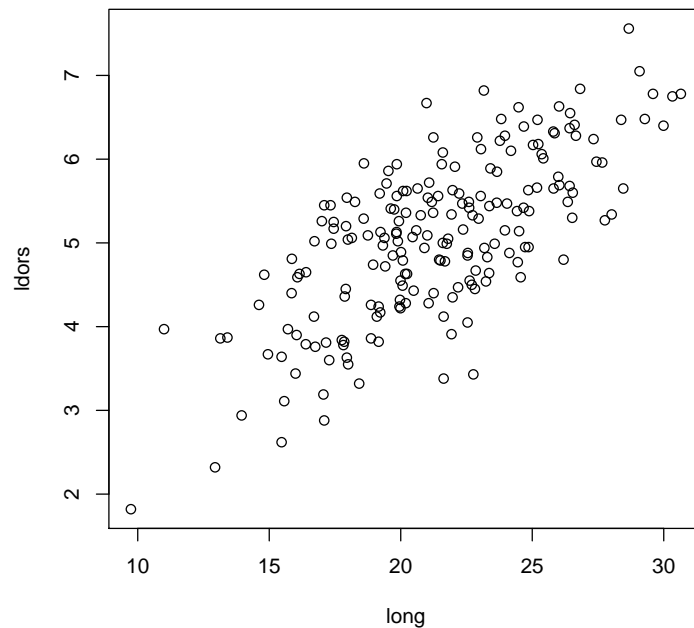


# Ejemplo de Análisis de la Covarianza

---

Utilizando los datos del archivo [sargos.csv](#) representamos la longitud desde el morro hasta la aleta dorsal frente a la longitud total del pez:

```
> sargos = read.table(file = "http://dl.dropbox.com/u/7610774/sargos.csv",  
  header = T, sep = ";", dec = ",")  
> attach(sargos)  
> plot(long, ldors)
```



## Estimación de la recta de regresión.

Para estimar la recta de regresión:

$$\text{ldors} = \beta_0 + \beta_1 \text{long} + \varepsilon$$

que permite predecir [ldors](#) (que consideramos aquí como variable respuesta) en función de [long](#) (que consideramos como variable explicativa) utilizamos la función [lm](#):

```
> llmod = lm(ldors ~ long)  
> llmod
```

Call:

```
lm(formula = ldors ~ long)
```

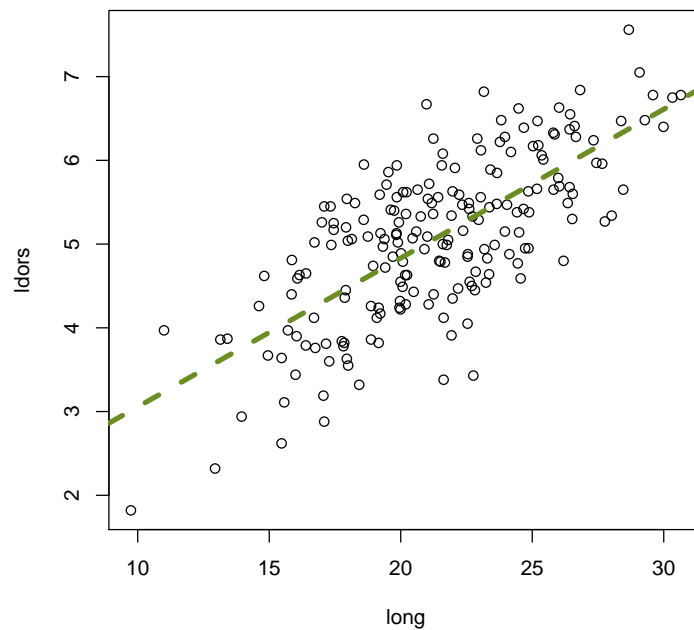
Coefficients:

```
(Intercept)      long  
    1.280      0.178
```

### Representación gráfica de la recta.

Podemos superponer la ecuación de esta recta a la nube de puntos mediante `abline()`:

```
> plot(long, ldors)  
> abline(l1mod, col = "olivedrab", lty = 2, lwd = 4)
```



El comando `summary()` nos permite contrastar la significación de los parámetros de ajuste; asimismo nos muestra el contraste de regresión y nos proporciona el coeficiente de determinación  $R^2$ :

```
> summary(l1mod)
```

Call:

```
lm(formula = ldors ~ long)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.89406	-0.49911	0.00266	0.48314	1.66225

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.2795	0.2678	4.78	3.5e-06	***
long	0.1777	0.0123	14.39	< 2e-16	***

---

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

Residual standard error: 0.677 on 198 degrees of freedom

Multiple R-squared: 0.511, Adjusted R-squared: 0.509

F-statistic: 207 on 1 and 198 DF, p-value: <2e-16

### Determinación de posibles diferencias entre sexos en la relación long-ldors.

Nos preguntamos ahora si la relación long-ldors es la misma para ambos sexos o no. En primer lugar resulta interesante explorar las diferencias entre sexos mediante un gráfico en el que cada sexo se represente de forma distinta, tal como se muestra en la figura 1. Esta figura se ha generado con la sintaxis:

```
> require(car)
> scatterplot(ldors ~ long | sexo, reg.line = lm,
              smooth = F)
```

Vemos que ambos sexos se ajustan a rectas de prácticamente la misma pendiente, pero con diferente ordenada en el origen. Podemos comprobar que los rangos de tallas y distancias a la aleta dorsal son muy similares en machos y hembras:

```
> aggregate(cbind(long, ldors) ~ sexo, data = sargos,
            function(x) {
              c(min = min(x), max = max(x))
            })
```

	sexo	long.min	long.max	ldors.min	ldors.max
1	Hembra	11.00	30.65	3.11	7.56
2	Macho	9.74	29.99	1.82	6.84

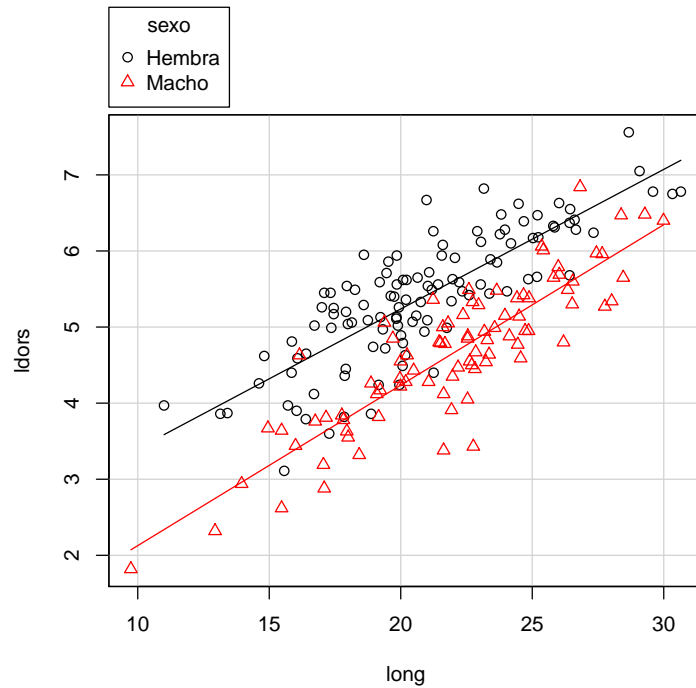


Figura 1: Relación long-ldors en machos y hembras

El gráfico, en cambio, nos muestra que, para una misma longitud, la distancia a la aleta dorsal en los machos es inferior a la distancia a la aleta dorsal en las hembras. Dicho de otro modo, los machos tienden a tener la aleta dorsal más adelantada que las hembras de la misma longitud.

Podemos ajustar una recta a cada grupo:

```
> require(nlme)
> llsex = lmList(ldors ~ long | sexo, data = sargos,
  na.action = na.exclude)
> llsex
```

Call:

```
Model: ldors ~ long | sexo
Data: sargos
```

Coefficients:

	(Intercept)	long
Hembra	1.8349	0.1725
Macho	0.3803	0.1970

Degrees of freedom: 102 total; 98 residual

Residual standard error: 0.4554

y mostrar intervalos de confianza para cada uno de los parámetros del modelo (pendientes y ordenadas en el origen):

```
> intervals(llsex)
```

```
, , (Intercept)
```

	lower	est.	upper
Hembra	0.7930	1.8349	2.877
Macho	-0.6507	0.3803	1.411

```
, , long
```

	lower	est.	upper
Hembra	0.1291	0.1725	0.2159
Macho	0.1540	0.1970	0.2400

La figura 2 muestra gráficamente estos intervalos. Estos intervalos sugieren que ambas rectas tienen efectivamente la misma pendiente y se diferencian sólo en la ordenada.

Podemos ahora tratar de ajustar un modelo en el que, efectivamente, haya una pendiente común para ambas regresiones, con diferencia sólo en la ordenada. Este modelo recibe el nombre de *modelo de Análisis de la Covarianza (ANCOVA)*, y puede estimarse de manera bastante sencilla mediante:

```
> llsexA = lm(ldors ~ long + sexo, data = sargos)
> summary(llsexA)
```

Call:

```
lm(formula = ldors ~ long + sexo, data = sargos)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.3771	-0.3058	0.0319	0.2960	1.2510

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.30849	0.19205	6.81	1.1e-10 ***
long	0.19592	0.00895	21.89	< 2e-16 ***
sexoMacho	-0.96063	0.07003	-13.72	< 2e-16 ***

---

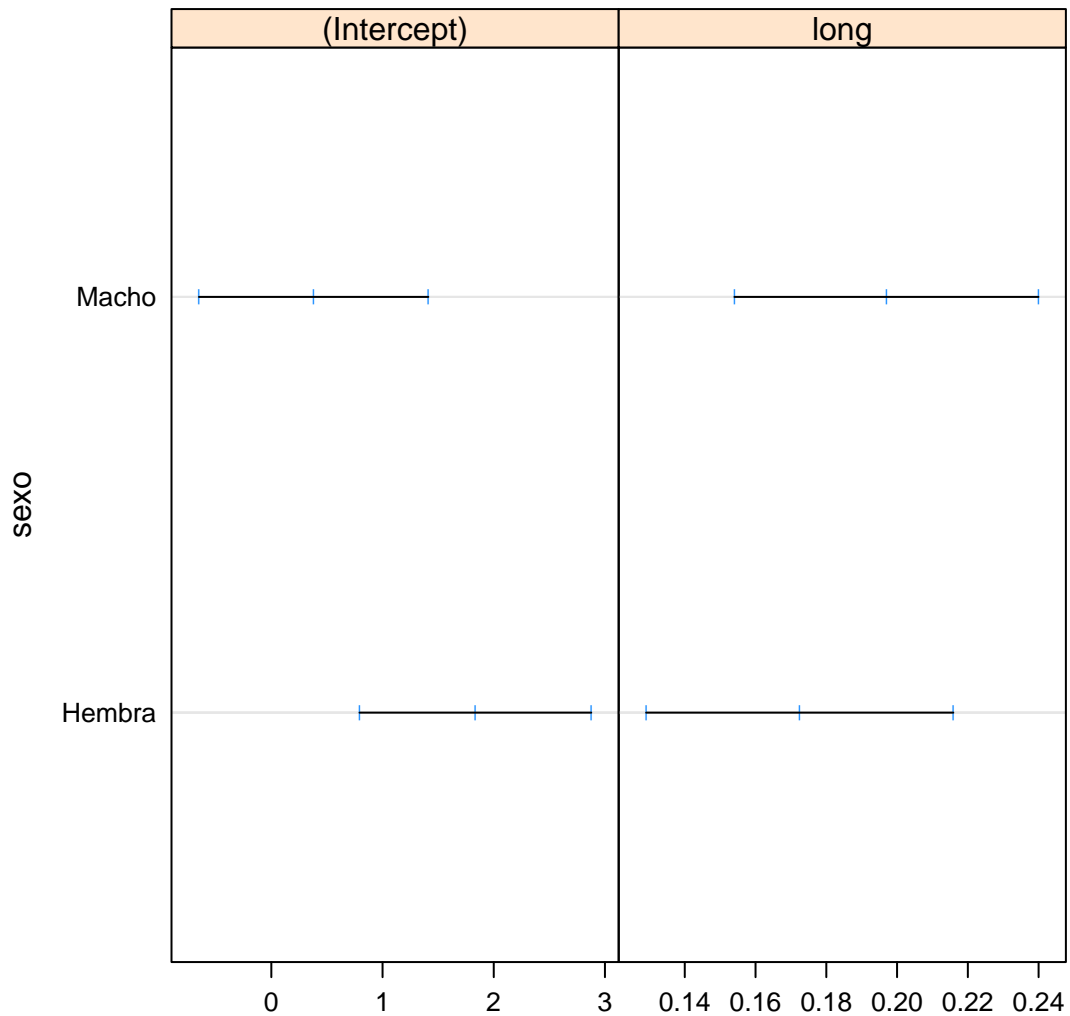


Figura 2: Intervalos de confianza para los parámetros de las regresiones de l dors frente a long según sexo.

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

Residual standard error: 0.486 on 197 degrees of freedom

Multiple R-squared: 0.75, Adjusted R-squared: 0.748

F-statistic: 296 on 2 and 197 DF, p-value: <2e-16

Como vemos, este modelo queda caracterizado por tres parámetros: la *pendiente común* (0.1959), la *ordenada* para la regresión correspondiente a las hembras (1.308) y la diferencia entre la ordenada correspondiente a las hembras y la correspondiente a los machos (-0.9606). Este último término nos indica que, para una misma longitud, la distancia a la aleta dorsal en un macho es, por término medio, 0.9606 inferior a la distancia equivalente en las hembras. Como podemos ver en la tabla anterior, los tres coeficientes son significativos, esto es, su valor es no nulo.

Podemos preguntarnos si este modelo es realmente mejor que alguno de los siguientes:

1. El modelo más simple en el que un único modelo de regresión permite explicar la relación entre `long` y `ldors` sin tener en cuenta el sexo (el modelo de una única recta de regresión estimado más arriba)
2. El modelo más complejo en el que son precisas dos ecuaciones de regresión distintas (diferentes en pendientes y ordenadas) para representar la relación entre ambas variables en cada sexo.

Ambos modelos ya han sido estimados más arriba. En particular, para comparar con el modelo de una sólo recta con el modelo de dos rectas de la misma pendiente, llevamos a cabo el *test de razón de verosimilitudes*:

```
> anova(llmod, llsexA)
```

```
Analysis of Variance Table
```

```
Model 1: ldors ~ long
```

```
Model 2: ldors ~ long + sexo
```

```
  Res.Df  RSS Df Sum of Sq  F Pr(>F)
1     198 90.8
2     197 46.4  1    44.4 188 <2e-16 ***
```

```
---
```

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

que nos indica que el modelo `llsexA` es mejor que el modelo de una única recta (ya que la diferencia entre ambos es significativa y el modelo 2, el `llsexA` produce un mejor ajuste, medido en términos de una suma de cuadrados residual, RSS, inferior). Podemos proceder de igual manera para comparar con el modelo con dos rectas diferentes; en este caso, debemos estimar el modelo de la forma:

```
> llsex2R = lm(ldors ~ long * sexo, data = sargos)
```

```
> summary(llsex2R)
```

```
Call:
```

```
lm(formula = ldors ~ long * sexo, data = sargos)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-1.3884 -0.3052  0.0334  0.3124  1.2528
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.5677	0.2555	6.13	4.6e-09	***
long	0.1835	0.0121	15.21	< 2e-16	***
sexoMacho	-1.5497	0.3911	-3.96	0.00010	***
long:sexoMacho	0.0274	0.0179	1.53	0.12744	

---

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

Residual standard error: 0.484 on 196 degrees of freedom

Multiple R-squared: 0.753, Adjusted R-squared: 0.749

F-statistic: 199 on 3 and 196 DF, p-value: <2e-16

Esta estimación nos proporciona la pendiente para la recta de las hembras (0.1835) y la diferencia con la pendiente de la recta para los machos (0.02743); asimismo nos da la ordenada para la recta de las hembras ((1.568) y la diferencia con la ordenada de los machos (-1.550). Como vemos, el contraste ya nos indica que la diferencia entre la pendiente de las hembras y la de los machos no es significativa, y sí lo es la diferencia entre ordenadas. El test de razón de verosimilitudes para comparar este último modelo con [llsexA](#), nos indica que entre ambos no hay diferencias significativas:

```
> anova(llsexA, llsex2R)
```

Analysis of Variance Table

Model 1: ldors ~ long + sexo

Model 2: ldors ~ long \* sexo

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	197	46.4				
2	196	45.9	1	0.549	2.34	0.13

Por tanto, siguiendo el principio de parsimonia consideramos como modelo más adecuado el más simple, esto es, [llsexA](#) (el modelo ANCOVA).