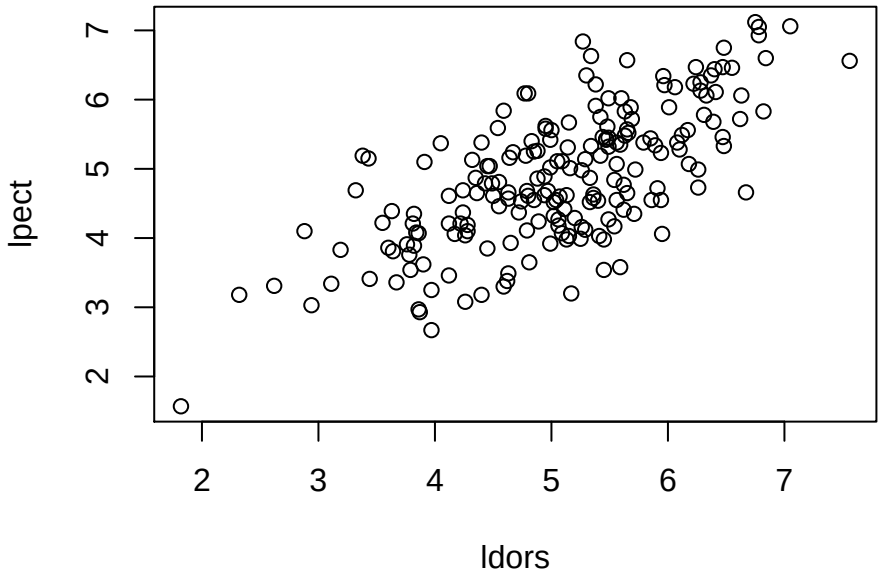
Ejemplo de Regresión Lineal Simple.

El archivo *sargos.csv* contiene datos morfométricos de una muestra de 200 sargos. Estos datos pueden leerse en R mediante la sintaxis:

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

En la siguiente gráfica representamos la longitud en centímetros desde el morro hasta la aleta pectoral frente a la longitud desde el morro a la dorsal:

```
plot(ldors,lpect)
```



Estimación de la recta de regresión.

Para estimar la recta de regresión:

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

que permite predecir *lpect* (que consideramos aquí como variable respuesta) en función de *ldors* (que consideramos como variable explicativa) utilizamos la función *lm*:

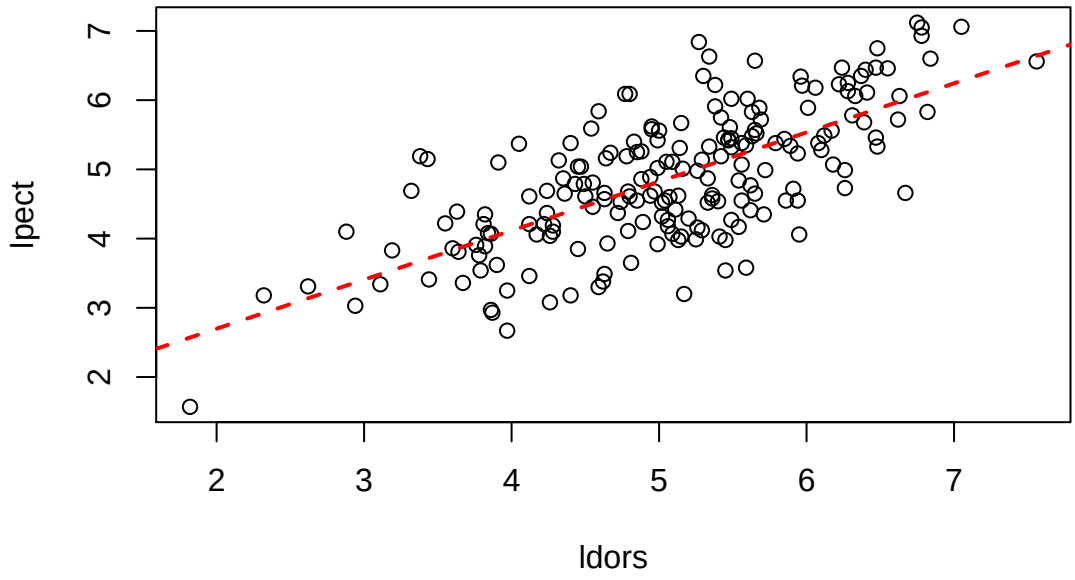
```
recta=lm(lpect~ldors)
recta

##
## Call:
## lm(formula = lpect ~ ldors)
##
## Coefficients:
## (Intercept)        ldors
##       1.281        0.709
```

Representación gráfica de la recta.

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

```
plot(ldors,lpect)
abline(recta,col="red",lty=2,lwd=2)
```



Interpretación de los coeficientes y sus intervalos de confianza.

El valor del coeficiente de regresión estimado $\hat{\beta}_1$ =0.7089 significa que por cada unidad (en este caso por cada centímetro) que aumenta la distancia del morro a la dorsal, `ldors`, la distancia a la pectoral, `lpect`, se incrementa, *por término medio* en 0.7089 cm.

Podemos (y debemos) preguntarnos con qué margen de error se realiza esta estimación de los coeficientes. El siguiente comando nos calcula los correspondientes intervalos de confianza para β_0 y β_1 respectivamente:

```
confint(recta)

##                2.5 % 97.5 %
## (Intercept) 0.7518 1.8093
## ldors       0.6065 0.8113
```

De esta forma, concluimos que por cada cm. que se incremente `ldors`, con un 95 % de confianza podemos asegurar que `lpect` se incrementa por término medio entre 0.606 y 0.811 cm. (el intervalo de confianza para β_1).

En cuanto al valor de $\hat{\beta}_0$ (1.2806, con un intervalo de confianza al 95 % que va de 0.752 a 1.809), es un coeficiente que en este caso se interpreta simplemente como coeficiente de ajuste para que la recta pase por el centro de la nube de puntos. En teoría este valor correspondería al valor de `lpect` cuando `ldors=0`; pero como no tiene sentido hablar de peces para los que `ldors=0`, no tiene tampoco sentido esta interpretación teórica.

Contraste de la regresión.

Aunque de hecho el intervalo de confianza para β_1 ya nos informa de que este parámetro es casi seguro no nulo (ya que el intervalo al 95 % no contiene al cero) es frecuente en los trabajos de investigación presentar el p-valor del contraste de regresión:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Este p-valor se encuentra en la última linea de la salida del siguiente comando:

```
summary(recta)

##
## Call:
## lm(formula = lpect ~ ldors)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7456 -0.5020  0.0193  0.5202  1.8235
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.2806     0.2681    4.78  3.5e-06 ***
## ldors         0.7089     0.0519   13.65  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.708 on 198 degrees of freedom
## Multiple R-squared:  0.485,Adjusted R-squared:  0.482
## F-statistic:  186 on 1 and 198 DF,  p-value: <2e-16
```

Como el p-valor es menor que 2.2e-16 (a los efectos prácticos es casi como si fuese cero), el p-valor es menor que cualquier nivel de significación α que quisiéramos elegir para el contraste anterior. Como $p - valor < \alpha$ concluimos que existe evidencia suficiente para rechazar H_0 y concluir por tanto que $\beta_1 \neq 0$.

Predicción de valores individuales.

Se desea utilizar la recta anterior para predecir las distancias desde el morro a la aleta pectoral de peces cuyas distancias desde el morro a la dorsal son de 4, 5 y 6 cm. (**Nota:** obsérvese en el gráfico que estos tres valores están dentro del rango observado de valores para la distancia del morro a la dorsal. Aunque podríamos utilizar la recta para hacer predicciones cuando dicha distancia es por ejemplo 40 cm, esta predicción es absurda por varias razones: porque previsiblemente no existen sargos con esta distancia desde el morro a la dorsal; y si existieran no los hemos observado; y como no los hemos observado no sabemos si para tal valor la relación va a seguir siendo la misma que en el rango que sí hemos observado, de 2 a 8 cm. aproximadamente. En conclusión: nunca debe utilizarse una recta para extrapolar más allá del rango de las observaciones). Para hacer la predicción utilizamos los siguientes comandos:

```r
xpred=data.frame(ldors=c(4,5,6))
predict(recta,xpred,interval="prediction",level=0.95)
```

```
##     fit   lwr   upr
## 1 4.116 2.712 5.520
## 2 4.825 3.426 6.224
## 3 5.534 4.131 6.937
```

De esta forma, *para un sargo particular* de 4 cm en `ldors` cabe esperar que `lpect` valga, por término medio, 4.116 cm, si bien con un 95 % de confianza el valor de `lpect` puede de hecho estar entre 2.712 y 5.520 cm. Del mismo modo puede interpretarse la predicción y el intervalo de confianza para los otros dos valores.

Predicción del valor medio condicionado.

Si la pregunta que nos hacemos, sin embargo, es cuál es la distancia media desde el morro hasta la aleta pectoral de todos los sargos que miden, respectivamente, 4, 5 o 6 cm, utilizaremos la siguiente sintaxis:

```
xpred=data.frame(ldors=c(4,5,6))
predict(recta,xpred,interval="confidence",level=0.95)

##      fit   lwr   upr
## 1 4.116 3.968 4.264
## 2 4.825 4.726 4.924
## 3 5.534 5.397 5.671
```

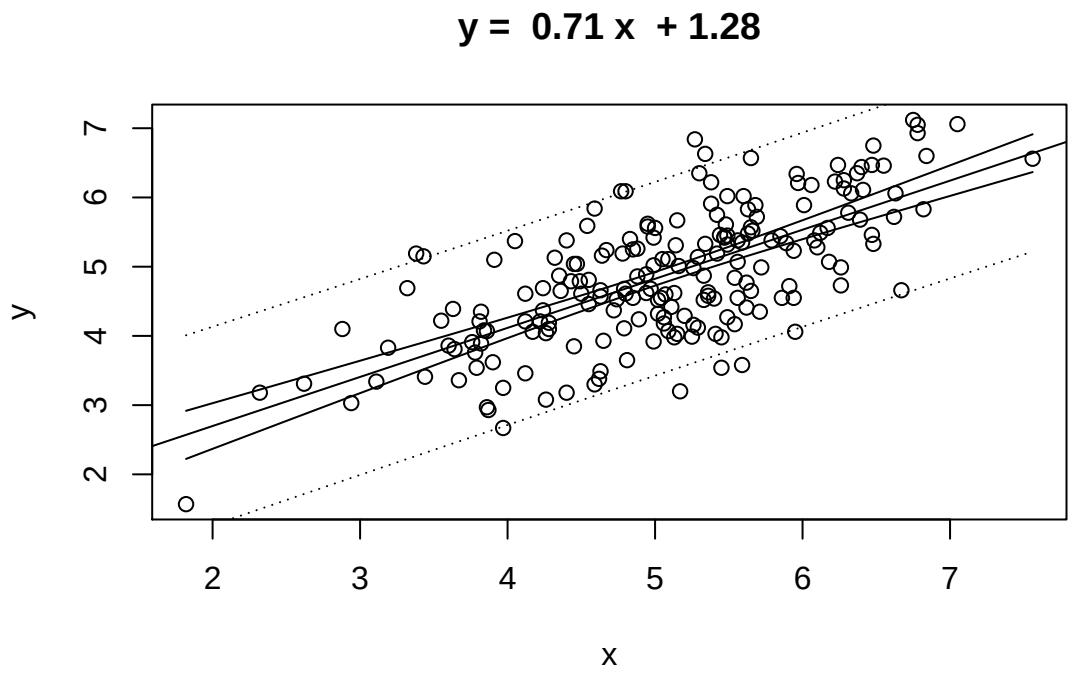
En este caso concluiríamos que *para todos los sargos cuya* `ldors` *mide 4 cm*, el valor medio de `lpect` es aproximadamente 4.116 cm, si bien, con una confianza del 95 % dicho valor medio podría estar entre 3.968 y 4.263 cm. Idéntica interpretación puede hacerse de los valores obtenidos para `ldors`=5 y 6.

Gráfico de las bandas de confianza.

Los intervalos de confianza para las predicciones, tanto de los valores individuales como de los valores medios condicionados, pueden representarse en forma de *bandas de confianza* para todo el recorrido de la recta de regresión usando la función `simple.lm` de la librería UsingR:

```
require(UsingR)
simple.lm(ldors,lpect,show.ci=TRUE,conf.level=0.95)
```

```
## Loading required package: UsingR
## Loading required package: MASS
```




```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##       1.281        0.709
```

La banda más próxima a la recta (en trazo continuo) corresponde a los intervalos de confianza para los valores medios; la más externa (en trazo discontinuo) a la predicción de los valores individuales.

Validación del ajuste de la recta.

La primera validación es gráfica. Para que una recta esté correctamente ajustada debe apreciarse que los puntos se distribuyen en torno a la recta y que la variabilidad es aproximadamente la misma a lo largo de todo el recorrido de ésta. Esto es precisamente lo que se ve en el gráfico anterior, por lo que el ajuste lineal parece razonablemente bueno.

No obstante para la validez de toda la inferencia hecha hasta ahora es preciso que los residuos de la recta tengan distribución normal (ya que los intervalos y contrastes que hemos calculado se apoyan, desde el punto de vista teórico, en la normalidad). La normalidad puede contrastarse mediante el test de Shapiro-Wilk:

$$\begin{cases} H_0: & \text{Los residuos siguen una distribución normal} \\ H_1: & \text{La distribución de los residuos no es normal} \end{cases}$$

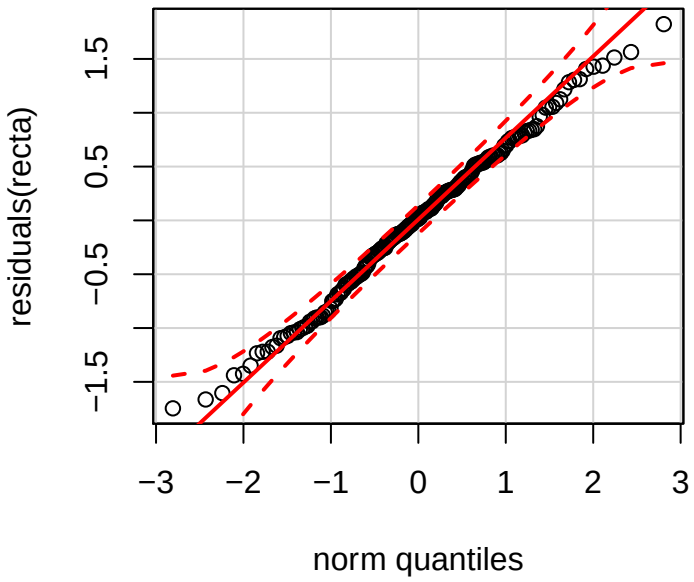
En R este test se lleva a cabo mediante:

```
shapiro.test(residuals(recta))

##
##  Shapiro-Wilk normality test
##
## data:  residuals(recta)
## W = 0.9947, p-value = 0.7014
```

o mediante un gráfico *rcodeqqPlot* del paquete `car`:

```
library(car)
qqPlot(residuals(recta))
```



Como vemos, el p-valor del contraste (0.70) es mayor que α (que podemos elegir como 0.05), por lo que nada se opone a aceptar H_0. Como puede aceptarse que los residuos son normales, la inferencia realizada hasta ahora (intervalos de confianza, contraste de la regresión) es válida.

Inferencia sobre el coeficiente de correlación.

Podemos valorar la intensidad de la asociación lineal entre `lpect` y `ldors` mediante el coeficiente de correlación lineal de Pearson:

```r
cor(lpect,ldors)

## [1] 0.6963
```

Como vemos, la correlación está en torno a 0.7, que es un valor intermedio (ni muy alto ni muy bajo). ¿Cuál es el error de estimación de este coeficiente? Para ello podemos calcular un intervalo de confianza. La función en R que calcula este intervalo es `cor.test()`, que también realiza el *test de independencia lineal* (que en regresión simple es equivalente al contraste de la regresión):

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$$

Para nuestros datos:

```r
cor.test(lpect,ldors)

##
##  Pearson's product-moment correlation
##
## data:  lpect and ldors
## t = 13.65, df = 198, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6172 0.7615
## sample estimates:
##    cor
## 0.6963
```

Así pues, vemos que aunque nuestra estimación de $\hat{\rho}$ es 0.696, con un 95 % de confianza el valor de este parámetro podría estar en el intervalo 0.617 a 0.761. Asimismo el p-valor para el test anterior es inferior a 2.2e-16, lo que indica que es menor que cualquier α que queramos usar y por tanto los datos muestran evidencia suficiente para rechazar H_0. De esta forma concluimos que ρ es significativamente distinto de cero (esto es, el valor 0.696 no se debe al azar sino a que de hecho ambas variables guardan realmente una relación lineal).