BIOMETRÍA II
CLASE 1
EL MODELO LINEAL

Adriana Pérez
Depto de Ecología, Genética y Evolución
FECN, UBA

Efecto de la exposición postnatal a etanol sobre el volumen del cerebro en ratones

- La exposición intrauterina al etanol causa alteraciones cognitivas y conductuales persistentes
- Se desea estudiar los efectos neuroestructurales asociados a esta exposición en ratones
- 18 ratones de 7 días (equivalente al 3er trimestre de gestación en humanos) fueron divididos al azar en 3 grupos de igual tamaño. A cada grupo se le aplicó uno de los siguientes tratamientos: a) Solución salina, b) Etanol 1 g/kg, c) Etanol 2 g/kg. A los 82 días se determinó el volumen cerebral por resonancia magnética (en cm^3)
    - Unidad experimental
    - Variable respuesta VR (Y)
    - Variable explicativa VE (X)
    - Réplicas

Modelo?

Modelos

- Simplificaciones de la realidad
- Todos los modelos son incorrectos…
- …pero algunos modelos son más útiles que otros
- El modelo correcto no puede ser conocido con exactitud
- Cuanto más simple sea un modelo (menos parámetros), mejor (Principio de parsimonia)

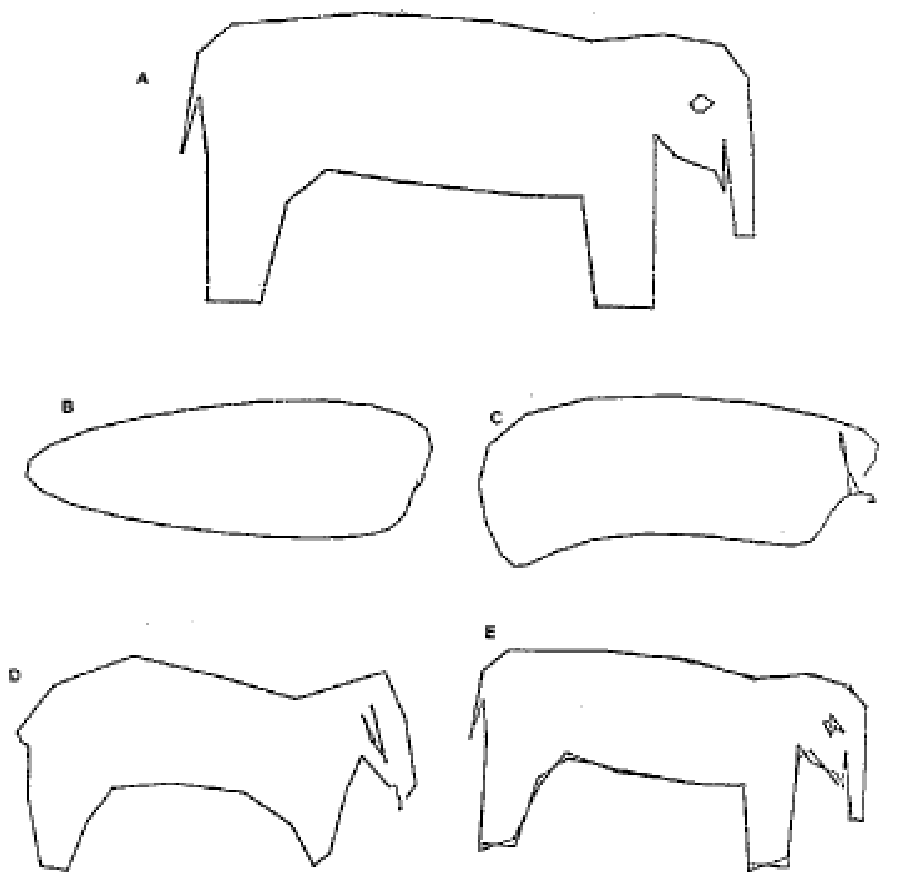


FIGURE 1.2. "How many parameters does does it take to fit an elephant?" was answered by Wel (1975). He started with an idealized drawing (A) defined by 36 points and used least squares Fourier sine series fits of the form $x(t) = \alpha_0 + \sum \alpha_i \sin(it\pi/36)$ and $y(t) = \beta_0 + \sum \beta_i \sin(it\pi/36)$ for $i = 1, \ldots, N$. He examined fits for $K = 5, 10, 20,$ and 30 (shown in B–E) and stopped with the fit of a 30 term model. He concluded that the 30-term model "may not satisfy the third-grade art teacher, but would carry most chemical engineers into preliminary design."

Burnham, K. P., & Anderson, D. (2003). Model selection and multi-model inference. *A Pratical informatio-theoric approch. Sringer*.

Modelo estadístico

Es una expresión matemática que indica cómo una variable aleatoria (VR, Y), con una distribución de probabilidades dada, se relaciona con una o más variables predictoras o explicativas (VE, X) consideradas en el diseño experimental

Y = función (X o Xs)

$$Y \sim X$$

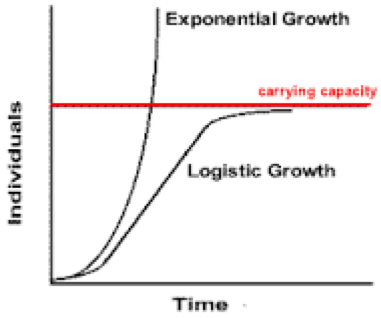
Modelos lineales

- Modelos lineales en los parámetros! La linealidad se refiere a los parámetros, no a la X. Los parámetros aparecen sumando; ningún parámetro aparece como exponente o multiplicado o dividido por otro parámetro.
- La VR es una combinación lineal de las VE

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 {X_i}^2 + \varepsilon_i$$

- En un modelo no lineal: los parámetros aparecen en la ecuación en forma no-lineal

$$Y_i = \beta_0 e^{\beta_1 X_i} + \varepsilon_i$$

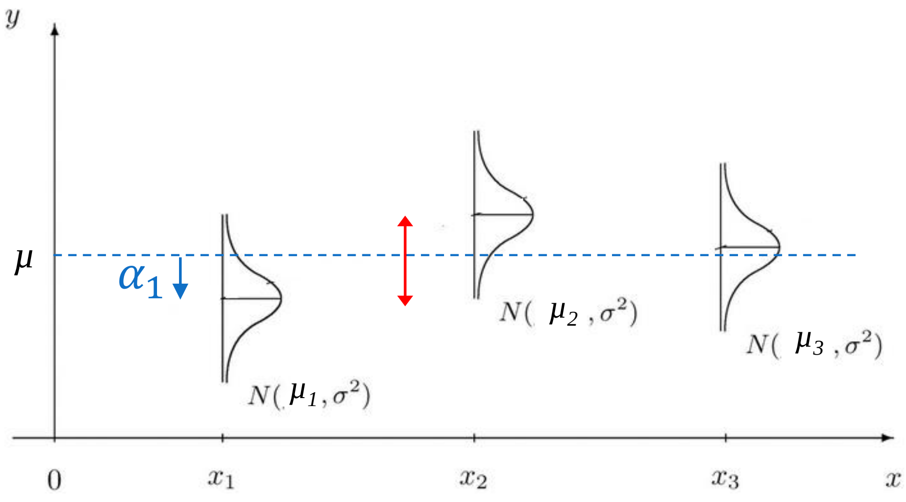


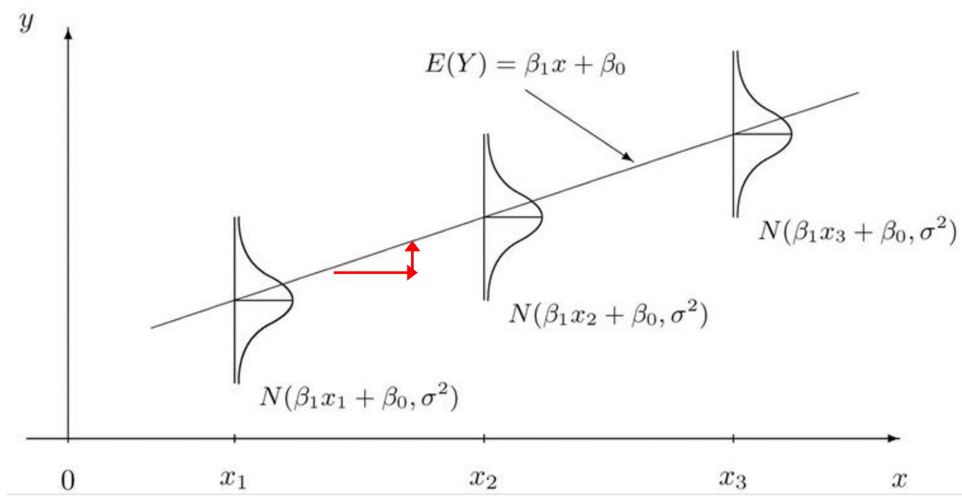
Parametrización del modelo:

X cuali (modelo de comparación de medias) o X cuanti (modelo de regresión)?

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

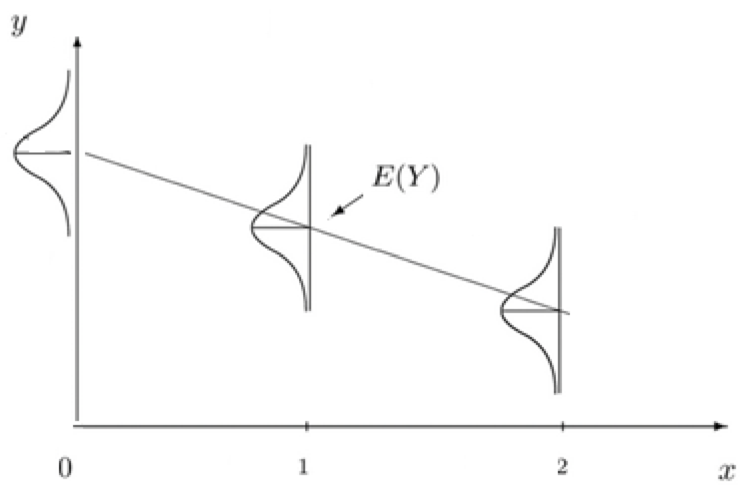
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$





- ✓ En la comparación de medias ("Anova") las VE se denominan **factores** y se las trata como cualitativas. La magnitud del efecto se mide como diferencia de medias
- ✓ En Regresión las VE son cuantitativas. La magnitud del efecto se mide mediante pendientes o coeficientes de regresión. Las VE cualitativas pueden ser incluidas previo transformación en variables indicadoras o dummy

Regresión lineal
Parametrización del modelo



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \varepsilon_i \sim Normal(0, \sigma^2)$$

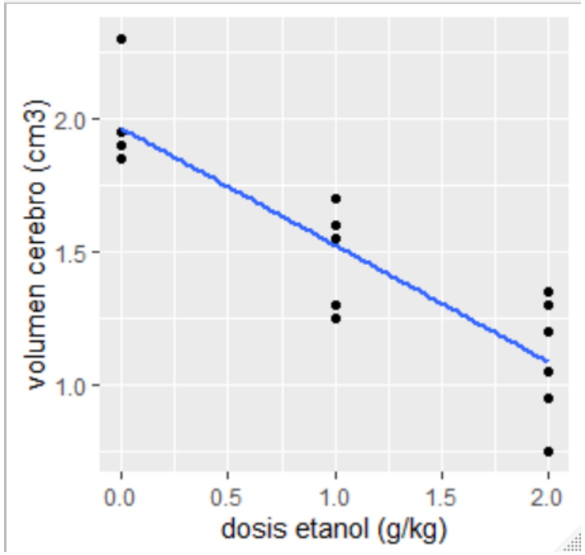
$$E(Y_i) = \beta_0 + \beta_1 X_i \qquad Y_i \sim Normal(\mu_i, \sigma^2)$$

Dado un valor de X, la esperanza de Y queda determinada unívocamente (componente sistemático).

Existe variación aleatoria (error) que responde a una distribución de probabilidades (componente aleatorio)

Modelo con 3 parámetros

- β_0 es el valor esperado de Y cuando X vale 0
- β_1 es el cambio esperado en Y por cada aumento unitario en X
- σ^2 es la varianza de Y para cada valor de X, común a todos



	etanol	vol.mean	vol.sd
1	0	1.975	0.1635543
2	1	1.500	0.1816590
3	2	1.100	0.2280351

Ecuación estimada?
Interpretación de intercepto y pendiente?

$$E(Y_i) = \beta_0 + \beta_1 \text{etanol}_i$$

```
m1<-lm(vol~etanol, bd)
summary(m1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.96250    0.06999  28.038 4.96e-15 ***
etanol      -0.43750    0.05422  -8.069 4.96e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1878 on 16 degrees of freedom
Multiple R-squared:  0.8028,	Adjusted R-squared:  0.7904 
F-statistic: 65.12 on 1 and 16 DF,  p-value: 4.956e-07
```

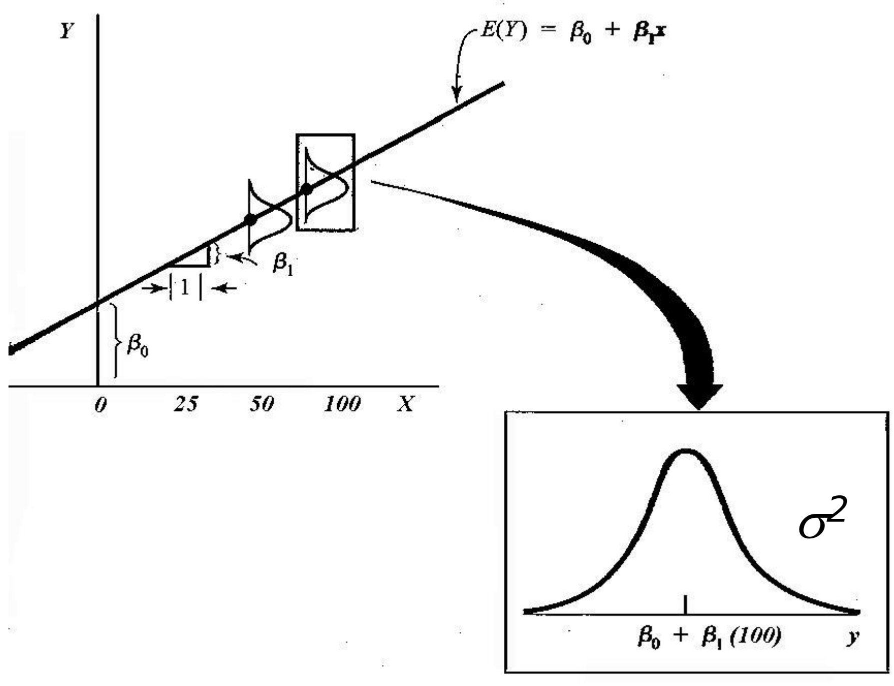
S_e estimador de σ

ratones_etanol.csv

Supuestos del modelo

No es necesarios para estimar los parámetros pero sí para que la inferencia sea válida

- Para cada valor de X existe una subpoblación de Y
  - La media de cada una de estas subpoblaciones es $E_{Y/X} = \beta_0 + \beta_1 X_i$ (linealidad)
  - La distribución de cada subpoblación es normal $Y_{i/X} \approx NID\ (\mu_{Y/X}, \sigma^2)$
  - las varianzas de las subpoblaciones son iguales, es decir que el modelo asume una varianza constante σ^2, sin importar el nivel de X Var [Y/X] = σ^2



Inferencia sobre los coeficientes de regresión

Ho: $\beta_1 = 0$ la variación de *Y* no se explica linealmente por la variación de *X*
H_1: $\beta_1 \neq 0$ la variación de *Y* sí se explica linealmente por la variación de *X*

Dos opciones (equivalentes)

A) Test *t* para β_i (en `summary`)
B) Anova (en `anova(modelo)`)

Test *t* para β_i

Se basa en la distribución del estimador $\hat{\beta}_1$

Si la distribución de $Y_{/X}$ es normal, $\hat{\beta}_1$ sigue una distribución normal, con esperanza

β_1 y EE = $\sqrt{\dfrac{\sigma^2}{\Sigma(x_i - \bar{x})^2}}$

EE: **error estándar** (de un estimador). Es una medida de la precisión en la estimación del parámetro

Se demuestra que $\hat{\beta}_1$ sigue una distribución aproximadamente normal cuando n es grande (extension del Teorema Central del Límite)

$$t_{n-k-1} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{S_e^2}{\Sigma(x_i - \bar{x})^2}}}$$

$$t_{n-k-1} = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{S_e^2}{\Sigma(x_i - \bar{x})^2}}}$$

k= cantidad de VE

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.96250    0.06999  28.038 4.96e-15 ***
etanol      -0.43750    0.05422  -8.069 4.96e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Anova

Se basa en descomponer la variabilidad de la VR en sus distintas fuentes:

- Variabilidad explicada por la/s VE
- Variabilidad no explicada o aleatoria (error /residual)

El estadístico es F (cociente de varianzas) y su distribución es F de Fisher (GL numerador, GL denominador)

Prueba t y anova son equivalentes ($t^2 = F$)

Anova

Ho: $\beta_i = 0$
H1: $\beta_i \neq 0$

Varianzas

variación controlada, impuesta por el investigador

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios	F
Explicada por las VE	$\sum(\hat{y}_i - \bar{y}_i)^2$	k	$\frac{SCexpl}{GLexpl}$	$\frac{CMexpl}{CMerror}$
No explicada por las VE, aleatoria o error	$\sum(y_i - \hat{y}_i)^2$	$n\text{-}k\text{-}1$	$\frac{SCerror}{GLerror}$	
Total	$\sum(y_i - \bar{y})^2$	$n\text{-}1$		

Variación aleatoria o no controlada
Estima σ^2

k= cantidad de VE

anova(m1)

```
Analysis of Variance Table

Response: vol
          Df  Sum Sq Mean Sq F value    Pr(>F)    
etanol     1 2.29688 2.29688  65.116 4.956e-07 ***
Residuals 16 0.56437 0.03527                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

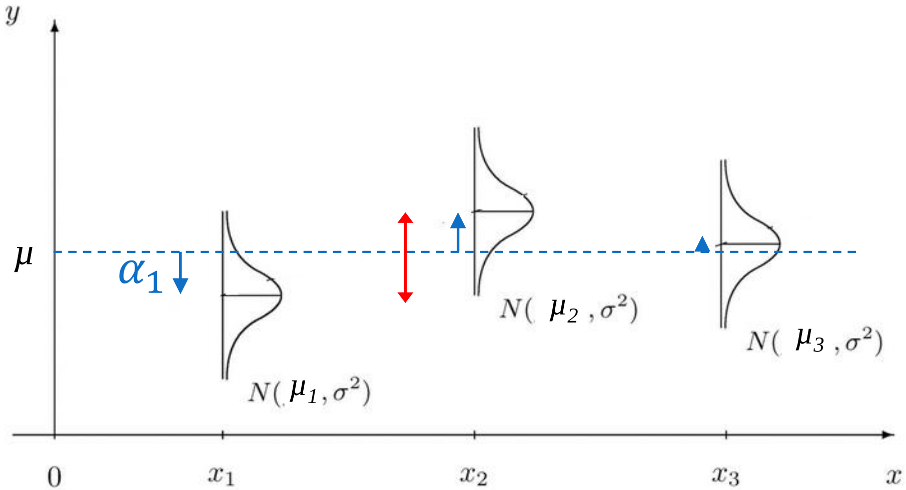
Efecto de la exposición postnatal a etanol sobre el volumen del cerebro en ratones

- 18 ratones de 7 días fueron divididos al azar en 3 grupos de igual tamaño. A cada grupo se le aplicó uno de los siguientes tratamientos: a) Solución salina, b) Etanol 1 g/kg, c) vino tinto con la misma concentración de etanol. A los 82 días se determinó el volumen cerebral por resonancia magnética (en cm^3)

  - Unidad experimental
  - Variable respuesta VR (Y)
  - Variable explicativa VE (X)
  - Réplicas

Modelo?

Modelo de comparación de medias
Parametrización



$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim Normal(0, \sigma^2)$$

$$E(Y_i) = \mu + \alpha_i \qquad Y_{ij} \sim Normal(\mu_i, \sigma^2)$$

Dado un nivel de X, la esperanza de Y queda determinada unívocamente (componente sistemático).

Existe variación aleatoria (error) que responde a una distribución de probabilidades (componente aleatorio)

Modelo con 4 parámetros:

- α_i es el efecto del nivel *i-ésimo* del factor sobre la esperanza de Y. Alternativamente puede pensarse en μ_i, la esperanza de Y para cada nivel del factor
- σ^2 es la varianza de Y para cada nivel del factor, común a todos

Los supuestos de este modelo son exactamente los mismos que para el modelo de regresión, salvo que aquí no aplica el supuesto de linealidad

Inferencia sobre los efectos α_i

Ho: $\alpha_i = o$ no existe efecto del factor // las medias poblacionales de los grupos son iguales
H_1: Al menos un $\alpha_i \neq o$ existe efecto del factor // al menos un grupo difiere en su media poblacional

No hay una prueba *t* equivalente

Una opción: Anova (en `anova(modelo)`)

La variación en la VR se particiona en:

- variación explicada por la VE (factor/tratamientos)
- variación no explicada o error

Y luego aplicar un método de comparaciones

ANOVA

Ho: Todos los $\alpha_i = 0$
H1: Algún $\alpha_i \neq 0$

Varianzas

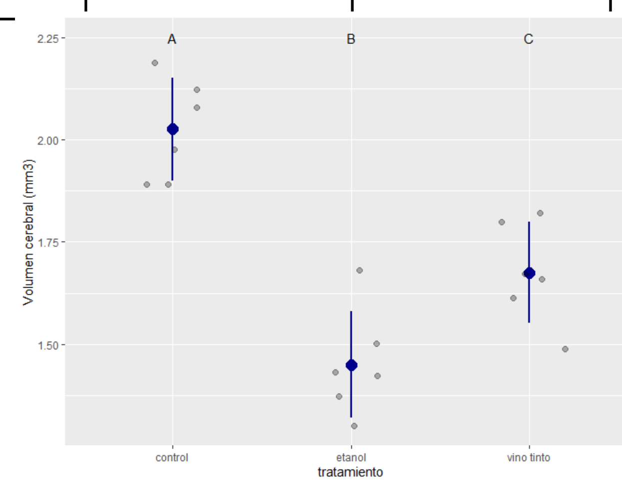
variación controlada, impuesta por el investigador

Fuente de Variación	Suma de Cuadrados	Grados de libertad	Cuadrados medios	F
Entre Tratamientos /grupos	$\Sigma n_i(\bar{y}_i - \bar{\bar{y}})^2$	$t-1$	$\frac{SCtrat}{GLtrat}$	$\frac{CMtrat}{CMerror}$
Dentro de tratamientos o error	$\Sigma(y_{ij} - \bar{y}_i)^2$	$(n_i-1)t = n-t$	$\frac{SCerror}{GLerror}$	
Total	$\Sigma(y_{ij} - \bar{\bar{y}})^2$	$n-1$		

Variación aleatoria o no controlada
Estima σ^2

```
m2<-lm(vol~tratamiento, bd)
anova(m2)
Analysis of Variance Table

Response: vol
            Df Sum Sq Mean Sq F value   Pr(>F)
tratamiento  2 1.0075 0.50375  31.656 4.14e-06 ***
Residuals   15 0.2387 0.01591
---
```

S^2_e estimador de σ^2



Pero, podemos analizar estos datos con un modelo de regresión?

- Las regresiones solo admiten VE cuantitativas
- Las v. cualitativas deben ser codificadas numéricamente para poder ser incluidas en la regresión (v. auxiliares, indicadoras o dummy)

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$
$$E(Y_i) = \beta_0 + \beta_1 etanol_i + \beta_2 vino\ tinto_i$$

Variables auxiliares, indicadoras o "Dummy"

Tratamiento	volumen
etanol	2.4
etanol	3.3
etanol	2.4
etanol	1.4
etanol	2.6
vino tinto	3.6
vino tinto	1.1
vino tinto	3.5
vino tinto	3.6
vino tinto	3.4
control	2
control	1.3
control	4.6
control	1.7
control	2.2

VE modelada cualitativa

Una de las variables auxiliares no aporta información novedosa ya que puede deducirse a partir de las otras dos (nivel de referencia)

Cuando el tratamiento es el control

$$E(Y_i) = \beta_0 = \mu_{control}$$

Cuando el tratamiento es etanol

$$E(Y_i) = \beta_0 + \beta_1 = \mu_{etanol}$$

Cuando el tratamiento es vino tinto

$$E(Y_i) = \beta_0 + \beta_2 = \mu_{vino\ tinto}$$

Modelo de 4 parámetros:

- β_0 es el valor esperado del nivel de referencia (control)
- β_1 es la diferencia de medias entre el tratamiento con etanol y el control
- β_2 es la diferencia de medias entre el tratamiento con vino tinto y el control (control)
- σ^2 es la varianza de Y para tratamiento, constante

```
tratamiento vol.mean    vol.sd vol.
     control    2.025 0.1246996
      etanol    1.450 0.1308434
 vino tinto    1.675 0.1227599
```

Es el resumen de un modelo de regresión, donde las VE cuali son convertidas en v.indicadoras

$$E(Y_i) = \beta_0 + \beta_1 dosis1_i + \beta_2 dosis2_i$$

```
m2<-lm(vol~tratamiento, bd)
summary(m2)
```

Magnitud del efecto (diferencia de medias con el grupo de referencia)

EE para la diferencia de medias

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            2.02500    0.05150  39.321  < 2e-16 ***
tratamientoetanol     -0.57500    0.07283  -7.895 1.01e-06 ***
tratamientovino tinto -0.35000    0.07283  -4.806 0.000231 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1261 on 15 degrees of freedom
Multiple R-squared:  0.8085,    Adjusted R-squared:  0.7829
F-statistic: 31.66 on 2 and 15 DF,  p-value: 4.14e-06
```

S_e estimador de σ

- Salvo cuando hay solo dos niveles, no hay una prueba “global” sobre el efecto de la VE
- Los coeficientes son diferencias de medias con respecto al nivel de referencia; no se informan otras comparaciones
- No son comparaciones ortogonales
- No controlan el error global

Inferencia sobre la diferencia de medias

```
m2<-lm(vol~tratamiento, bd)
anova(m2)
Analysis of Variance Table

Response: vol
            Df Sum Sq Mean Sq F value    Pr(>F)    
tratamiento  2 1.0075 0.50375  31.656 4.14e-06 ***
Residuals   15 0.2387 0.01591                      
---
```

Prueba global: Al menos una media poblacional difiere significativamente del resto

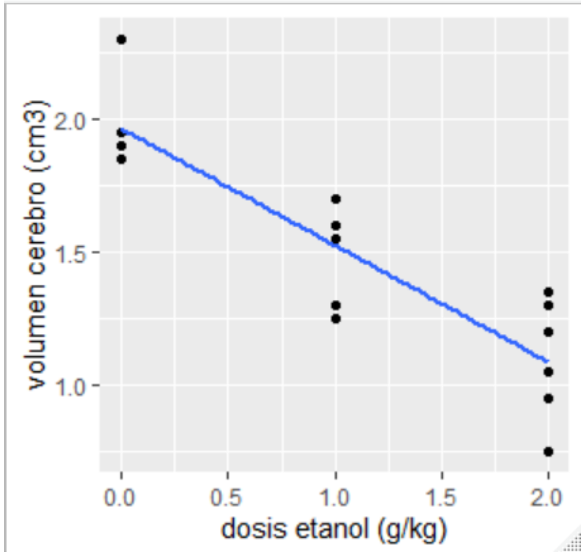
```
emmeans(m2, pairwise ~ tratamiento)

$emmeans
 tratamiento emmean     SE df lower.CL upper.CL
 control       2.02 0.0515 15     1.92     2.13
 etanol        1.45 0.0515 15     1.34     1.56
 vino tinto    1.68 0.0515 15     1.57     1.78

Confidence level used: 0.95

$contrasts
 contrast              estimate     SE df t.ratio p.value
 control - etanol         0.575 0.0728 15   7.895  <.0001
 control - vino tinto     0.350 0.0728 15   4.806  0.0006
 etanol - vino tinto     -0.225 0.0728 15  -3.089  0.0193

P value adjustment: tukey method for comparing a family of 3 estimates
```

Magnitud del efecto (diferencia entre medias)

Comparaciones de Tukey

Volviendo al ejemplo de regresión (VE cuantitativa)

- Y si la relación no fuese lineal?
- Y si considerase poco prudente ajustar una regresión lineal con solo 3 niveles de X?
- Y si me interesase la diferencia de medias entre dosis (tratamientos)?

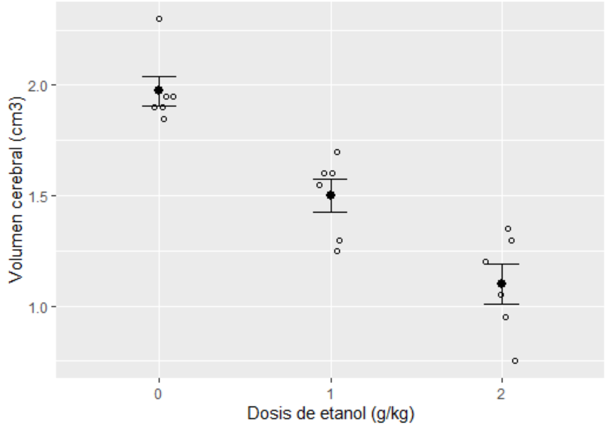


Es decir, si se desea incluir a una VE cuantitativa como cualitativa:

Modelo de comparación de medias

```
m3<-lm(vol~factor(etanol), bd)
```

Convierte a la variable en cualitativa



Inferencia sobre los efectos α_i

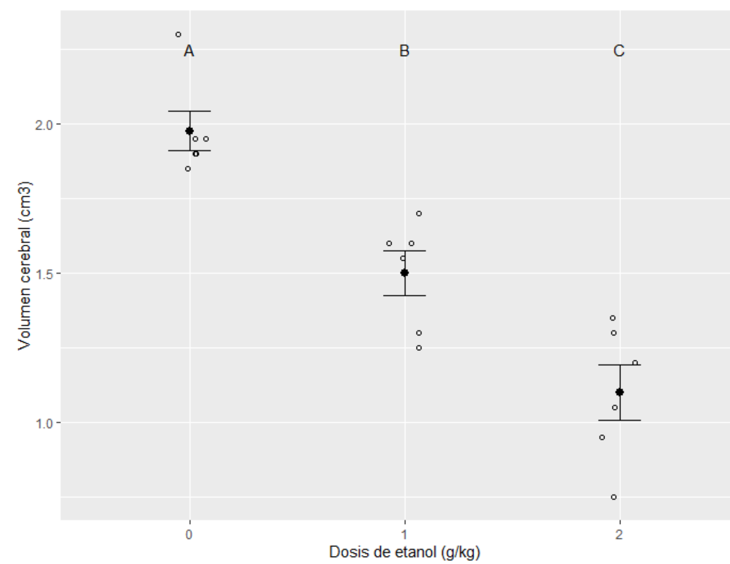
Ho: $\alpha_i = 0$ no existe efecto del factor // las medias poblacionales de los grupos son iguales
H_1: Al menos un $\alpha_i \neq 0$ existe efecto del factor // al menos un grupo difiere en su media poblacional

```
anova(m3)
```

```
Analysis of Variance Table

Response: vol
                Df  Sum Sq Mean Sq F value    Pr(>F)    
factor(etanol)   2 2.30250 1.15125  30.906 4.786e-06 ***
Residuals       15 0.55875 0.03725                      
---
```

```
emmeans(m3, pairwise ~ factor(etanol))
```

```
$emmeans
 etanol emmean     SE df lower.CL upper.CL
      0   1.98 0.0788 15    1.807     2.14
      1   1.50 0.0788 15    1.332     1.67
      2   1.10 0.0788 15    0.932     1.27

Confidence level used: 0.95

$contrasts
 contrast estimate    SE df t.ratio p.value
 0 - 1       0.475 0.111 15 4.263   0.0019
 0 - 2       0.875 0.111 15 7.852   <.0001
 1 - 2       0.400 0.111 15 3.590   0.0071

P value adjustment: tukey method for comparing a family of 3 estimates
```



En resumen

`lm(VR~VE)`

VE cuantitativa

Modelo de regresión

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

VE cualitativa

Modelo de comparación de medias

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

no son equivalentes

son equivalentes

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \varepsilon_i$$

- Más parsimonioso (menos parámetros)
- Permite interpolar
- Modelos lineales en los parámetros
- Primera opción si la VE es cuanti

- Más parámetros
- Inferencia solo para los niveles estudiados
- No implica una función entre Y y X
- Primera opción si la VE es cuali

Para la próxima

- Leer Perelman S y Garibaldi L. 2019. Capítulo 1. Introducción a la estadística experimental.
- Responder ejercicios 1.1, 1.3 y 1.4

Bibliografía general

- Quinn, G. P., & Keough, M. J. (2002). Experimental design and data analysis for biologists. Cambridge University Press
- Agresti A. (2015). Foundations of Linear and Generalized Linear Models . Wiley
- Zuur, A., Ieno, E. N., & Smith, G. M. (2007). Analyzing ecological data. Springer Science & Business Media.
- Faraway, JJ (2002) Practical regression and Anova using R