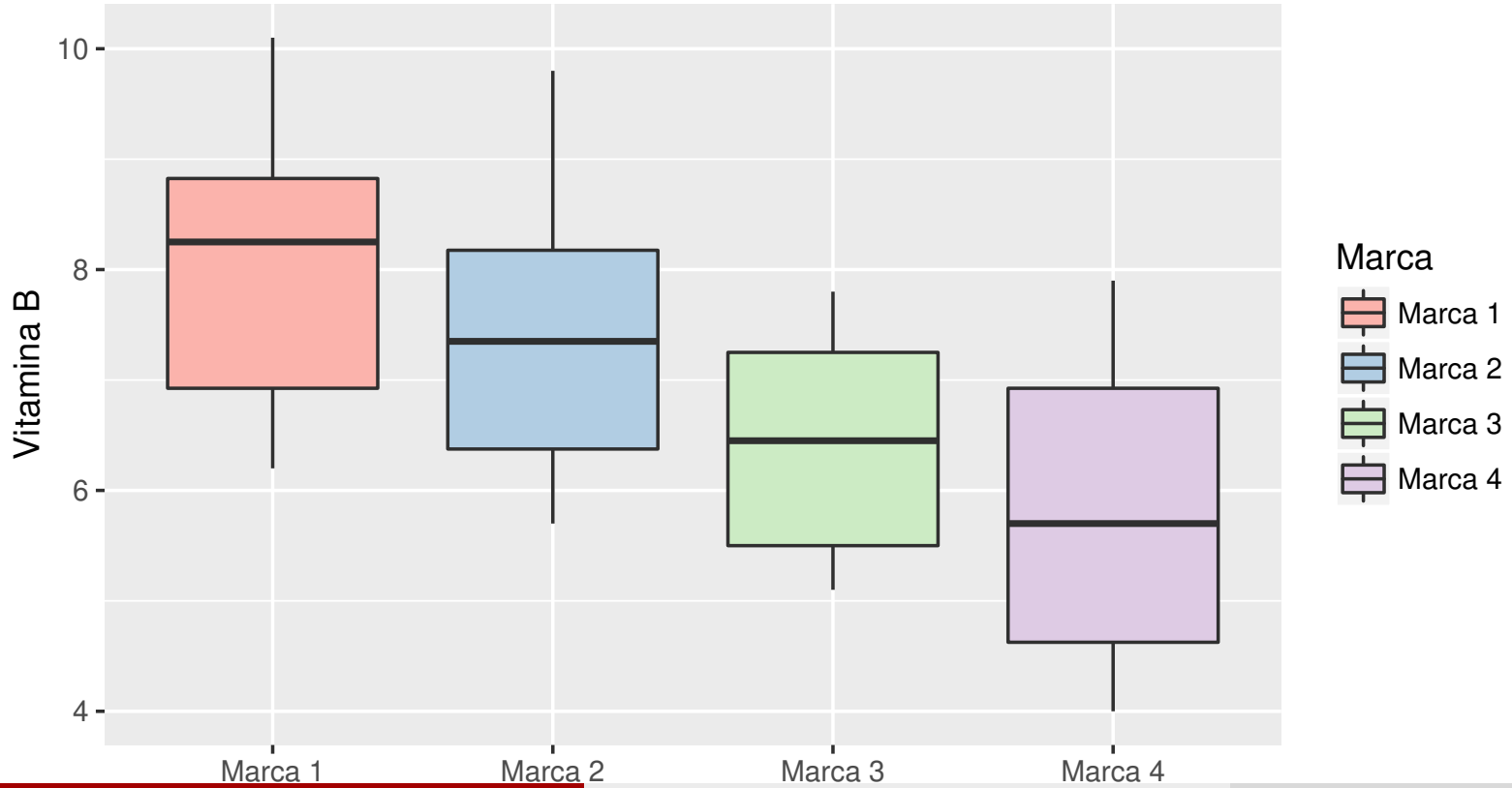
	Marca 1	Marca 2	Marca 3	Marca 4
	7.9	5.7	6.8	6.4
	6.2	7.5	7.8	7.1
	6.6	9.8	5.1	7.9
	8.6	6.1	7.4	4.5
	8.9	8.4	5.3	5.0
	10.1	7.2	6.1	4.0
	9.6			
Media	8.0500	7.4500	6.4167	5.8167
Desv.Std	1.4680	1.5083	1.1053	1.551

Tabla: Vitamina B en el té

Ejemplo Té

Inspección Visual

Analicemos gráficamente si se observan diferencias importantes entre los contenidos medios de vitamina B de las distintas marcas.



Comparación estadística de las Medias

Son estadísticamente significativas las diferencias que apreciamos visualmente?

Para generalizar lo antes expuesto, supongamos el siguiente modelo aplicado a la observación de k muestras normales independientes con varianzas iguales

$$\begin{array}{lll}
\text{Muestra } 1 & X_{11}, X_{12}, \cdots, X_{1n_1} \text{ v.a.i.i.d.} & X_{1j} \sim N(\mu_1, \sigma^2); \\
\vdots & \vdots & \vdots \\
\text{Muestra } i: & X_{i1}, X_{i2}, \cdots, X_{in_i} \text{ v.a.i.i.d.} & X_{ij} \sim N(\mu_i, \sigma^2); \\
\vdots & \vdots & \vdots \\
\text{Muestra } k: & X_{k1}, X_{k2}, \cdots, X_{kn_k} \text{ v.a.i.i.d.} & X_{kj} \sim N(\mu_k, \sigma^2);
\end{array}$$

donde v.a.i.i.d significa variables aleatorias independientes e idénticamente distribuidas o, equivalentemente, una muestra aleatoria.

El Modelo

Supuestos

$$X_{ij} = \mu_i + \varepsilon_{ij} \text{ para } 1 \leq i \leq k, 1 \leq j \leq n_i$$

siendo $\varepsilon_{ij} \sim N(0, \sigma^2)$ independientes.

Las variables aleatorias observadas son normales, independientes entre sí dentro de las muestras y entre las muestras y homocedásticas, lo que significa que sus varianzas son iguales. Este es un supuesto bastante fuerte que, en caso de no satisfacerse, se deberá realizar una transformación de los datos o aplicar técnicas no paramétricas.

Cuando las transformaciones disponibles no son efectivas para que los supuestos se satisfagan, veremos más adelante, una alternativa interesante conocida como el test de Kruskall-Wallis

Notación

Introducimos la siguiente notación: $\overline{X}_{i.}$ y S_i^2 para indicar respectivamente las variables media y varianza de la i-ésima muestra, con $1 \leq i \leq k$.
El estimador de σ^2 se puede obtener calculando un promedio ponderado de las varianzas de cada muestra s_i^2, lo que es una generalización de la idea de la varianza amalgamada o *pooleada*.
El mejor estimador insesgado de σ^2 bajo este modelo es

$$S_P^2 = \frac{SSW}{n-k} = \frac{(n_1-1)S_1^2 + \cdots + (n_k-1)S_k^2}{n_1+n_2+\cdots+n_k-k} = \frac{\sum_{i=1}^k (n_i-1)S_i^2}{n-k},$$

donde SSW (*sum squares within*), o suma de cuadrados dentro de los grupos y $n = \sum_{i=1}^{k} n_i$.

Las hipótesis a testear son

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1: & \mu_i \neq \mu_j \text{ para algún par } (i,j). \end{cases}$$

La media general de todas las observaciones se calcula como

$$\overline{X}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{X}_i = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}.$$

Descomposición de la suma de cuadrados totales

Suma de cuadrados totales (SST *total sum of squares*)

Se define la como la suma de los cuadrados de las diferencias a la media general de todas las observaciones,

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2$$

Esta suma se puede descomponer en la suma de cuadrados dentro de los grupos (SSW) y entre los grupos (SSB) como mostraremos a continuación

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [(X_{ij} - \bar{X}_{i.}) + (\bar{X}_{i.} - \bar{X}_{..})]^2.$$

Descomposición de la suma de cuadrados totales

Desarrollando el cuadrado del binomio

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i.} - \bar{X}_{..})^2 + 2\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..}).$$

Puede probarse que el último sumando es nulo; utilizando

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})(\bar{X}_{i.} - \bar{X}_{..}) = \sum_{i=1}^{k}(\bar{X}_{i.} - \bar{X}_{..})\left[\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})\right].$$

Descomposición de la SST

Podemos expresar a la SST como SSW y SSB. Simbólicamente,

$$SST = SSW + SSB.$$

Estadístico de Contraste

Es el cociente entre dos estimaciones de la varianza común de los grupos

La estimación del numerador considera la varianza entre grupos mientras que la del denominador considera la varianza dentro de los grupos. La distribución de este estadístico es F-Fisher Snedecor, por ser un cociente de variables aleatorias con distribución Chi cuadrado normalizadas por sus respectivos grados de libertad.

La suma de cuadrados entre los grupos SSB (*sum squares between*)

$$SSB = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{..})^2}{k-1}.$$

Estadístico del Test

Como cociente $\frac{SSB}{SSW}$

Su expresión es:

$$F = \frac{\frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X}_{..})^2}{k-1}}{S_P^2}$$

Para decidir si las medias son o no iguales en las distintas subpoblaciones, debemos aplicar un test F. Como las variables Chi cuadrado son positivas, la variable F asume solamente valores positivos también.

ANOVA

Secuencia del Procedimiento

- **Primer paso:** se establecen la hipótesis de nulidad y la alternativa

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \cdots = \mu_k, \\ H_1: & \exists (i,j) : \mu_i \neq \mu_j. \end{cases}$$

- **Segundo paso:** se calcula el estadístico F el cual tiene distribución $F_{k-1,n-k}$ cuyos grados de libertad se corresponden con los grados de libertad de los estimadores de la varianza del numerador y del denominador.
- **Tercer paso:** se decide con la siguiente regla si $F_{obs} > F_{k-1,n-k,\alpha}$ entonces se rechaza H_0 con un nivel de significación α.

¿Por qué se rechaza para valores grandes del estadístico? O equivalentemente, ¿por qué se trata de una prueba unilateral derecha?

La respuesta se basa en que estamos comparando dos estimadores de la misma varianza, en el numerador utilizamos las diferencias entre las medias de los grupos y la media general, mientras que en el denominador amalgamamos las varianzas estimadas para cada subgrupo.
El hecho de que el numerador sea mucho mayor que el denominador indica que las medias son muy distintas entre sí.
Suponiendo en primera instancia que se verifican los supuestos del modelo del análisis de la varianza para el ejemplo del té, armamos la base de datos y aplicamos el test F para decidir si existen diferencias entre las medias del contenido de vitamina B en las distintas marcas de té a nivel 0.05.

Tabla de ANOVA

Salida de ANOVA presencia de vitamina B en el té

	GL	Suma de cuad.	Media de cuad.	F	$Pr(>F)$ (p-value)
Marca	3	22.93	7.645	3.791	0.0256*
Residuos	21	42.35	2.016		

donde los códigos de significación son los siguientes

El test F rechaza la igualdad de medias a nivel 0.05. Ahora, antes de tomar una decisión, se debe estudiar si los supuestos del contraste se satisfacen con el objeto de ver si la conclusión es válida. Para ello se realiza el diagnóstico del modelo que será desarrollado luego.

Validez de la Salida de ANOVA

Sólo es válido si se satisfacen los supuestos

El test F es válido sólo si las observaciones son independientes, las muestras tienen distribución normal y las varianzas de los grupos son iguales. Al igual que con el test t, hay que observar los datos para detectar si existe alguna razón para pensar que este modelo es o no el adecuado.

Inspeccionemos la homocedasticidad gráficamente

Los diagramas de caja aparecen a diferentes alturas pero el tamaño de las cajas se ve muy similar y tampoco se detecta la presencia de *outliers*. Por estas razones, no hay motivos para sospechar, a partir del gráfico, que no se cumple el supuesto de homocedasticidad.

Inspección Analítica de la Homocedasticidad

Test de Bartlett

Las hipótesis a contrastar en el **test de Bartlett** son

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2, \\ H_1: & \exists (i,j): \sigma_i^2 \neq \sigma_j^2. \end{cases}$$

Salida de R del Test de Bartlett

```
Bartlett test of homogeneity of variances
data:  te$Vitamina.B and Marca
Bartlett's K-squared = 0.6168, df = 3, p-value = 0.8926
```

El test de Bartlett no rechaza la hipotesis de nulidad; es decir, no hay evidencia estadística significativa de que la varianza de alguno de los subgrupos difiera de las otras.

Inspección Analítica de la Homocedasticidad

El test de Bartlett es muy sensible a la falta de normalidad. Es decir que puede rechazar la homocedasticidad por no cumplirse el supuesto de normalidad en lugar de rechazarla por no cumplirse el supuesto de homocedasticidad.
La alternativa más robusta, es el test de Levenne.

Test de Levene

Esta prueba es un nuevo ANOVA sobre los valores absolutos de los residuos de las observaciones respecto de la mediana (media), de su grupo.

```
Levene's Test for Homogeneity of Variance (center = median)
       Df  F value  Pr(>F)
group   3   0.2949  0.8286
       21
```

Inspección Analítica de la Homocedasticidad

Conclusión

Como el p-valor de la prueba es 0.8286, no se rechaza la hipótesis de homocedasticidad. Esto significa que el test de Levene no rechaza la hipótesis nula de homocedasticidad, lo que brinda la misma conclusión que el test de Bartlett. Por lo tanto, podemos suponer que se cumple la hipótesis de homocedasticidad.

Hipótesis de Normalidad de los Residuos

Faltaría analizar el cumplimiento del supuesto de normalidad de la distribución de los residuos, que es equivalente a analizar el supuesto de normalidad de la distribución de la variable original.

Análisis de la Normalidad de los residuos

Dentro de las herramientas conocidas, se dispone de distintos tests de normalidad así como de un gráfico que compara los cuantiles empíricos con los esperados, en el caso de que el supuesto se verifica. Este gráfico se denomina **QQ-plot** o **gráfico de cuantil-cuantil**.

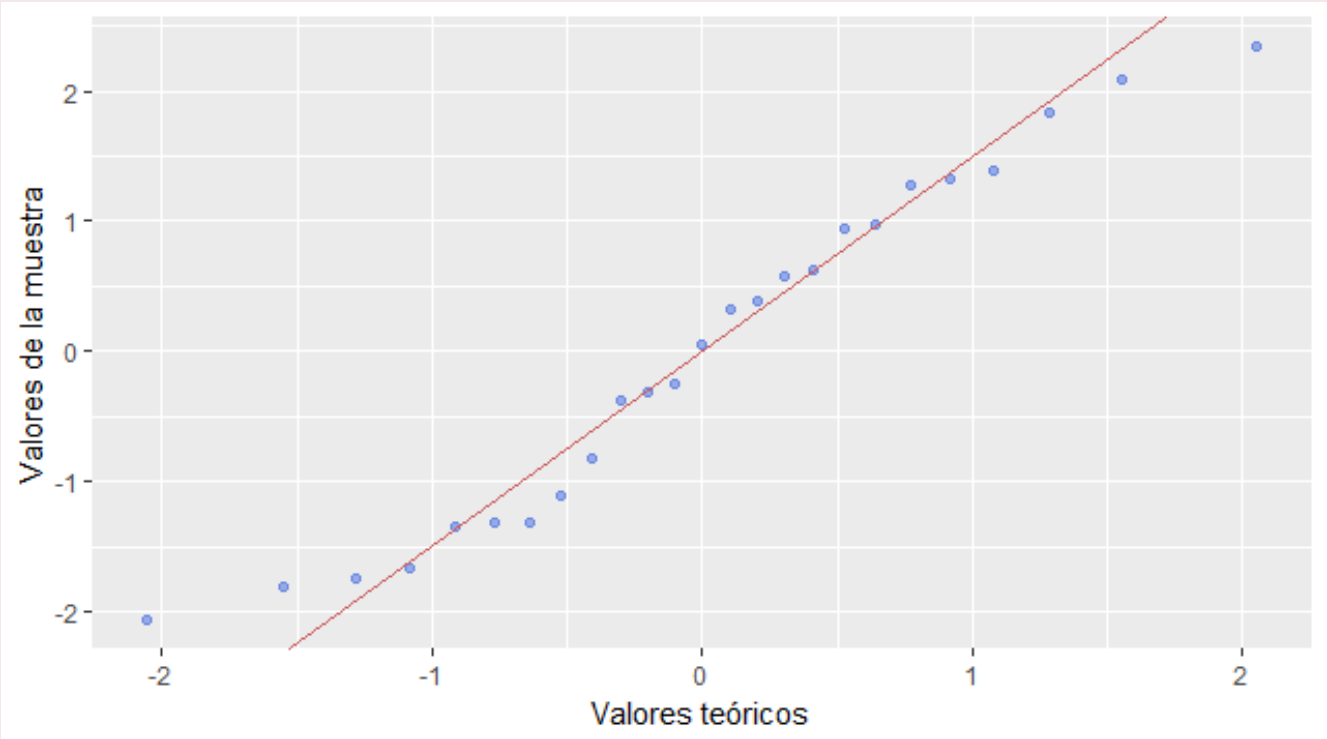
Salida de R de test de normalidad

```
Shapiro-Wilk normality test
data:  residuals(te.anova)
W = 0.95307, p-value = 0.2937

Anderson-Darling normality test
A = 0.36947, p-value = 0.3995

D'Agostino skewness test
skew = 0.065564, z = 0.160000, p-value = 0.8729
alternative hypothesis:  data have a skewness
```

Análisis de la Normalidad de los Residuos

Gráficos de cuantil-cuantil

Qué se debe observar en este gráfico?



Cumplimiento de Supuestos

Consideraciones Finales sobre el Diagnóstico

Si los datos fueran normales, los puntos que corresponden a las observaciones deberían posicionarse sobre la recta. Esto en la realidad no ocurrirá nunca, dado que se trata de una muestra aleatoria. Lo que debemos determinar es si el alejamiento observado de los puntos es significativo o no.

Conclusión

En base a las salidas de `R` ya estudiadas, no existe evidencia empírica en contra de la normalidad de la distribución de la variable o de los residuos. Luego, podemos dar por válido el rechazo de la hipótesis de nulidad del test F de análisis de la varianza. Por esto, concluimos que al menos una de las medias de los contenidos de vitamina B de las marcas de té es significativamente distinta de las demás.

Comparaciones a posteriori

Cuándo?

Si no se rechaza la hipótesis nula, el análisis finaliza en esa instancia. Si por el contrario se rechaza, resulta lógico que el experimentador no se conforme con esta respuesta, sino que desee comparar las medias de a pares en general y de algunas otras formas en casos específicos. En el ejemplo del té se encontró evidencia en contra de la igualdad de contenido medio de vitamina B en las distintas variedades de té consideradas.

Intervalo de confianza para la diferencia de dos medias

El propósito radica ahora en comparar las medias de dos grupos, digamos i e i^*. A estas comparaciones se las conoce como **comparaciones a posteriori** o **post-hoc**.

Comenzamos construyendo un intervalo de confianza para $\mu_i - \mu_{i^*}$. El estimador puntual a considerar es $\bar{X}_i - \bar{X}_{i^*}$.

Comparaciones Post-Hoc

IC para la diferencia de medias de poblaciones normales

En caso de normalidad, homocedasticidad e independencia de las variables, el IC de nivel $1-\alpha$ para la diferencia de medias es

$$\left[\bar{X}_i - \bar{X}_{i^*} - t_{n-k,1-\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_i}+\frac{1}{n_{i^*}}},\ \bar{X}_i - \bar{X}_{i^*} + t_{n-k,1-\frac{\alpha}{2}} S_P \sqrt{\frac{1}{n_i}+\frac{1}{n_{i^*}}}\right].$$

A partir de este intervalo, podemos deducir un test para estudiar las siguientes hipótesis

$$\begin{cases} H_0: & \mu_i = \mu_{i^*}, \\ H_1: & \mu_i \neq \mu_{i^*}. \end{cases}$$

Sin embargo..

... el problema radica en el nivel global de estos intervalos!

Comparaciones Post-Hoc

Nivel Global

Este intervalo tiene nivel $1 - \alpha$ para un par de medias pero deja de tener este nivel cuando se quiere comparar varios pares. Si se planean uno o muy pocos intervalos o tests, se puede usar intervalos de a pares ajustando el nivel de confianza de los mismos, aunque en caso contrario, conviene emplear un método para **intervalos de confianza simultáneos**.

Alternativas disponibles

Podemos citar entre muchas, las de Dunnet, Newman–Keuls, Tukey o LSD (*least signficative difference*). Todos estos procedimientos tienen un valor crítico que es luego comparado con las diferencias entre pares de los promedios muestrales. Cuando la diferencia supera este valor crítico significa que los valores medios de esos dos grupos son significativamente distintos.

Recomendaciones para la selección de la prueba a posteriori

Idea General

Básicamente, todas las pruebas son mejoras de la prueba original de t de Student. La prueba de Dunnet se emplea cuando se comparan todos los competidores contra uno original.

Algunas Consideraciones

- La prueba LSD debe emplearse sólo si se desean unas pocas comparaciones establecidas a priori antes de realizar el análisis de la varianza).
- Las pruebas de Scheffe y de Tukey están diseñadas para comparar todos los pares de medias.
- La prueba de Dunnet se recomienda cuando se comparan todos los tratamientos con un tratamiento base (tipo placebo).

Ejemplo: Dieta Conejos

Enunciado

Se llevó a cabo un estudio a fin de determinar el efecto de la fracción lipoproteica HDL-VHDL sobre lesiones ateroscleróticas en conejos. Se escogieron 24 conejos que fueron asignados aleatoriamente y, en forma balanceada, a una de las siguientes dietas aterogénicas que consisten de un conjunto de alteraciones con el fin de generar un depósito de lípidos en la pared de las arterias, que finalmente facilitará la pérdida de elasticidad arterial y otros trastornos vasculares.

Dietas

- **Dieta 1:** 60 días de dieta rica en colesterol 0.5%.
- **Dieta 2:** 90 días de dieta rica en colesterol 0.5%.
- **Dieta 3:** 90 días de dieta rica en colesterol 0.5% y luego 30 días con 50 mg de fracción lipoproteica HDL-VHDL por semana.

Ejemplo: Dieta Conejos

Las observaciones

Luego del experimento, los animales fueron sacrificados. En todos los casos se comprobaron lesiones aterogénicas en la arteria aorta. Se midió el contenido de colesterol en la aorta en mg/g obteniéndose los siguientes resultados.

Dieta 1	Dieta 2	Dieta 3
13.4	10.4	7.5
11.0	14.2	7.2
15.3	20.5	6.7
16.7	19.6	7.6
13.4	18.5	11.2
20.1	24.0	9.6
13.6	23.4	6.8

Ejemplo: Dieta Conejos

El Modelo

El modelo que vamos a aplicar es

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

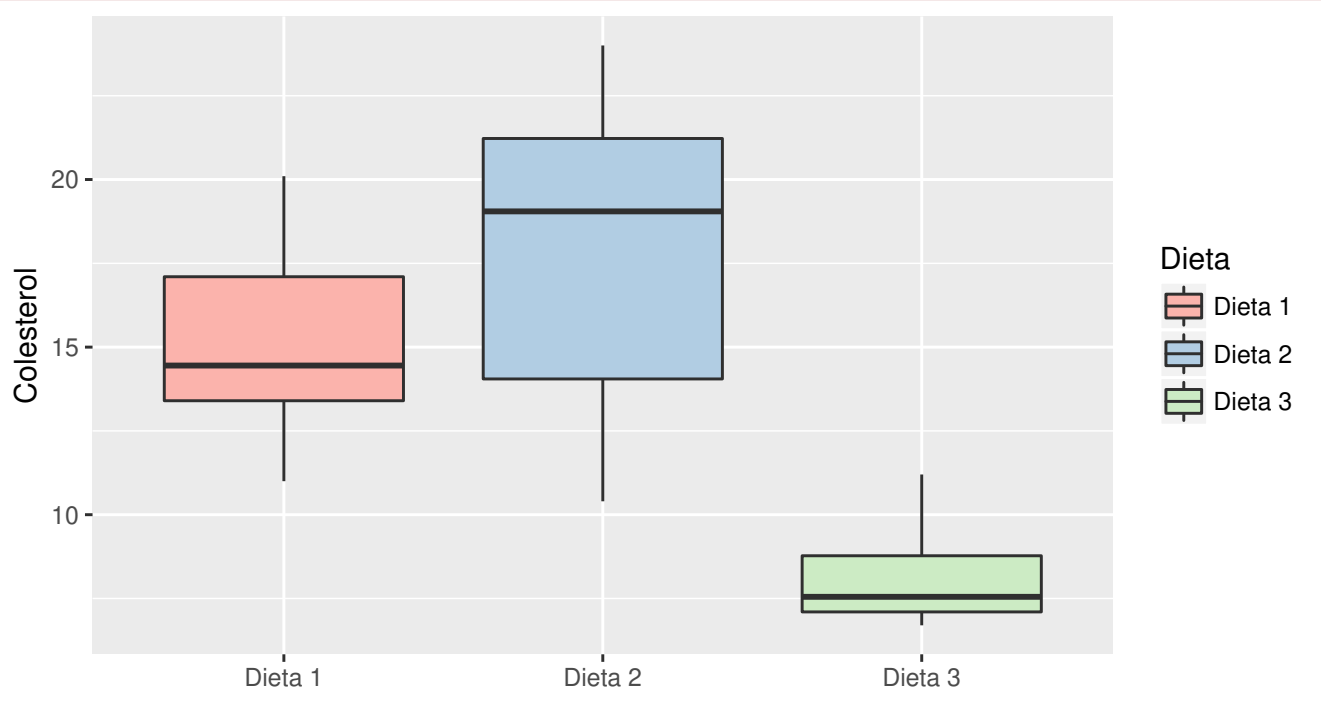
donde $\varepsilon_{ij} \sim N(0, \sigma^2)$ para $1 \leq i \leq 3$ y $1 \leq j \leq 8$. .

Resúmenes: media y el desvío del contenido de colesterol por dieta

Tratamiento	Media	Desviación típica	*n*
Dieta 1	15.225	2.9894	8
Dieta 2	18.025	4.8664	8
Dieta 3	8.138	1.5620	8
Totales	13.796	5.3608	24

Ejemplo: Dieta Conejos

Realizamos un *boxplot* para apreciar gráficamente si existen diferencias entre los contenidos medios de colesterol en la aorta de las dietas y, también ver si hay presencia de *outliers* en las distribuciones o asimetrías y si tiene sentido pensar que las varianzas son iguales.



Ejemplo: Dieta Conejos

Observaciones del Gráfico

En el boxplot se aprecia que las varianzas no parecen ser similares. No se observan *outliers* en ninguno de los diagramas de caja. Ahora, para comprobar si estas sospechas tienen significación estadística, vamos a ensayar la prueba de Levene.

El resumen es

```
            Df  Sum Sq  Mean Sq  F value    Pr(>F)
Dieta        2   415.6   207.78    17.78  3.03e-05      ***
Residuals   21   245.4    11.69
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘ ’ 1
```

Ejemplo: Dieta Conejos

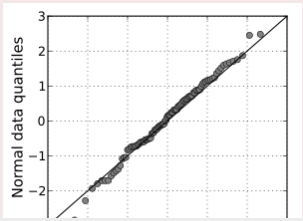
Análisis de la Normalidad de los Residuos

Esta salida indica que las diferencias entre los diámetros aórticos de los conejos sometidos a las distintas dietas son diferentes. Sin embargo, estos resultados sólo serán válidos si se satisfacen los supuestos del modelo de análisis de la varianza.

La prueba de Shapiro-Wilk produce la siguiente salida

```
Shapiro-Wilk normality test
data:  residuals(colesterol.anova)
W = 0.97939, p-value = 0.8843
```

Esta salida indica que puede sostenerse el supuesto de normalidad distribucional de los residuos.



Ejemplo: Dieta Conejos

La prueba de Levene produce la siguiente salida

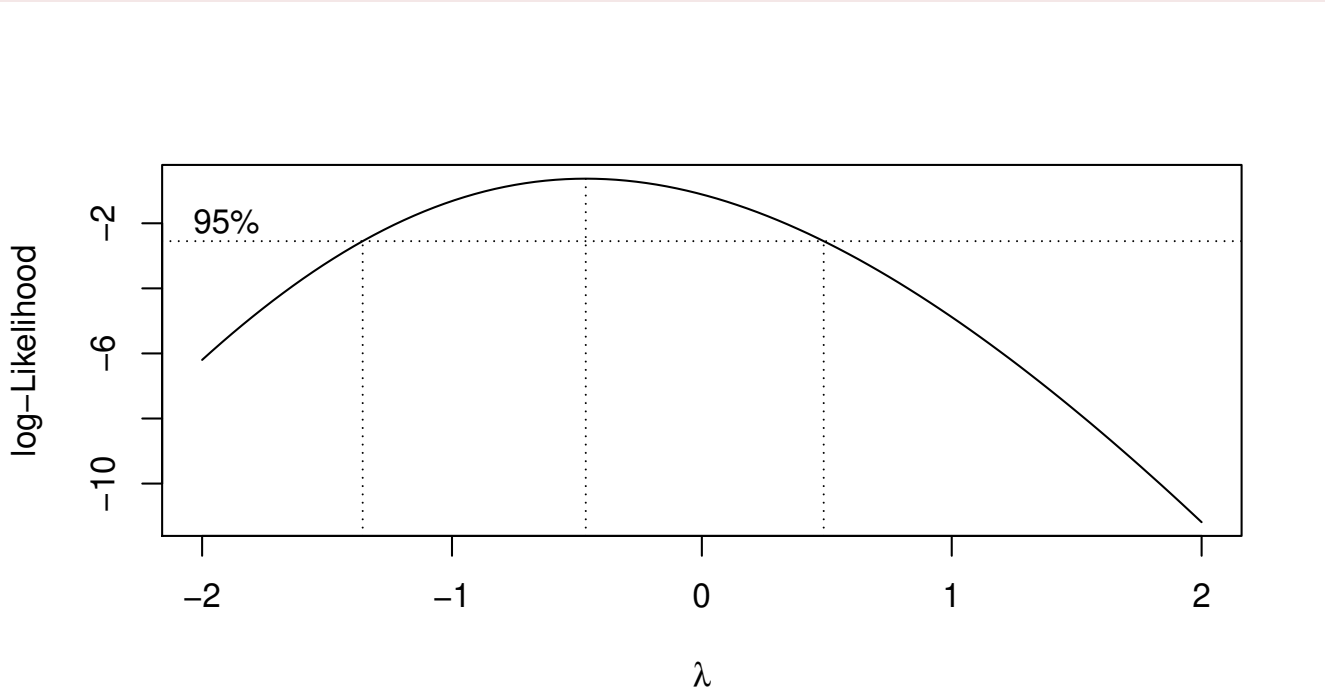
```
Levene's Test for Homogeneity of Variance (center = median)
       Df  F value   Pr(>F)
group   2    3.639  0.04396
       21
- - -
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 '
```

De esta última salida se interpreta que no es posible suponer homocedasticidad en la distribución de los residuos.

Ejemplo: Dieta Conejos

Transformación de la Variable Respuesta

Para decidir el exponente de la transformación aplicamos el test de Box & Cox. La salida de este test sugiere una transformación de la variable respuesta con un exponente cercano a -0.5.



Ejemplo Dieta Conejos

Datos Transformados

Realizamos la transformación sugerida y un nuevo análisis de la varianza, que origina nuevas salidas.

```
            Df   Sum Sq   Mean Sq  F value    Pr(>F)
Dieta        2  0.05841  0.029207    29.75  7.45e-07
Residuals   21  0.02062  0.000982
- - -
Signif.  codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’  0.1 ‘
```

En esta salida se aprecia que las diferencias siguen siendo significativas aún con los datos transformados. Falta realizar el análisis diagnóstico del modelo para los nuevos datos.

Ejemplo: Dieta Conejos

Supuesto de Normalidad

```
Shapiro-Wilk normality test
data:  residuals(tcolesterol.anova)
W = 0.97902, p-value = 0.8771
```

Se cumple el supuesto de normalidad requerido para el modelo de análisis de la varianza para los datos transformados.

También se cumple el supuesto de homocedasticidad, se testearon los residuos respecto de la mediana.

```
Levene's Test for Homogeneity of Variance
       Df  F value                  Pr(>F)
group   2   0.2626                  0.7715
       21
```

Ejemplo: Dieta Conejos

Nos preguntamos por último, cuáles son las dietas que difieren entre sí.

Realizamos comparaciones a posteriori con intervalos de Tukey.

Salida de R

```
Tukey multiple comparisons of means
95% family-wise confidence level
$Dieta
                    diff      lwr       upr     p adj
Dieta 2-Dieta 1  -0.0174  -0.0569    0.0220   0.516
Dieta 3-Dieta 1   0.0948   0.0553    0.1343   0.000
Dieta 3-Dieta 2   0.1122   0.0727    0.1517   0.000
```

Ejemplo: Dieta Conejos

Consideraciones Finales

De donde se puede apreciar que la Dieta 3 produce niveles de colesterol inferiores a los de las otras dos dietas. Más aún, los dos últimos intervalos no contienen al 0, lo cual indica que la Dieta 3 es diferente de las Dietas 1 y 2.

¿Qué sucede si no se verifican los supuestos del análisis de la varianza ni para los datos originales ni para los datos transformados?

La alternativa en este caso son las pruebas no paramétricas. Siendo las más usadas para ANOVA, la prueba de la mediana y la prueba de Kruskal-Wallis, también conocida como **análisis de la varianza no paramétrico**. De estas dos pruebas, la más potente resulta ser la de Kruskal-Wallis siendo una generalización del test de Wilcoxon de rangos signados.

Test de Kruskal-Wallis no paramétrico para muestras independientes

Idea General

Esta prueba contrasta la hipótesis nula que establece que las k muestras independientes proceden de la misma población y, en particular, todas ellas tienen la misma posición central. La misma se basa en los rangos de las observaciones y no requiere el cumplimiento del supuesto de normalidad ni del supuesto de homocedasticidad.

El modelo que supone este test consiste en

$$\begin{array}{lll} \text{Pob 1:} & Y_{11}, \cdots, Y_{1n_1} & \text{v.a.i.i.d. con escala al menos ordinal;} \\ \text{Pob 2:} & Y_{21}, \cdots, Y_{2n_2} & \text{v.a.i.i.d. con escala al menos ordinal;} \\ \vdots & \vdots & \vdots \\ \text{Pob } k\text{:} & Y_{k1}, \cdots, Y_{kn_k} & \text{v.a.i.i.d. con escala al menos ordinal.} \end{array}$$

Test de Kruskal Wallis

Supuestos

Las distribuciones de todas las subpoblaciones deben ser semejantes, de lo contrario, el rechazo de la hipótesis de nulidad implicaría que las distribuciones son distintas y no que sus medianas difieren, al igual que en la prueba de Wilcoxon-Mann-Whitney.

Las hipótesis a contrastar son

$$\begin{cases} H_0: & \theta_1 = \theta_2 = \cdots = \theta_k, \\ H_1: & \exists (i,j) : \theta_i \neq \theta_j. \end{cases}$$

Test de Kruskal Wallis

El estadístico de contraste

cuando hay pocos o ningún empates, es

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_{i.}^2}{n_i} - 3(N+1),$$

donde N es el total de observaciones y $R_{i.}$ es la suma de los rangos de la muestra i, dados por R_{ij}, el rango en la distribución conjunta de la observación j del grupo i.

La distribución del estadístico de contraste

Bajo H_0, este estadístico tiene distribución aproximada Chi cuadrado con $k-1$ grados de libertad.

Por este motivo, la regla de decisión resulta:

- se rechaza H_0 cuando $H_{obs} > \chi^2_{k-1,1-\alpha}$.
- no se rechaza H_0 cuando $H_{obs} < \chi^2_{k-1,1-\alpha}$.

Pasos del Procedimiento

- Se ordenan todas las observaciones en sentido creciente y se reemplazan por su rango R_{ij} $(i = 1, \cdots, k,\ j = 1, \cdots, n_i)$, en la muestra conjunta ordenada.
- En caso de empates, se asigna a cada una de las observaciones empatadas el rango promedio de ellas.
- Se suman los rangos de cada grupo de observaciones. La suma de los rangos en la muestra combinada del i-ésimo grupo se designa con $R_{i.}$ y el rango promedio del i-ésimo grupo se denota con $\bar{R}_{i.}$.
- Se calcula el estadístico de contraste H.
- Se toma una decisión y se concluye

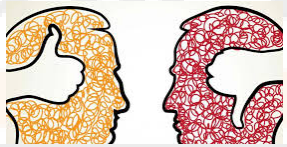
Ejemplo: Calificaciones

Aclaración

Cabe aclarar que decir que "las poblaciones tienen la misma posición central" es equivalente a decir que tienen "el mismo valor esperado o media aritmética de los rangos", o que "las poblaciones tienen igual mediana".

Los Datos

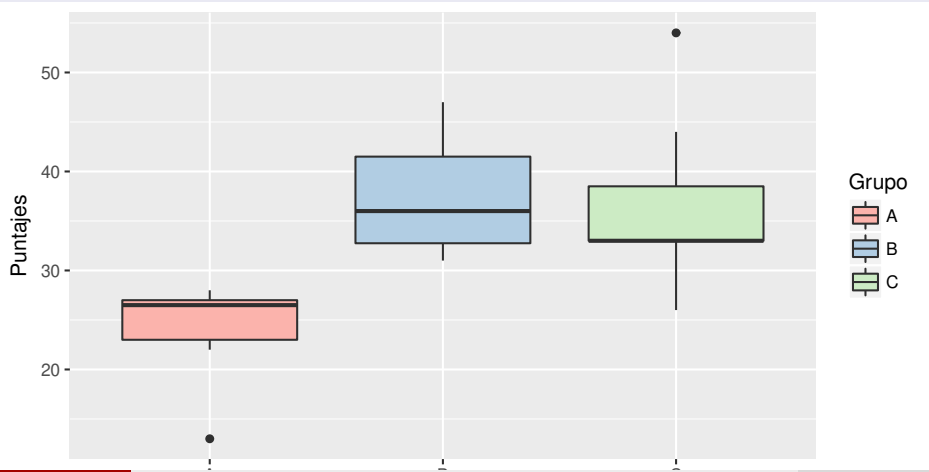
Se realizó una intervención educativa innovadora para mejorar el rendimiento de los estudiantes. Dentro de los grupos de clasificación, el A es el grupo de control y los restantes, B y C, son los grupos con distintas innovaciones. Se evaluó a los alumnos mediante una prueba objetivo sobre un total de 60 puntos.



Ejemplo: Calificaciones

Los Datos

Grupo	Puntajes						
A	13	27	26	22	28	27	
B	43	35	47	32	31	37	
C	33	33	33	26	44	33	54



Análisis de Normalidad

Datos	W	p-valor
Grupo Total	0.76163	0.02583

Estos Datos son esencialmente discretos

Además se rechaza la hipótesis de normalidad Por ende, aplicamos un análisis no paramétrico mediante el test de Kruskal-Wallis.

Las Hipótesis

$$\begin{cases} H_0: & \text{los tres grupos tienen la misma distribución de puntaje} \\ H_1: & \text{al menos un grupo tiene diferente distribución de puntaje} \end{cases}$$

Ejemplo: Calificaciones

Los datos ordenados los datos y sus rankeamientos.

X	13	22	26	27	27	28	31	32	35	37	
G	A	A	A	A	A	A	B	B	B	B	
Rgo	1	2	3.5	5.5	5.5	7	8	9	14	15	
X	43	47	26	33	33	33	33	44	54	33	44
G	B	B	C	C	C	C	C	C	C	C	C
Rgo	16	18	3.5	11.5	11.5	11.5	11.5	17	19	11.5	17

Rangos Totales por Grupo

Grupo	**A**	**B**	**C**
RangodelGrupo	**24.5**	**80**	**85.5**

Ejemplo: Calificaciones

Regla de Decisión

Rechazamos H_0 si $H_{obs} > \chi^2_{2,0.95} = 5.99$, siendo el estadístico de contraste de nuestra prueba

$$H_{obs} = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_{i.}^2}{n_i} - 3(N+1) = \frac{12}{19(19+1)} \left(\frac{24.5^2}{6} + \frac{80^2}{6} + \frac{85.5^2}{7} \right) - 3(19+1) = 9.92.$$

Por lo tanto, la decisión es rechazar H_0 debido a que $9.92 > 5.99$. Luego, no puede suponerse que la distribución de los rendimientos de las innovaciones sean iguales.

Sin embargo...

Con el análisis realizado hasta el momento, no se puede inferir acerca de la mediana ya que, como puede apreciarse en el boxplot comparativo, las distribuciones en los distintos grupos no son similares.

Ejemplo: Calificaciones

Test de Kruskal-Wallis de la suma de los rangos

Salidas del test y de las comparaciones múltiples

- Para la prueba de Kruskal-Wallis

```
Kruskal-Wallis rank sum test
Kruskal-Wallis chi-squared = 9.9265, df = 2, p-value =
0.00699
```

- Para la prueba de comparaciones múltiples

```
Multiple comparison test after Kruskal-Wallis
                p.value:  0.05
Comparisons  obs.dif  critical.dif  difference
A-B            9.250         7.777        TRUE
A-C            8.130         7.494        TRUE
B-C            1.119         7.494       FALSE
```

Ejemplo: Calificaciones

Conclusiones

A partir de esta última salida, surge que las diferencias de las distribuciones son estadísticamente significativas y que el grupo A difiere significativamente de los grupos B y C, mientras que los grupos B y C no difieren significativamente entre sí.

Organización

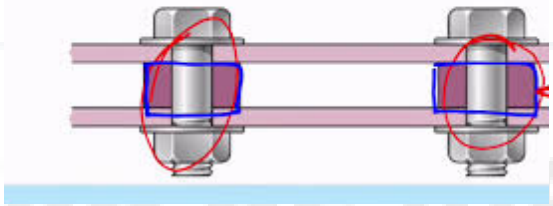
ANOVA dos vías

El análisis de varianza de dos vías, también conocido como plan factorial con dos factores, sirve para estudiar la relación entre una variable dependiente cuantitativa y dos variables independientes cualitativas (factores) cada una con varios niveles. El ANOVA de dos vías permite estudiar cómo influyen por si solos cada uno de los factores sobre la variable dependiente (modelo aditivo) así como la influencia de las combinaciones que se pueden dar entre ellas (modelo con interacción).;

El Problema

Ejemplo: Resistencia

Una empresa de materiales de construcción quiere estudiar la influencia que tienen el grosor y el tipo de templado sobre la resistencia máxima de unas láminas de acero. Para ello miden el estrés hasta la rotura (variable cuantitativa dependiente) para dos tipos de templado (lento y rápido) y tres grosores de lámina (8mm, 16mm y 24 mm).



Definición del Modelo

Este modelo tienen dos factores cruzados y se plantea inicialmente la interacción entre ambos que puede resultar significativa o no.

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + (\alpha\gamma)_{ij} + \varepsilon_{ijk}$$

con $1 \leq i \leq 2, 1 \leq j \leq 3, 1 \leq k \leq 5$

Y_{ijk} es la resistencia observada en la k-ésima lámina que recibió la combinación del nivel i-ésimo del factor templado con el nivel j-ésimo del factor grosor.

μ valor medio poblacional de la resistencia de las láminas.

Los efectos del Modelo

α_i es el efecto del i-ésimo nivel del factor templado sobre la variable resistencia de las láminas.

γ_j es el efecto del j-ésimo nivel del factor grosor sobre la variable resistencia de las láminas.

$(\alpha\gamma)_{ij}$ es el efecto de la interacción del i-ésimo nivel del factor templado con el j-ésimo nivel del factor grosor sobre la variable resistencia de las láminas.

ε_{ijk} es la contribución propia o error aleatorio de la k-ésima lámina que recibió la combinación del i-ésimo nivel del factor templado con el j-ésimo nivel del factor grosor sobre la resistencia de las láminas.

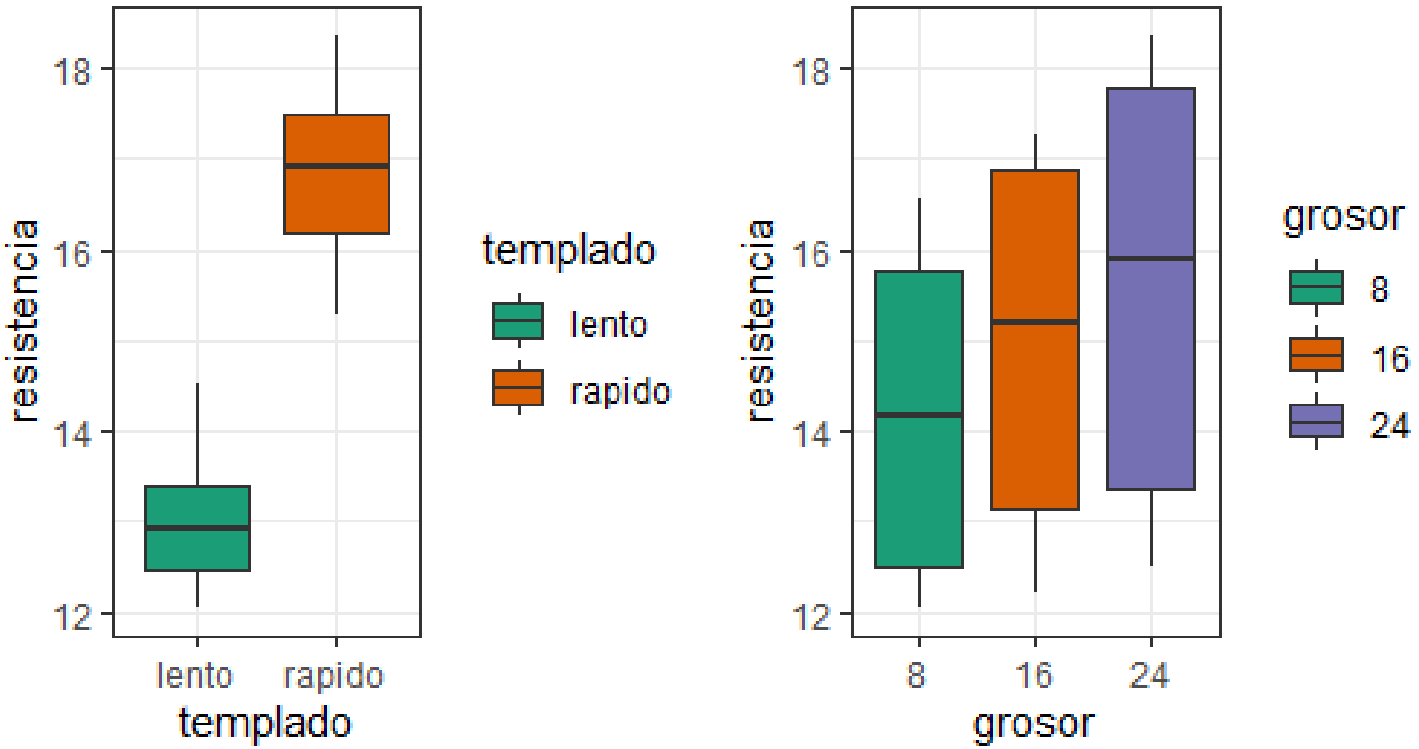
Los Datos

Ingresamos los Datos

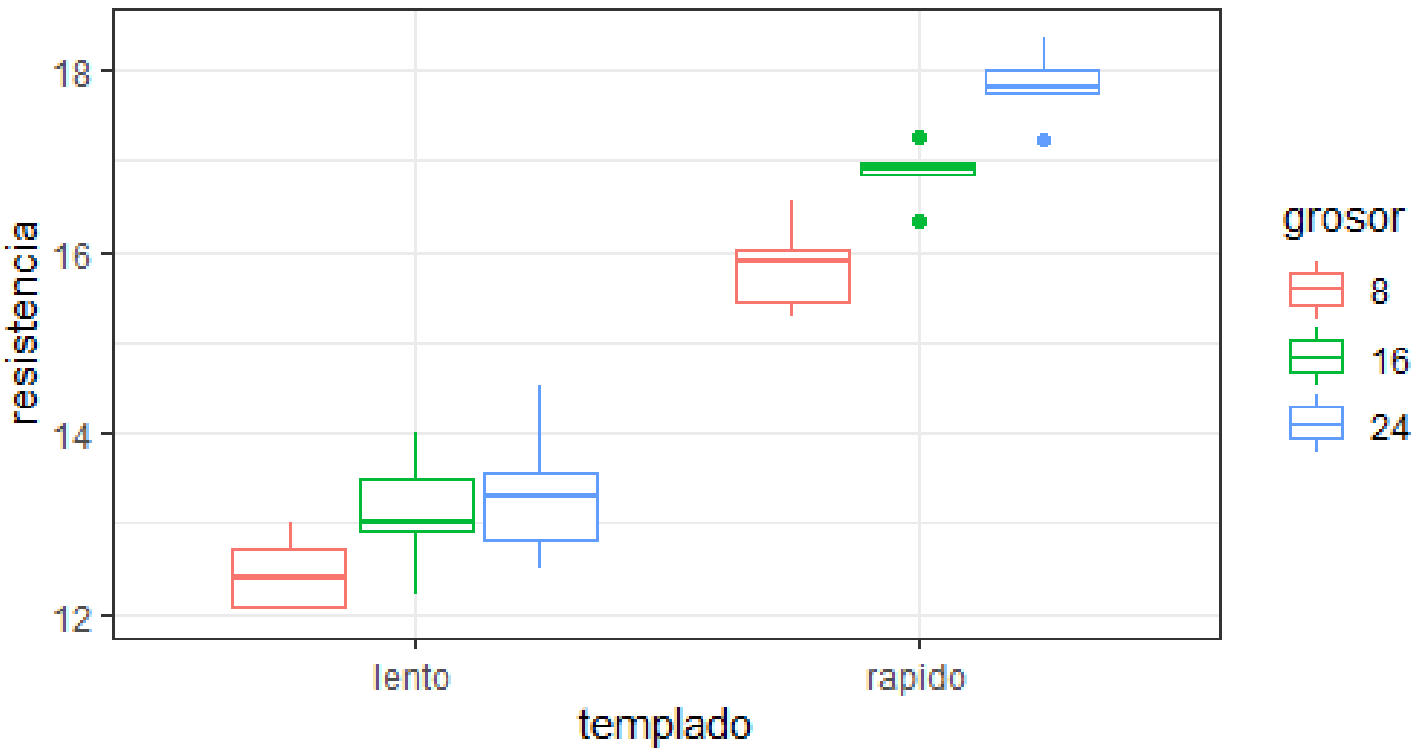
```
resistencia<- c(15.29, 15.89, 16.02, 16.56, 15.46, 16.91, 16.99,
17.27, 16.85, 16.35, 17.23, 17.81, 17.74, 18.02, 18.37, 12.07, 12.42,
12.73, 13.02, 12.05, 12.92, 13.01, 12.21, 13.49, 14.01, 13.30, 12.82,
12.49, 13.55, 14.53)
templado <- c(rep(c("rapido", "lento"), c(15,15)))
grosor<-rep(c(8, 16, 24), each = 5, times = 2)
datos<- data.frame(templado = templado, grosor =
as.factor(grosor), resistencia = resistencia)
head(datos)
```

Visualización de los Datos

En primer lugar se generan los diagramas "Boxplot" para identificar posibles diferencias significativas, asimetrías, valores atípicos y homogeneidad de varianza entre los distintos niveles.



Visualizando los efectos de ambos factores simultaneamente



El código de la visualización

```
p1 <- ggplot(data = datos, aes(x = templado, y =
resistencia,fill=templado)) + geom_boxplot() +
theme_bw()+scale_fill_brewer(palette="Dark2")
p2 <- ggplot(data = datos, aes(x = grosor, y = resistencia, fill =grosor))
+ geom_boxplot() + theme_bw()+scale_fill_brewer(palette="Dark2")
p3 <- ggplot(data = datos, aes(x = templado, y = resistencia, colour =
grosor)) + geom_boxplot() +
theme_bw()+scale_fill_brewer(palette="Dark2")
p3
grid.arrange(p1, p2, ncol = 2)
```

Resúmenes Básicos de la Variable objetivo por Factores

```
with(data = datos,expr = tapply(resist, templado, mean))
with(data = datos,expr = tapply(resist, templado, sd))
with(data = datos,expr = tapply(resist, grosor, mean))
with(data = datos,expr = tapply(resist, grosor, sd))
with(data = datos,expr = tapply(resist, list(templado,grosor), mean))
with(data = datos,expr = tapply(resist, list(templado,grosor), sd))
```

Medias	**X8**	**X16**	**X24**
lento	12.458	13.128	13.338
rapido	15.844	16.874	17.834

Desvíos	**X8**	**X16**	**X24**
lento	0.4208	0.6725	0.7834
rapido	0.5000	0.3342	0.4172

Observaciones

ⓐ A partir de la representación gráfica y el cálculo de las medias se puede intuir que existe una diferencia en la resistencia alcanzada dependiendo del tipo de templado.

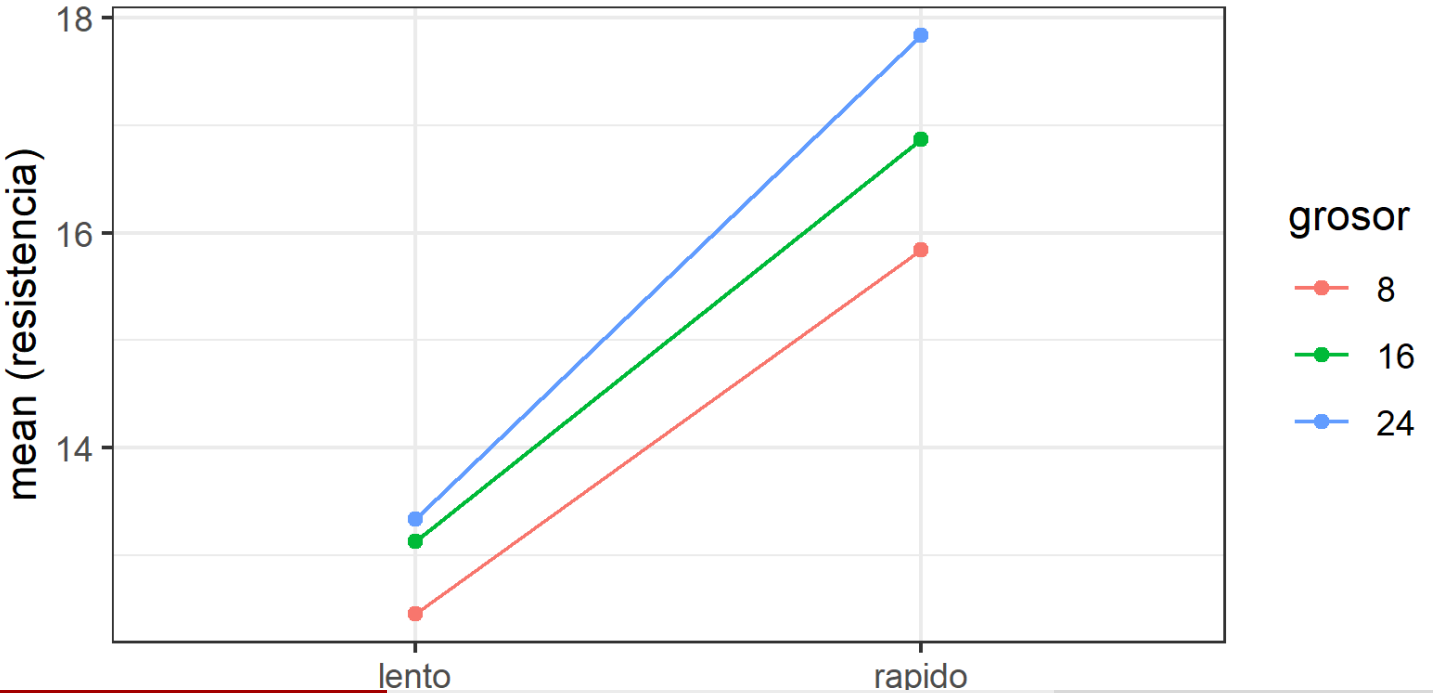
ⓑ La resistencia se incrementa a medida que aumenta el grosor de la lámina, si bien no es seguro que las medias sean significativamente distintas.

ⓒ La distribución de las observaciones de cada nivel parece simétrica sin presencia de valores atípicos. Falta testear el cumplimiento de los supuesto del modelo de ANOVA.

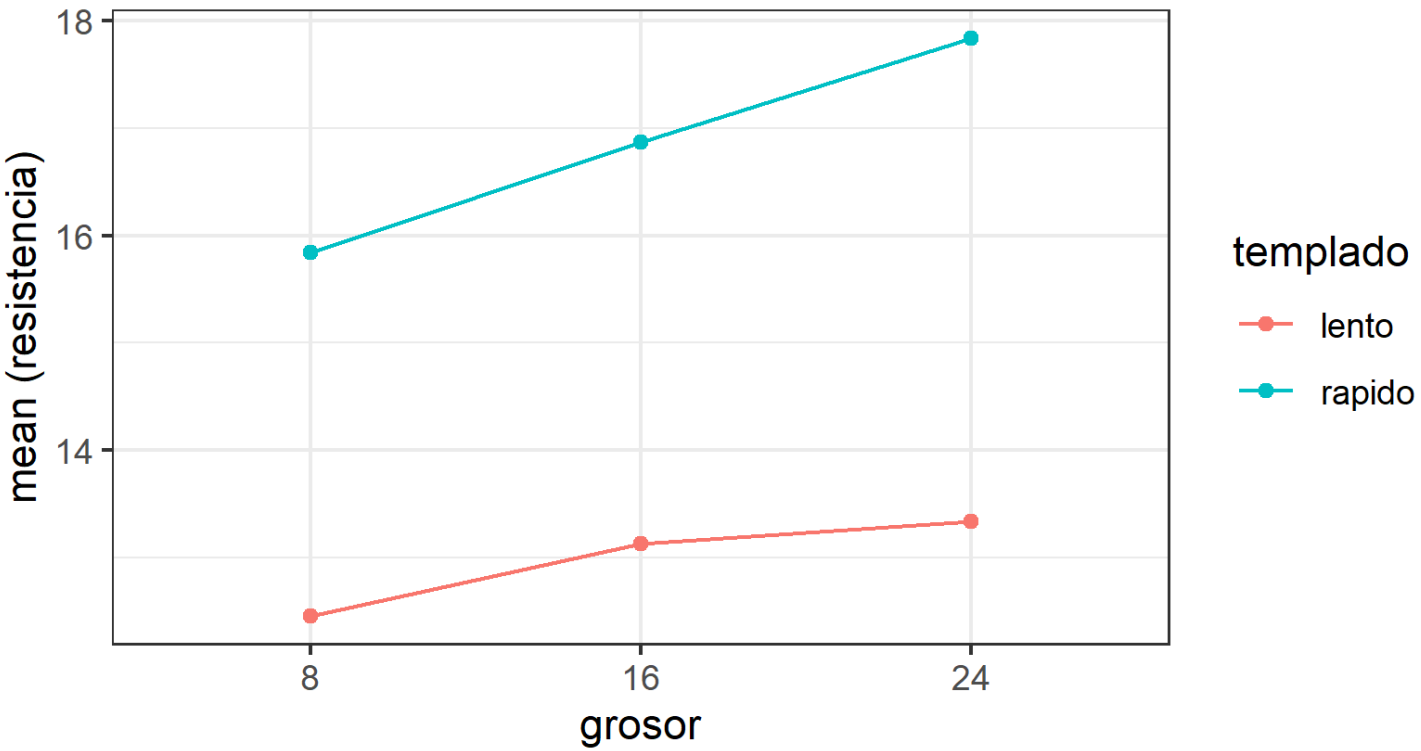
ⓓ Para identificar las interacciones una herramienta muy útil son los gráficos de interacción. Si en ellos las líneas son paralelas significa que el efecto de un factor es similar para los distintos niveles del otro factor y por lo tanto no hay interacción.

Gráficos de Interacción

```
ggplot(data = datos, aes(x = templado, y = resistencia, colour = grosor,
group = grosor)) +
stat_summary(fun = mean, geom = "point") +
stat_summary(fun = mean, geom = "line") +
labs(y = 'mean (resistencia)') + theme_bw()
```



```
ggplot(data = datos, aes(x = grosor, y = resistencia, colour = templado,
group = templado)) +
stat_summary(fun = mean, geom = "point") +
stat_summary(fun = mean, geom = "line") +
labs(y = 'mean (resistencia)') + theme_bw()
```



Estimamos los coeficientes del Modelo

anova $<-$ aov(resistencia $\sim$ templado * grosor, data $=$ datos)
summary(anova)

	Df	Sum Sq	Mean Sq	F	pvalue	
templado	1	112.68	112.68	380.082	3.19E-16	***
grosor	2	10.41	5.21	17.563	2.00E-05	***
templado:grosor	2	1.6	0.8	2.705	0.0873	.
Residuals	24	7.11	0.3			

Interpretación de la Salida

El análisis de varianza confirma que existe una influencia significativa sobre la resistencia de las láminas por parte de ambos factores (templado y grosor) con tamaños de efecto η^2 grande y mediano respectivamente, pero que no existe interacción significativa entre ellos.

Tamaño de los Efectos

La eta cuadrado parcial es una medida del tamaño del efecto en ANOVA (proporción de varianza explicada por las variables predictoras).
Suele considerarse que una eta cuadrada parcial en torno a 0,01 es poco efecto, que una eta cuadrada en torno a 0,06 indica un efecto medio y que una eta cuadrada superior a 0.14 es ya un efecto grande.
La principal ventaja de los índices (η^2 y R^2) es su fácil interpretación ya que se puede multiplicar por 100 y hablar en términos de porcentaje de varianza explicada por el efecto de la variable independiente. Por ejemplo, si el valor de $\eta^2 = 0.15$ entonces el 15% de las diferencias encontradas entre los dos grupos se atribuye al efecto de la intervención o tratamiento.

Eta Cuadrado

$$R^2 = \eta^2 = \frac{Suma\ de\ Cuadrados_{TRATAMIENTO\ o\ EFECTO}}{Suma\ de\ Cuadrados_{TOTAL}}$$

$$\eta_p^2 = \frac{Suma\ de\ Cuadrados_{TRATAMIENTO\ o\ EFECTO}}{Suma\ de\ Cuadrados_{TRATAMIENTO\ o\ EFECTO} + Suma\ de\ Cuadrdos_{ERROR}}$$

Interpretación de eta cuadrado parcial

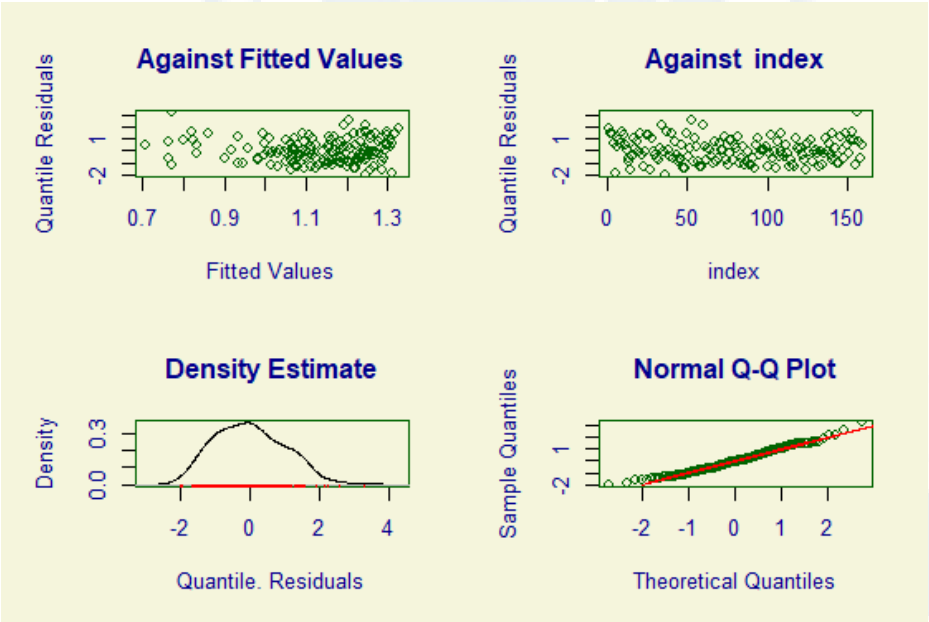
El estadístico de eta cuadrado parcial (η_p^2) es la proporción de varianza explicada por el efecto (efecto de A, efecto de B o efecto de interacción AB) más la del error que se puede atribuir a dicho efecto o fuente de varianza.

Salida de R

```
library(lsr)

etaSquared(anova)
```

	eta.sq	eta.sq.part
templado	0.8549	0.9406
grosor	0.0790	0.5941
templado:grosor	0.0122	0.1839

Diagnóstico



Observaciones y Conclusiones

Los residuos muestran la misma varianza para los distintos niveles (homocedasticidad).

Los residuos se distribuyen en forma normal.

No se aprecia estructura en el orden de los residuos.

Aparece observaciones con influencia dado sus valores observados de distancia de Cook.

Dada la no significancia de la interacción se puede plantear un modelo sin interacción, que siempre ofrece una alternativa más sencilla para interpretar.

$$Y_{ijk} = \mu + \alpha_i + \gamma_j + \varepsilon_{ijk}$$

con $1 \leq i \leq 2$, $1 \leq j \leq 3$ y $1 \leq k \leq 5$.

Modelo sin interacción

Para el modelo sin interacción estimamos los coeficientes

	Df	Sum Sq	Mean Sq	F	pvalue	
templado	1	112.68	112.68	336.02	$< 210^{-16}$	***
grosor	2	10.41	5.21	15.53	3.65E-05	***
Residuals	26	8.72	0.34			

Tamaño de los Efectos

	eta.sq	eta.sq.part
templado	0.8548	0.9282
grosor	0.0790	0.5443

Organización

Conceptos

El τ-esimo cuantil de Z es un número $Q(z)$ tal que satisface:

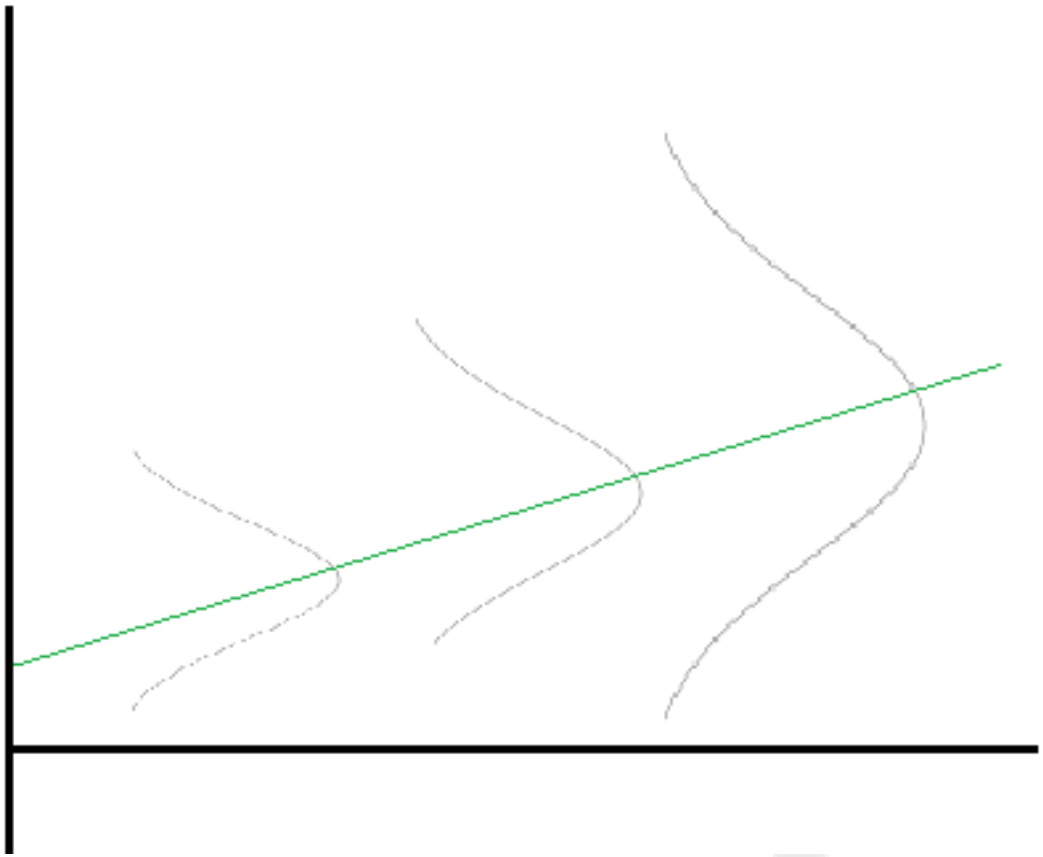
$$\tau = F(Q_z(\tau))$$

La idea de la regresión de cuantiles es que:

$$\frac{\partial Q_{y|x}(\tau)}{\partial x} = \beta(x)$$

Esto permite que el efecto de x resulte diferente para distintos valores de la variable x.

Varianza no constante



Necesidad de Regresión de Cuantiles

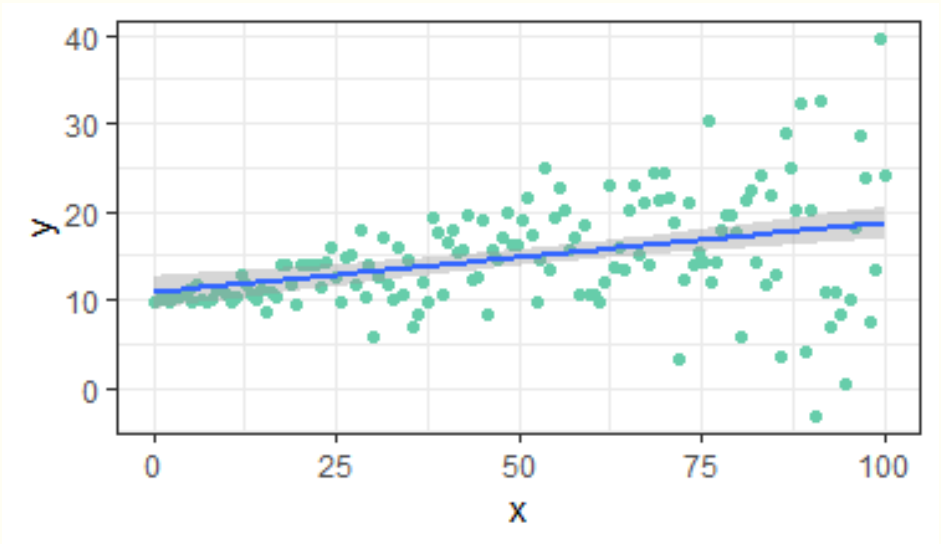
El método OLS se utiliza para estimar la media de una variable respuesta condicionada al valor de cierta cantidad de predictores.
Sin embargo, en algunas ocasiones es de mayor interés estimar algún cuantil de una distribución, o bien todos los cuantiles.
Los algoritmos que se utilizan son diferentes de los que se aplican en OLS pero la interpretación de los coeficientes es similar e interesa el efecto de un predictor sobre cada cuantil de la variable respuesta.

Estos modelos permiten evaluar la influencia de cada predictor sobre todo el rango de la variable respuesta y esto es muy apropiado en el caso de modelos de varianza no constante.

Ejemplo Simulado

```
# Generamos una base de datos con varianza no constante
set.seed(3) # fijamos una semilla para replicabilidad
x <- seq(0, 100, length.out = 150) # variable independiente
dst <- 0.2 + 0.1 * x # varianza no constante
b0 <-5 # intercepto teórico
b1 <-2 # pendiente teórica
err <- rnorm(150, mean = 0, sd = dst) # errores normales
y <-b0 + b1 * x + err # variable dependiente
base <- data.frame(x,y) # armamos la base de datos
p1<-ggplot(base, aes(x,y)) + geom_point(color='aquamarine3')+
geom_smooth(method='lm', level = 0.95)+theme_bw() # visual-
izamos
```

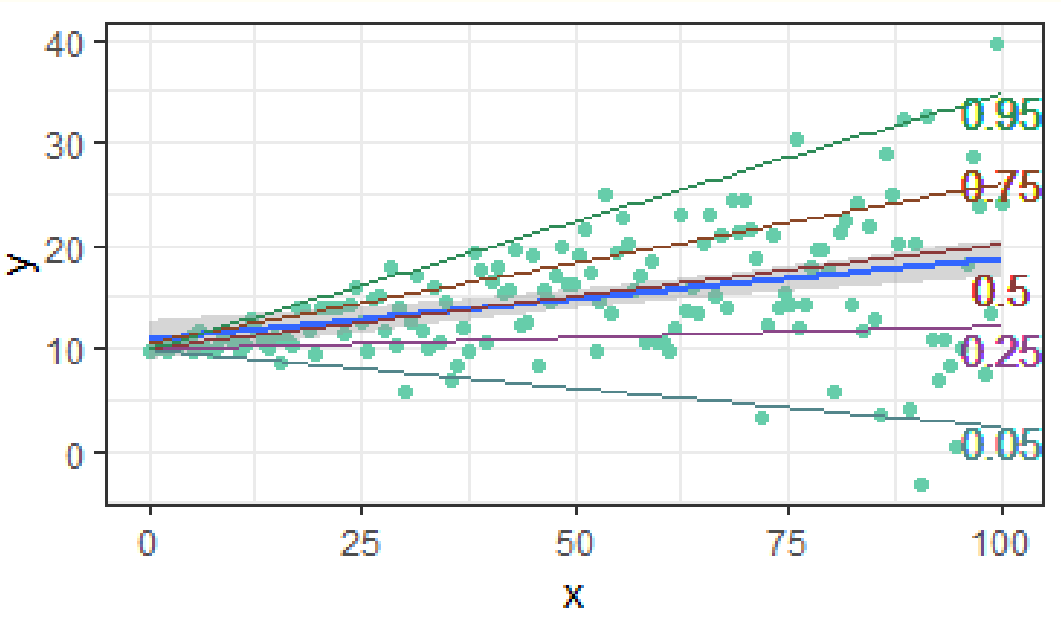
Ejemplo Simulado



Estimamos y visualizamos los cuantiles

```
p1+ geom_quantile(quantiles =0.05, color = 'cadetblue4')+
geom_text(aes(x=100,y=1,label="0.05"), color = 'cadetblue4')+
geom_quantile(quantiles =0.25, color = 'orchid4')+
geom_text(aes(x=100,y=10,label="0.25"),color='orchid4')+
geom_quantile(quantiles =0.50, color = 'indianred4')+
geom_text(aes(x=100,y=16,label="0.5"),color='indianred4')+
geom_quantile(quantiles =0.75, col = 'sienna4')+
geom_text(aes(x=100,y=26,label="0.75"),color='sienna4')+
geom_quantile(quantiles =0.95, color = 'seagreen4')+
geom_text(aes(x=100,y=33,label="0.95"),color='seagreen4')
```

Visualizamos las estimaciones de algunos cuantiles



Estimación de los coeficientes de la regresión

Si estamos interesados en la estimación de los coeficientes de la regresión:

```
library(quantreg)

qs <- seq(0.1, 0.9, by=0.1) # indicamos los cuantiles que nos interesan
qr2 <- rq(y ~ x, data=base, tau = qs)# estimamos estos cuantiles
coef(qr2)# pedimos los coeficientes de las rectas estimadas
```

Los coeficientes de las rectas estimadas

	(Intercept)	x
tau= 0.1	9.92	-0.03
tau= 0.2	9.95	0.01
tau= 0.3	9.97	0.03
tau= 0.4	10.07	0.06
tau= 0.5	10.01	0.10
tau= 0.6	10.02	0.13
tau= 0.7	10.53	0.15
tau= 0.8	10.42	0.17
tau= 0.9	10.20	0.22

Regresión de cuantiles para comparar medianas

Cómo se vincula con el test de Mann Whitney?

Cuando se quiere comparar dos poblaciones pero los datos no siguen la distribución normal, porque por ejemplo presentan colas asimétricas, y además el tamaño muestral es pequeño, suele ser adecuado comparar medianas en lugar de medias.

Con frecuencia se recurre al test de Mann–Whitney–Wilcoxon para contrastar la hipótesis nula de dos medianas son iguales. Este test sólo puede utilizarse con este fin si se cumple la exigente condición de que la única diferencia entre las poblaciones es su localización, el resto de características (asimetría, dispersión) tienen que ser idénticas.

La regresión de cuantiles permite comparar medianas (cuantil 0.5) sin necesidad de que se cumpla esta condición.

Ejemplo: Tiempos

Los Datos

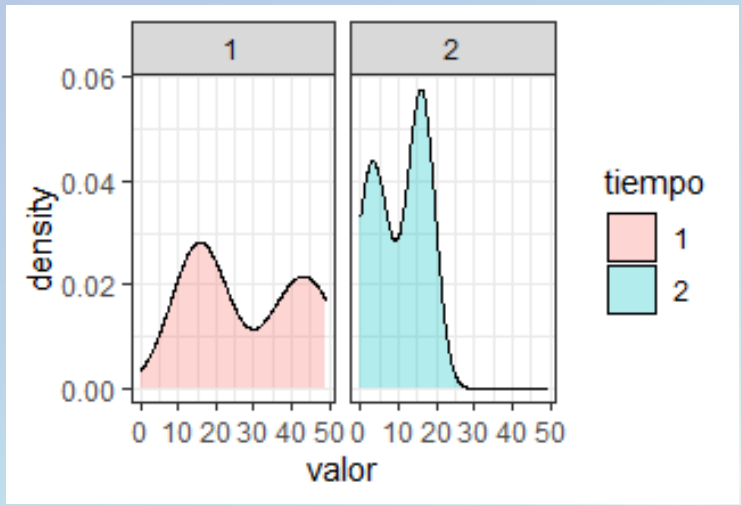
Supóngase que se dispone de dos grupos de estudiantes a los que se le asignó la misma tarea y se registró en cada caso el tiempo que tardaron en resolverla.
Queremos comparar los tiempos medianos de las muestras correspondientes a ambos grupos.

Ejemplo: Tiempos

```
# Cargamos los datos y visualizamos las dos distribuciones
tiempos1 <- c(15.4, 15.5, 15.6, 15.7, 15.8, 15.9, 16.0, 16.0, 38.3,
39.5, 40.0, 43.0, 45.0, 49.0, 48.0)
tiempos2 <- c(0,1.1,2.7, 3.2, 3.3, 5.4, 7.5, 16, 16.1, 16.1, 16.2,
16.3, 16.4, 16.5, 16.6)
todos <- data.frame(tiempos = c(rep("1", 15), rep("2", 15)),
valor = c(tiempos1, tiempos2))
p1=ggplot(todos,aes(x=valor,fill=tiempos))+
geom_density(alpha=0.3)+ facet_grid( . tiempos )+theme_bw()

p1
```

Ejemplo: Tiempos

Qué se observa?



El segundo grupo es claramente asimétrico por lo cual el t-test no resultaría adecuado. Como las formas distribucionales son diferentes tampoco es apropiado el test de Mann–Whitney–Wilcoxon.

Ejemplo: Tiempos

Cuál es el recurso apropiado en este caso?

Si se emplea de Mann–Whitney–Wilcoxon, debido a la diferencia en forma de las distribuciones, el resultado es significativo, lo que llevaría a conclusiones erróneas. Además en este caso las medianas son iguales!!

Conviene entonces recurrir a la regresión de cuantiles para realizar la comparación.

```
# Calculamos la mediana de cada grupo y luego aplicamos la prueba de Mann Whitney-Wilcoxon
median(tiempos1)
median(tiempos2)
wilcox.test(tiempos1,tiempos2)
```

```
median(tiempos1) [1] 16
median(tiempos2) [1] 16
wilcox.test(tiempos1,tiempos2)
Wilcoxon rank sum test with continuity

data: tiempos1 and tiempos2
W = 162, p-value = 0.042
alternative hypothesis: true location shift is not equal to 0
```

Ejemplo: Tiempos

```
# Ajustamos un modelo de regresión para la mediana con la predictora categórica del grupo
mod_q50 <- rq(valor ~ tiempos, tau = 0.5, data = todos)
summary(mod_q50, se = 'boot')
```

```
Call: rq(formula = valor ~ tiempos, tau = 0.5, data = todos)
tau: [1] 0.5
Coefficients:
             Value    Std. Error  t value  Pr(>|t|)
(Intercept) 16.00000 11.14023     1.43624  0.16201
tiempos2     0.00000 12.18324     0.00000  1.00000
```

Ejemplo: Tiempos

El resultado no muestra evidencias en contra de la hipótesis nula de que las medianas de ambos grupos son iguales.
También puede observarse que la mediana de ambos grupos es la misma.