



Estadística aplicada en R

-Marzo 2022-

Carlota Solano
carlota.solano.udina@upm.es



T-test

- Comparar **dos** grupos
H0= las medias de los dos grupos son iguales
Ha= las medias de los dos grupos son distintas

>t.test(Y ~ X)

ANOVA (One-way)

- Comparar **más de dos** grupos
H0= La media de los grupos no difiere
Ha= La media de los grupos difiere al menos entre dos grupos

>aov(Y ~ X) %>%summary()

ANOVA (Two-way)

- Comparar el efecto de la **combinación** de varios factores
H0= La media de los grupos no difiere
Ha= La media de los grupos difiere al menos entre dos grupos

>aov(Y ~ X1* X2) %>%summary()

Modelo Lineal Simple

- Determinar cómo se relacionan dos variables continuas
H0= No existe relación entre variables $\rightarrow b_1 = 0$
Ha= Existe relación entre variables $\rightarrow b_1 \neq 0$

```
>lm( y ~ x) %>% summary()
```

Modelo Lineal Múltiple con Interacción

- Determinar cómo afecta múltiples variables categóricas y/o continuas a una variable respuesta
H0= Las vars. explicativas no afectan a la var. respuesta
Ha= Las vars. explicativas afectan a la var. respuesta

```
>lm( y ~ x1 * x2) %>% summary()
```

- Diferencia entre **R²** y **p-valor** en un Modelo Lineal

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

proporción de varianza de la var. respuesta
que está explicada por el modelo

P-valor: Significancia del modelo → Nos permite
aceptar o rechazar las hipótesis alternativas.

Earlier this week...

- Diferencia entre **R²** y **p-valor** en un Modelo Lineal

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

proporción de varianza de la var. respuesta
que está explicada por el modelo

P-valor: Significancia del modelo → Nos permite
aceptar o rechazar las hipótesis alternativas.

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.67412	-0.19634	0.00585	0.23708	0.65073

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.54664	0.98028	3.618	0.000437	***
BJ\$BillLength	0.18465	0.04050	4.559	1.25e-05	***
BJ\$KnownSexM	3.38540	1.37280	2.466	0.015086	*
BJ\$BillLength:BJ\$KnownSexM	-0.12525	0.05534	-2.263	0.025438	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

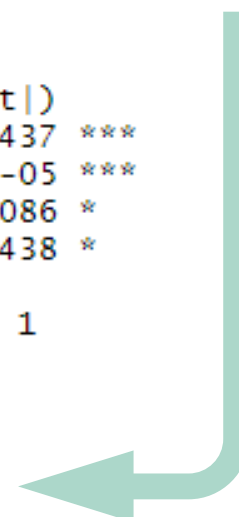
Residual standard error: 0.3003 on 119 degrees of freedom
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076
F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

H0: Las vars. explicativas no afectan a la variable respuesta

La anchura y la longitud no están relacionado ni en
hembras ni en machos

Ha: Las vars. explicativas afectan a la variable respuesta

La anchura y la longitud están relacionadas, y esta relación
difiere entre machos y hembras



- Diferencia entre **R²** y **p-valor** en un Modelo Lineal

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

proporción de varianza de la var. respuesta
que está explicada por el modelo

P-valor: Significancia del modelo → Nos permite
aceptar o rechazar las hipótesis alternativas.

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.67412	-0.19634	0.00585	0.23708	0.65073


Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.54664	0.98028	3.618	0.000437	***
BJ\$BillLength	0.18465	0.04050	4.559	1.25e-05	***
BJ\$KnownSexM	3.38540	1.37280	2.466	0.015086	*
BJ\$BillLength:BJ\$KnownSexM	-0.12525	0.05534	-2.263	0.025438	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom
Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076
F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14

H0: Las vars. explicativas no afectan a la variable respuesta

La anchura y la longitud no están relacionado ni en
hembras ni en machos →  ptes ≠ 0

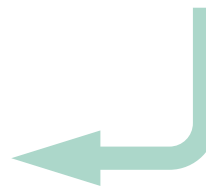
Ha: Las vars. explicativas afectan a la variable respuesta

La anchura y la longitud están relacionadas, y esta relación
difiere entre machos y hembras →



pte de hembras ≠ 0

pte de machos ≠ pte de hembras



Earlier this week...

- Diferencia entre **R²** y **p-valor** en un Modelo Lineal

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

proporción de varianza de la var. respuesta
que está explicada por el modelo

P-valor: Significancia del modelo → Nos permite
aceptar o rechazar las hipótesis alternativas.

```
> lm(BJ$BillDepth~BJ$BillLength*BJ$KnownSex)%>%summary()
```

Call:

```
lm(formula = BJ$BillDepth ~ BJ$BillLength * BJ$KnownSex)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.67412	-0.19634	0.00585	0.23708	0.65073

Coefficients:

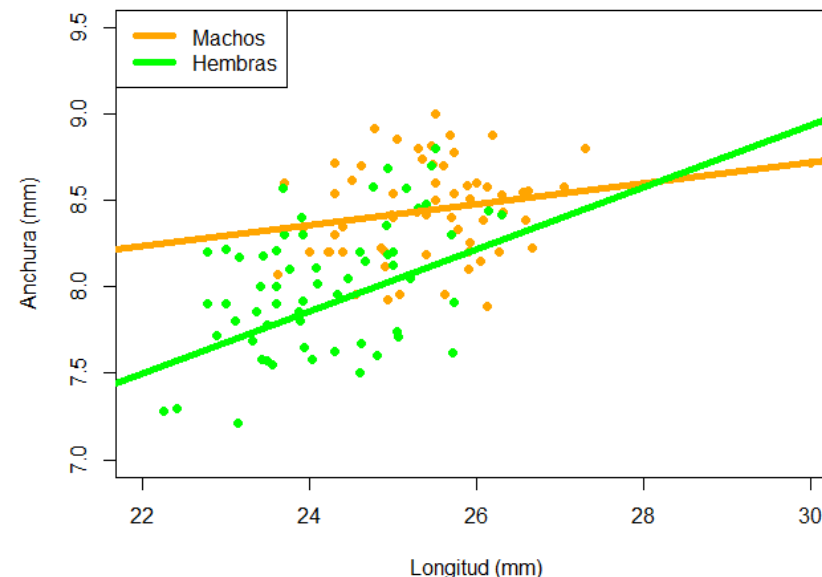
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.54664	0.98028	3.618	0.000437	***
BJ\$BillLength	0.18465	0.04050	4.559	1.25e-05	***
BJ\$KnownSexM	3.38540	1.37280	2.466	0.015086	*
BJ\$BillLength:BJ\$KnownSexM	-0.12525	0.05534	-2.263	0.025438	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3003 on 119 degrees of freedom

Multiple R-squared: 0.4221, Adjusted R-squared: 0.4076

F-statistic: 28.98 on 3 and 119 DF, p-value: 3.876e-14



- Modelo Lineal Sin Interacción

Cuando nuestro LM con interacción muestra que no existe una interacción entre las variables explicativas, es decir, que la pendiente del grupo no de referencia no difiere de la del grupo de referencia, podemos simplificar el modelo (*Principio de Parsimonia*) haciendo un modelo lineal sin interacción.

```
>lm( y ~ x1 + x2 ) %>% summary()
```


- Modelo Lineal Sin Interacción

```
> lm(data=iris, Sepal.Length~Petal.Length*Species)%>%summary()
```

Call:
lm(formula = Sepal.Length ~ Petal.Length * Species, data = iris)

Residuals:

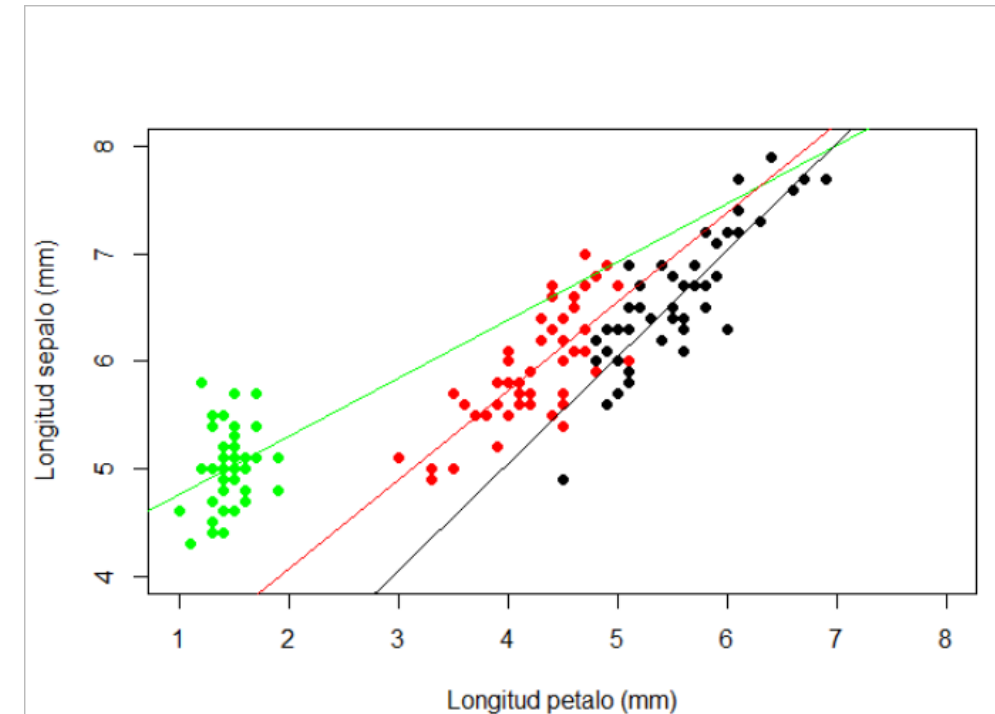
	Min	1Q	Median	3Q	Max
	-0.73479	-0.22785	-0.03132	0.24375	0.93608

Coefficients:

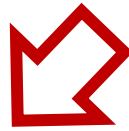
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.2132	0.4074	10.341	< 2e-16	***
Petal.Length	0.5423	0.2768	1.959	0.05200	.
Speciesversicolor	-1.8056	0.5984	-3.017	0.00302	**
Speciesvirginica	-3.1535	0.6341	-4.973	1.85e-06	***
Petal.Length:Speciesversicolor	0.2860	0.2951	0.969	0.33405	
Petal.Length:Speciesvirginica	0.4534	0.2901	1.563	0.12029	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3365 on 144 degrees of freedom
Multiple R-squared: 0.8405, Adjusted R-squared: 0.8349
F-statistic: 151.7 on 5 and 144 DF, p-value: < 2.2e-16



- Modelo Lineal Sin Interacción



```
> lm(data=iris, Sepal.Length~Petal.Length+Species)%>%summary()
```

Call:

```
lm(formula = Sepal.Length ~ Petal.Length + Species, data = iris)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.75310	-0.23142	-0.00081	0.23085	1.03100

Coefficients:

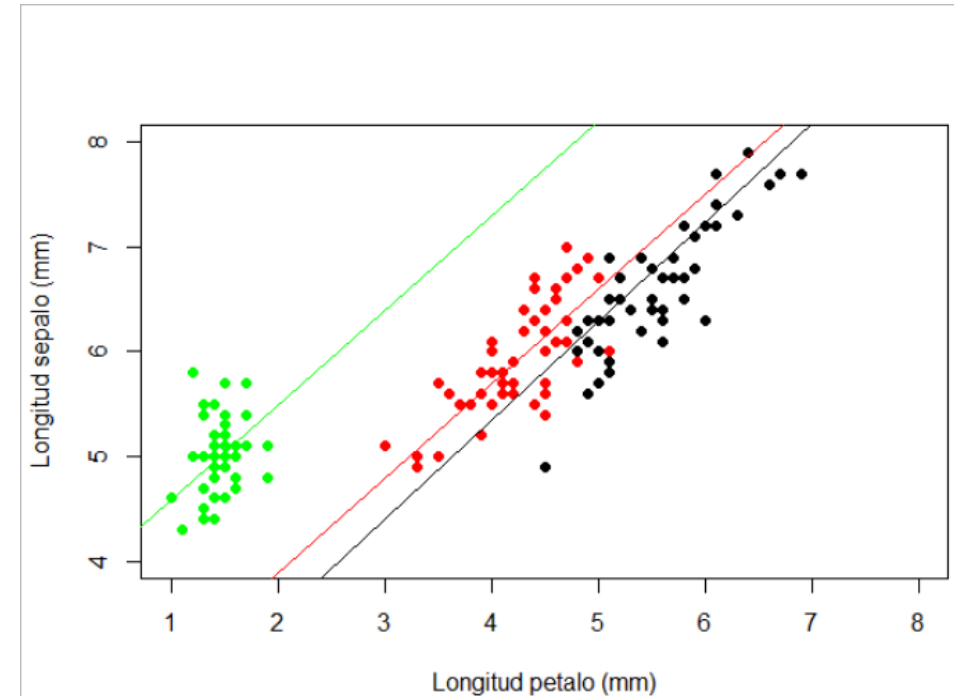
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.68353	0.10610	34.719	< 2e-16	***
Petal.Length	0.90456	0.06479	13.962	< 2e-16	***
Speciesversicolor	-1.60097	0.19347	-8.275	7.37e-14	***
Speciesvirginica	-2.11767	0.27346	-7.744	1.48e-12	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.338 on 146 degrees of freedom

Multiple R-squared: 0.8367, Adjusted R-squared: 0.8334

F-statistic: 249.4 on 3 and 146 DF, p-value: < 2.2e-16



- ¿Cómo cargar bases de datos en R?

```
>setwd("C:/Users/Usuario/Desktop/Talks/StatsIntroR")  
>miBDenR<-read.csv("nombreBaseDatos.csv",sep=";",dec=".")  
>miBDenR
```

- ¿Cómo cambiar el tipo de dato de nuestras variables?

```
>miBD$var_cat<-as.factor(miBD$var_cat)  
      as.numeric(...)  
      as.integer(...)
```

- *¿Qué hacer con variables respuesta que son datos numéricos NO continuos, i.e. enteros, binarios, etc.?*

Número de pie
Número de especies
Número de visitas a una flor
Número de respiraciones por minuto
...



Conteos = números enteros = **Conteos**

- *¿Qué hacer con variables respuesta que son datos numéricos NO continuos, i.e. enteros, binarios, etc.?*

Número de pie

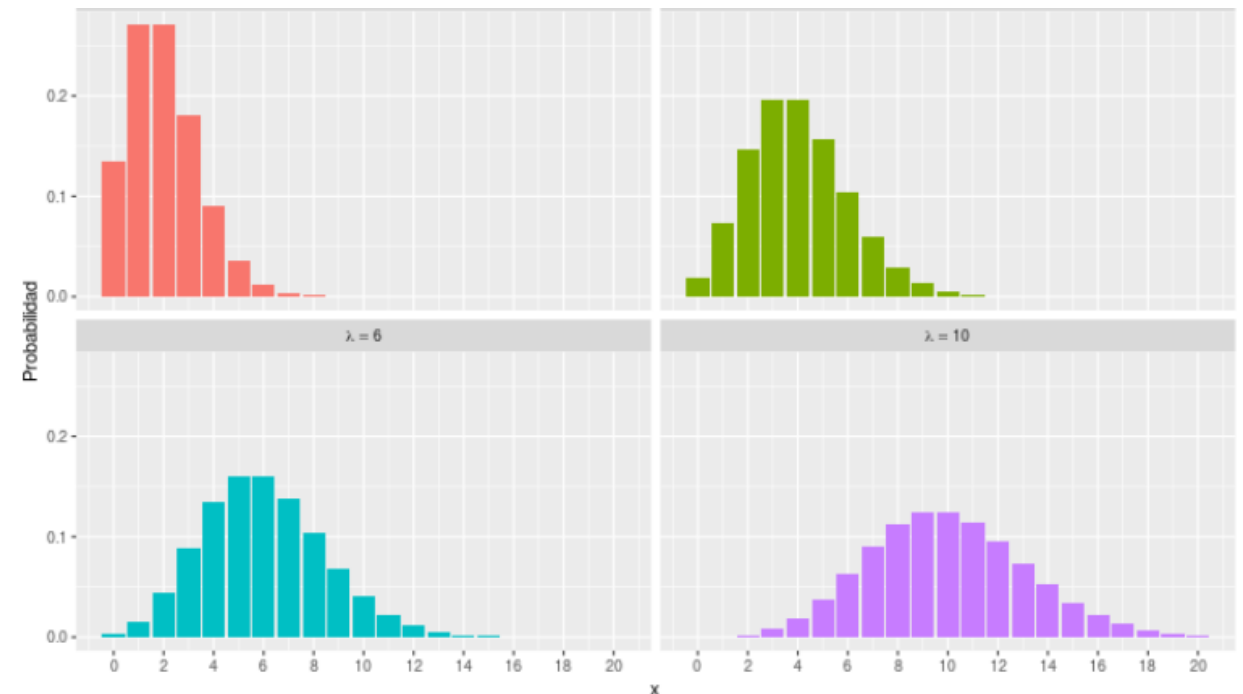
Número de especies

Número de visitas a una flor

Número de respiraciones por minuto

...

Conteos = números enteros = **Conteos**



- *¿Qué hacer con variables respuesta que son datos numéricos NO continuos, i.e. enteros, binarios, etc.?*

Número de pie
Número de especies
Número de visitas a una flor
Número de respiraciones por minuto
...



Conteos = números enteros = **Conteos**

GLM con distribución de Poisson

```
>glm(data, y~x1*x2,family="poisson")
```



LINK FUNCTIONS (Funciones de enlace): permite modificar una relación no lineal para que se ajuste a un modelo lineal.

¡Importante! Transformar de vuelta las estimaciones

GLM Poisson → Log → $\ln(x)$

- *¿Qué hacer con variables respuesta que son datos numéricos NO continuos, i.e. enteros, binarios, etc.?*

Presencia/Ausencia
Germinado/No germinado
Vivo/Muerto
Cara/Cruz
Éxito/Fracaso
...



0 vs 1= números que son categorías = **Binario**

- *¿Qué hacer con variables respuesta que son datos numéricos NO continuos, i.e. enteros, binarios, etc.?*

Presencia/Ausencia
Germinado/No germinado
Vivo/Muerto
Cara/Cruz
Éxito/Fracaso
...

0 vs 1= números que son categorías = **Binario**

GLM con distribución Binomial /Regresión logística
`>glm(data, y~x1*x2,family="binomial")`

R

LINK FUNCTIONS (Funciones de enlace):
permite modificar una relación no lineal
para que se ajuste a un modelo lineal.

¡Importante! Transformar de vuelta las estimaciones

GLM Binomial \rightarrow Logit $\rightarrow \ln\left(\frac{x}{1-x}\right)$

¡Muchas gracias por tu apoyo, interés y confianza!



Carlota Solano Udina



<https://www.linkedin.com/in/carlota-solano-udina/>



<https://github.com/Calamardotis>



<https://twitter.com/Calamardotis>