Estadística con R

intro-R

Febrero 2018

Secciones

0. ¿Qué vamos o no vamos a hacer en estos 10-15 minutos?
1. Estadística ¿para qué?
2. Mi hipótesis. Errores de tipo I y de tipo II
3. Comparación de la media dos grupos: Test de hipótesis: t-Student.
4. Modelos lineales generales (MLG)
    - Regresión lineal
    - Anova

Secciones

0. ¿Qué vamos o no vamos a hacer en estos 10-15 minutos?
1. Estadística ¿para qué?
2. Mi hipótesis. Errores de tipo I y de tipo II
3. Comparación de la media dos grupos: Test de hipótesis: t-Student.
4. Modelos lineales generales (MLG)
    - Regresión lineal
    - Anova

Secciones

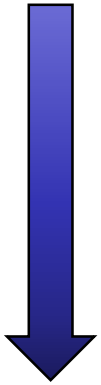
0. ¿Qué vamos o no vamos a hacer en estos 10-15 minutos?
1. Estadística ¿para qué?
2. Mi hipótesis. Errores de tipo I y de tipo II
3. Comparación de la media dos grupos: Test de hipótesis: t-Student.
4. Modelos lineales generales (MLG)
    - Regresión lineal
    - Anova

1. Estadística ¿para qué?

A) Saciar nuestra curiosidad.

- ¿Ocurren las cosas por azar? ¿Puede deberse a algún motivo?
- ¿Hay diferencias entre dos o más cosas?
- ¿Existe relación entre dos o más cosas?
- Si las cosas ocurren por azar o siguen patrones.

B) Ayudar a la toma de decisiones.



Pasos		
1)	Diseño del experimento	
2)	Observar algo	Toma de datos
3)	Análisis estadístico	
	1) Nuevo conocimiento	
	2) Predecicciones	

Secciones

0. ¿Qué vamos o no vamos a hacer en estos 10-15 minutos?
1. Estadística ¿para qué?
2. Mi hipótesis. Errores de tipo I y de tipo II
3. Comparación de la media dos grupos: Test de hipótesis: t-Student.
4. Modelos lineales generales (MLG)
    - Regresión lineal
    - Anova

2. Mi hipótesis. Errores de tipo I y de tipo II

Teorema del mono infinito

H_0 = El mono escribe por azar

Siendo cierta H_0, no es imposible que el mono escriba un libro.

Mi hipótesis: Quiero probar si ocurre algo.

H_0 = Hipótesis nula

H_1 = Hipótesis alternativa

→ **Aceptamos Ho**, cometemos un error de tipo II, porque existe la posibilidad de que escriba un libro.

→ **Rechazamos Ho**, cometemos un error de tipo I, porque el mono escribe por azar.

Test estadísticos
Tratan de minimizar los errores de tipo I (valor de α o p-valor).

Secciones

0. ¿Qué vamos o no vamos a hacer en estos 10-15 minutos?
1. Estadística ¿para qué?
2. Mi hipótesis. Errores de tipo I y de tipo II
3. Comparación de la media dos grupos: Test de hipótesis: t-Student.
4. Modelos lineales generales (MLG)
    - Regresión lineal
    - Anova

3. Comparación de la media de grupos: Test de hipótesis ¿t-Student o Anova?

Variable	Test paramétrico
2 poblaciones	t-Student
N poblaciones	Anova

Test de la t-Student: se utiliza para comparar los valores cuantitativos de la variable respuesta (VR) de **2** muestras/grupos.

Asunciones de la t-Student: hay que comprobarlo SIEMPRE

1. Normalidad de la variable respuesta (VR).
   - Test de Siegel
   - Test de Kolmogorov-Smirnov; **Test de Shapiro-Wilks**
2. Homocedasticidad: la varianza de los grupos son iguales.
   - Test de la F de Snedecor (poco recomendable porque es muy sensible a la violación de normalidad).
   - **Test de Levene**
   - Test de Brown-Forsythe
3. Las muestras son independientes **(→Muestreo aleatorio).**

Si esto **no** se cumple:

```
> for (i in 1:10^9){
      print("Debo
       cambiar de test")
    }
```

3. Comparación de la media de grupos: Test de hipótesis **t-Student**

Ejemplo (caso particular)

Observamos dos grupos de 30 niños de la misma edad y medimos su altura. Un grupo de niños juega al fútbol y el otro grupo de niños juega al baloncesto. ¿Serán sus alturas diferentes?

H_0 = la alturas medias no son distintas

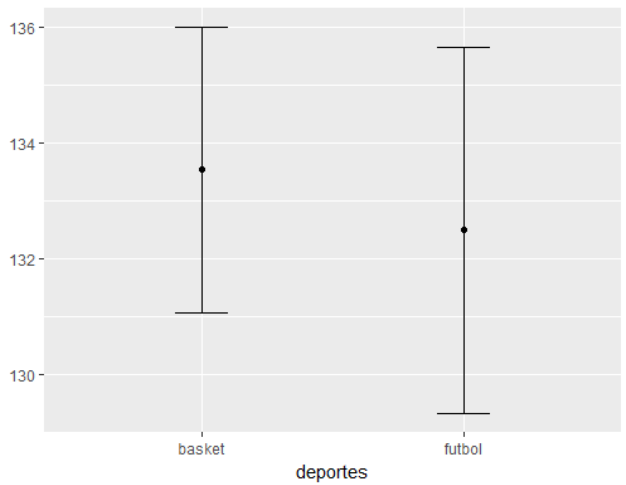
H_1 = la alturas medias no son distintas

```
basket<-rnorm(30, 133, 8) ; futbol<-rnorm(30, 130,10)
kids<-(c(basket,futbol)) ; deportes<-c(rep("basket",30),rep("futbol",30))
alturas<-data.frame(kids, deportes)
```

Exploro mis datos

```
> boxplot(alturas$niños~alturas$deportes)
```

#Comprueba siempre si en tus datos hay valores “raros/extremos”
#esos valores pueden ser verdaderos, es decir, ocurren en la realidad o
#puede ser un error al introducir tus datos. Si es/son un error estás a tiempo
#de corregirlo y evitar problemas futuros.



3. Comparación de la media de grupos: Test de hipótesis **t-Student**

H_0 = VR es normal
H_1 = VR no es normal

H_0 = varianzas de VR son iguales
H_1 = varianzas de VR no son iguales

Compruebo las asunciones del modelo

1) Normalidad de la variable respuesta (VR)

```
> shapiro.test(alturas$kids)
Shapiro-Wilk normality test data:
alturas$kids
W = 0.95322, p-value = 0.9995
```
>0.05

Acepto H_0 => No puedo decir que VR no es normal/varianzas no son iguales.

2) Homocedasticidad de mis variables

```
> leveneTest(alturas$niños, group=deportes)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  1    0.258 0.6134
       58
```
>0.05

```
> t.test(alturas$kids~ alturas$deportes)

Welch Two Sample t-test data: alturas$kids by alturas$deportes

t = 0.61362, df = 57.984, p-value = 0.5419
```
>0.05

```
 alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval: -1.015654 1.913598 sample estimates:

mean in group basket mean in group futbol

                133.2709          132.8219
```

Cuidado con poner el botón automático!!!

Recuerda:

H_0 = la alturas medias no son distintas

H_1 = la alturas medias son distintas

La probabilidad de equivocarnos al rechazar Ho es muy alta => Acepto Ho

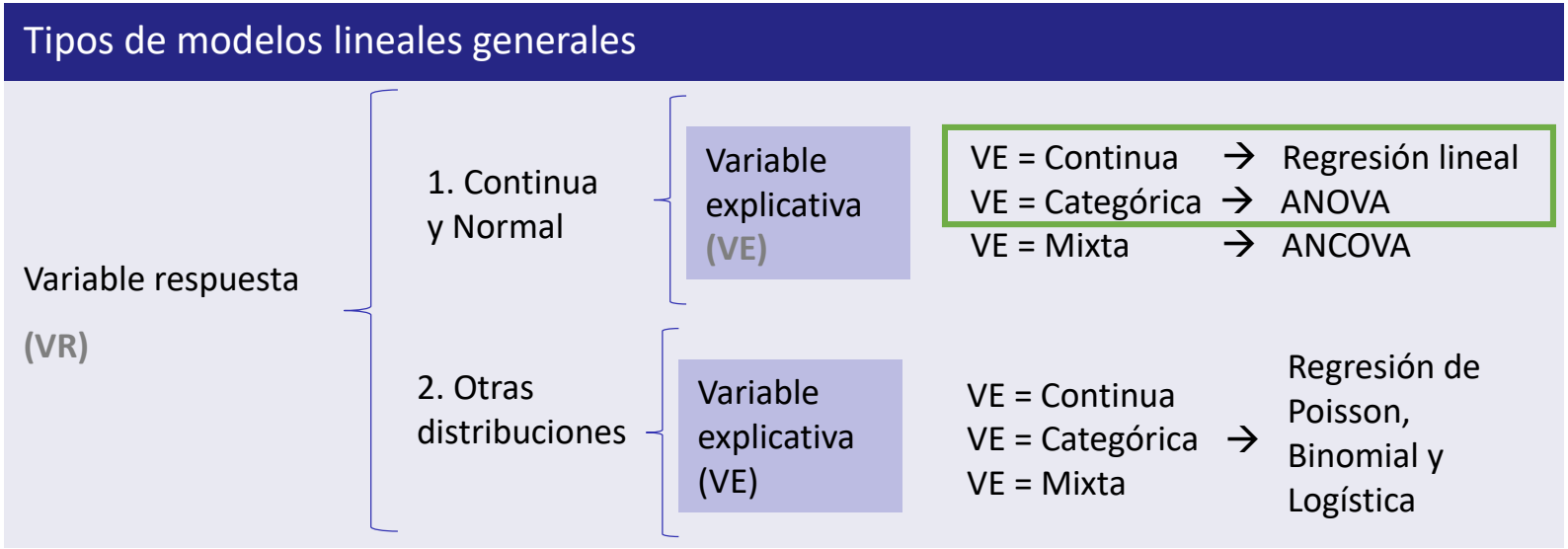
➔ Las alturas medias de niños de 10 años que juegan a baloncesto y fútbol no son distintas.

Secciones

0. ¿Qué vamos o no vamos a hacer en estos 10-15 minutos?
1. Estadística ¿para qué?
2. Mi hipótesis. Errores de tipo I y de tipo II
3. Comparación de la media dos grupos: Test de hipótesis: t-Student.
4. Modelos lineales generales (MLG)
   - Regresión lineal
   - Anova

4. Modelos lineales generales (MLG)

Relación entre una variable respuesta (VR) y una variable explicativa (VE)



4. Regresión lineal

Modelo predictivo en el que conozco ***y*** a través de ***x*.**

Realidad

$y = a + b \cdot x + Error$

Lo desconozco

→

Mi predicción

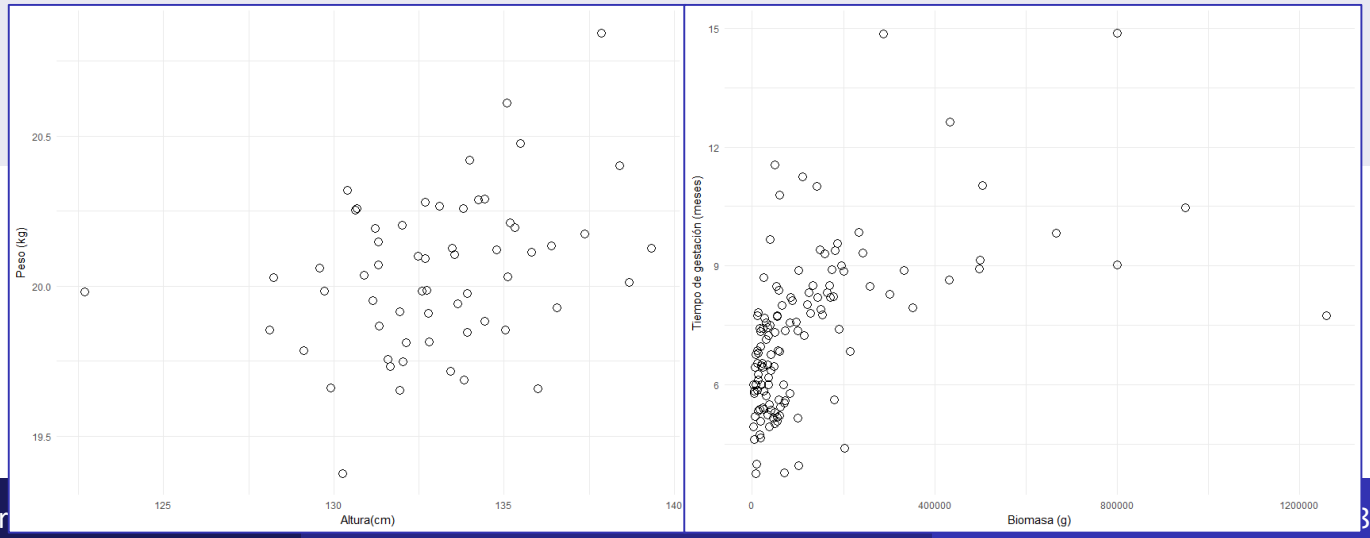
$y' = a + b \cdot x$

Intercepto:
Corte con x=0

Pendiente:
Cada unidad de Δ x
cuánto Δ ó ∇ y

Asunciones de los modelos lineales: hay que comprobarlo **SIEMPRE**

1. Relación lineal entre VR y VE



4. Regresión lineal

Modelo predictivo en el que conozco ***y*** a través de ***x***.

Realidad

$y = a + b \cdot x + Error$

Lo desconozco

Mi predicción

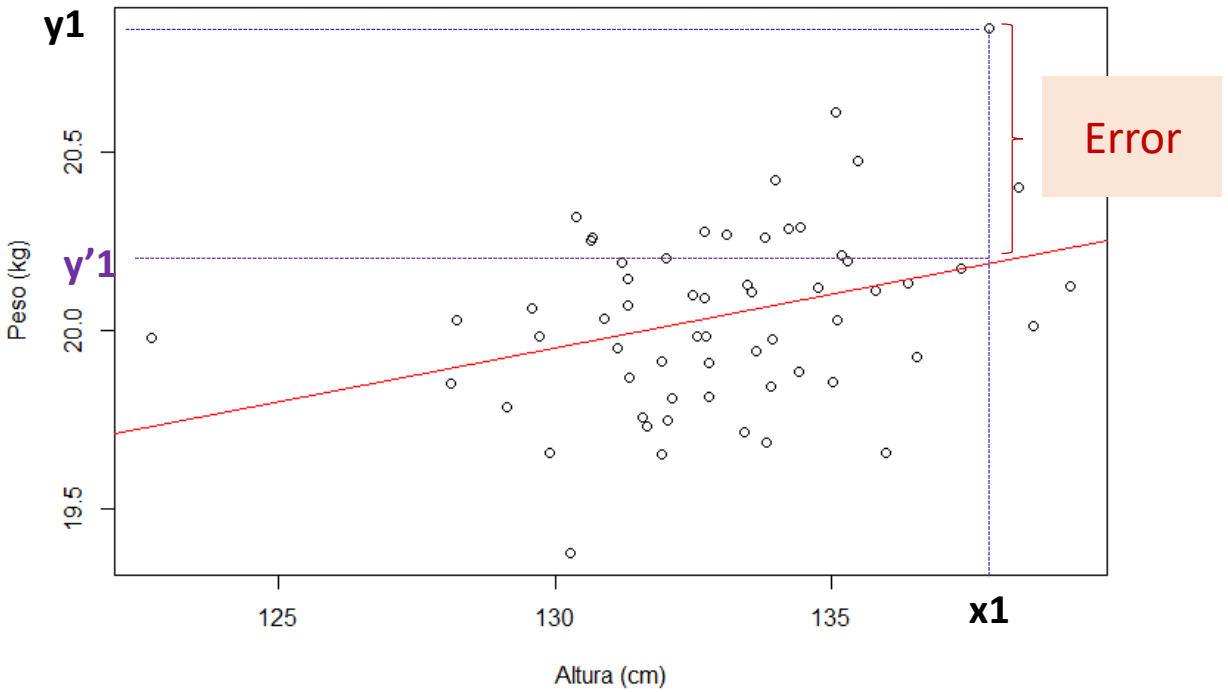
$y' = a + b \cdot x$

Intercepto: Corte con x=0

Pendiente: Cada unidad de Δ x cuánto Δ ó ∇ y

Objetivo de la regresión: mínimas distancias verticales (y) a la recta

Método de mínimos cuadrados (LM).



4. Regresión lineal

Modelo predictivo en el que conozco ***y*** a través de ***x***.

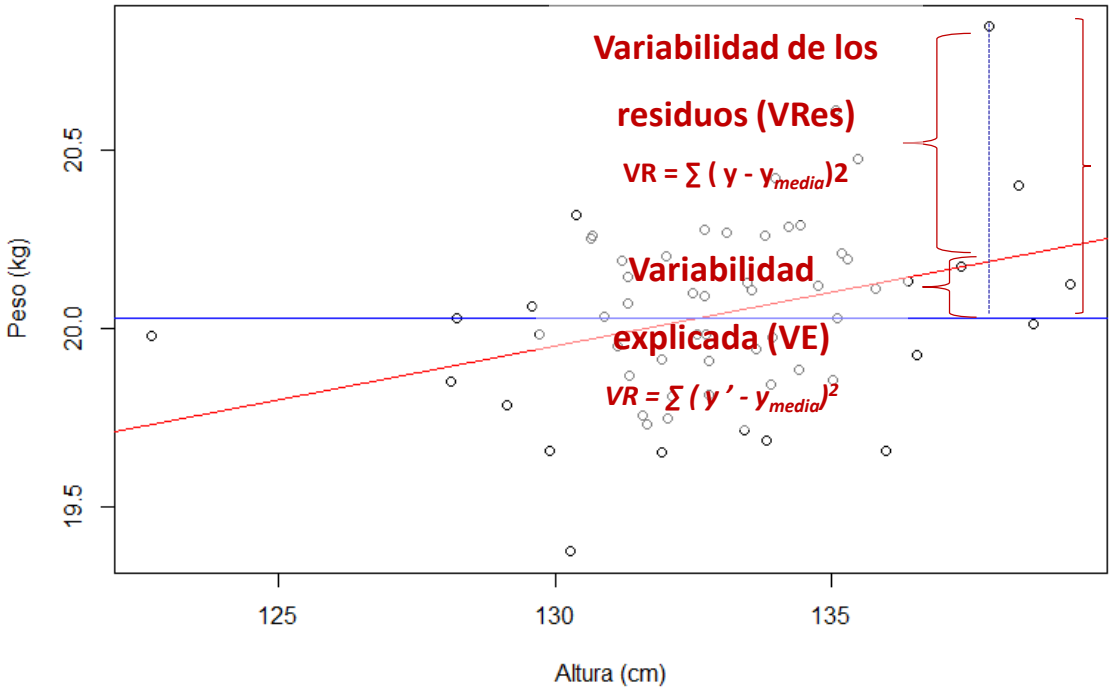
Realidad: $y = a + b \cdot x + Error$ (Lo desconozco)

→

Mi predicción: $y' = a + b \cdot x$ (Intercepto: Corte con x=0; Pendiente: Cada unidad de Δ x cuánto Δ ó ∇ y)

Objetivo de la regresión: mínimas distancias verticales (y) a la recta

Método de mínimos cuadrados (LM).



Variabilidad total (VT)

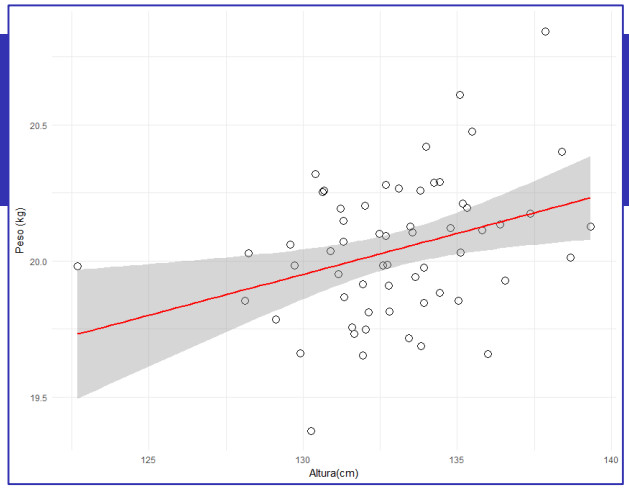
Valor de y – Valor de y_{media}

y_{media}

$$R^2 = \frac{VE}{VT}$$

Muy dependiente del tamaño de la muestra

4. Regresión lineal

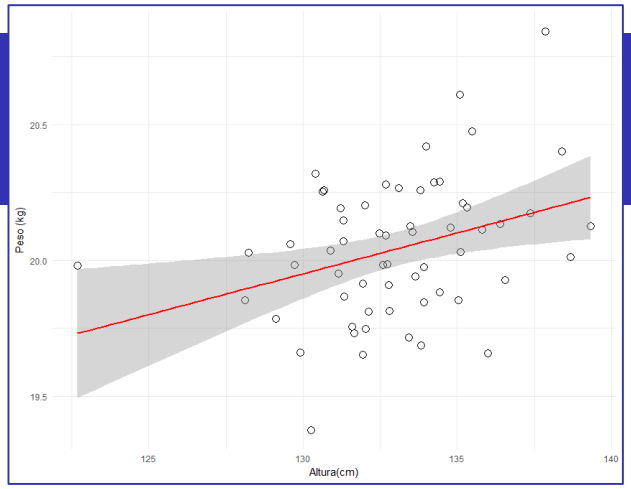



```
>alturas<-read_csv("alturas.csv")
>modelo1<-lm(alturas$peso~alturas$kids)

#compruebo las asunciones de mi modelo
>ggplot(alturas, aes(kids,peso))+
  geom_point(pch=1, size=3.5)+                    #La relación entre VR y VE parece ser lineal
  geom_smooth(method = "lm", col="red")+
  xlab("Altura(cm)")+ylab("Peso (kg)")+theme_minimal()
>qqnorm(residuals(modelo1))                       #Comprobación visual de normalidad. Los residuos se
>qqline(residuals(modelo1))                       #distribuyen de manera normal
>shapiro.test(residuals(modelo1))                 #Comprobación estadística de normalidad

#Igualdad de varianzas: si no hay patrón (el gráfico me da una nube de puntos), ¡genial! hay homocedasticidad
→ Los residuos se distribuyen con igual varianza, eso conlleva que mi modelo vaya a ser igual de bueno o malo
en todos los puntos.
> plot(residuals(modelo1),predict(modelo1))
```

4. Regresión lineal



#se puede ver tanto la comprobación de la normalidad como de la

#homocedasticidad a la vez:

>par(mfrow=c(1,4))

>plot(modelo1)

Nube de puntos. No existe un patrón definido.
→ Hay homocedasticidad

Mis valores siguen más o menos la línea guay, me preocupo cuando se alejen mucho.

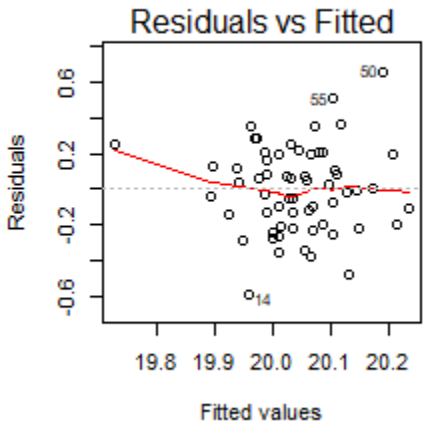
Distancia de Cook: sirve para ver datos aberrantes.
El valor más frecuente de los residuos estandarizados es 0, y no suelen sobrepasar ± 2.

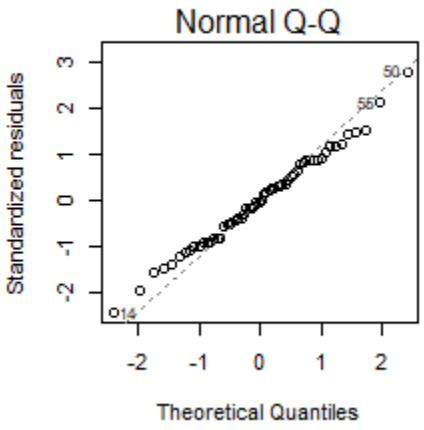
- Si encuentras patrones: no te fíes de las predicciones.

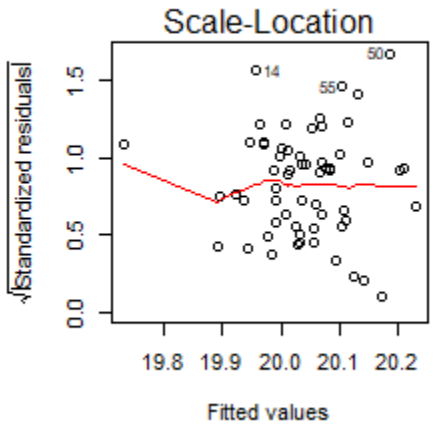
Índice de Leverage (hi)

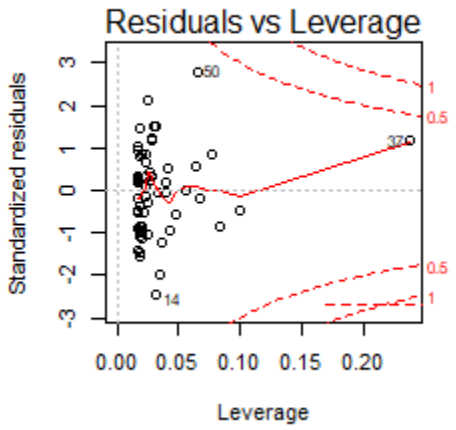
Me preocupo por ser un **punto aberrante en el eje de las X,** si hi > a 2 * p/n

p: nº de parámetros;
b: tamaño de la muestra.









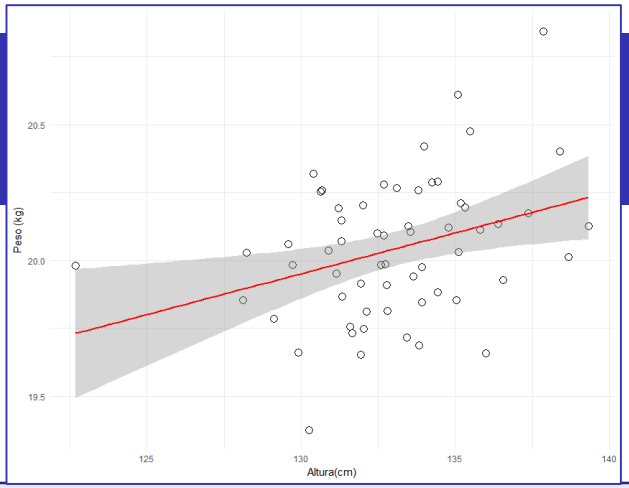
4. Regresión lineal

> anova(modelo1) #comprueba la significación del modelo

```
Analysis of Variance Table Response:
alturas$peso
              Df  Sum Sq  Mean Sq  F value  Pr(>F)
alturas$kids  1   0.4368  0.43677  7.4296   0.008469 **
Residuals     58  3.4097  0.05879
--- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

>summary(modelo1) #resumen del modelo

```
Call: lm(formula = alturas$peso ~ alturas$kids)
Residuals:
    Min      1Q      Median  3Q     Max
    -0.5826 -0.1994 -0.0039 0.1909 0.6566
Coefficients:
              Estimate Std.Error t value  Pr(>|t|)
(Intercept)  16.03683  1.46949   10.913   1.1e-15 ***
alturas$kids 0.03011  0.01105   2.726    0.00847 **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.2425 on 58 degrees of freedom
Multiple R-squared: 0.1136, Adjusted R-squared: 0.09827
F-statistic: 7.43 on 1 and 58 DF, p-value: 0.008469
```



¿Es válido mi modelo?

Un buen modelo debe explicar obligatoriamente mayor varianza que la varianza que explican los residuos.

4. Anova

El Anova es un **modelo predictivo lineal** que sirve para testar la hipótesis de si existen diferencias entre las medias de unas poblaciones debido a una variable explicativa cualitativa.

El resultado del Anova te dice si las medias son o no diferentes, **pero** no te dice cuánto lo son. Hay que hacer pruebas a posteriori (test post-hoc) para ver cómo de diferentes son las medias de los grupos *(e.g.* Test de Bonferroni, test de Tukey*).*

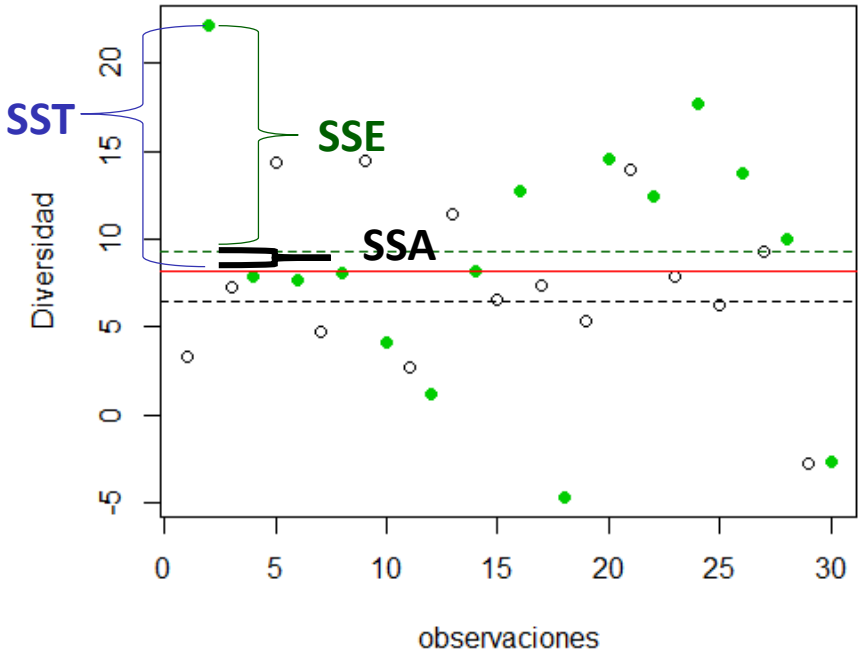
SST: Desviación individual respecto a la media total de los grupos

SSE: Desviación de cada grupo respecto a la media de su grupo.

SSA = SST – SSE

SSA tamaño del efecto de VE

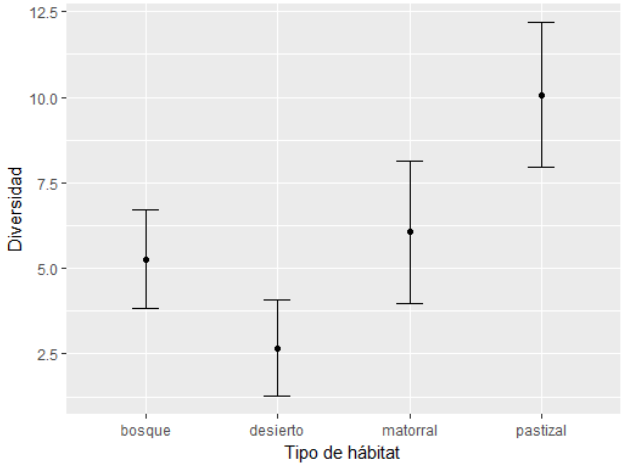
X_{media} **total** X_{media} **pastizal** X_{media} **bosque**



4. Anova

Asunciones de la Anova: hay que comprobarlo **SIEMPRE**
1. Normalidad de la variable respuesta (VR).
2. Homocedasticidad: la varianza de los grupos son iguales.
3. Las muestras son independientes **(→ Muestreo aleatorio).**
4. Independencia de los residuos: cualquier variación que hay en mi VR se debe **al azar** y no a otro factor que esté influyendo sin que nos demos cuenta.

4. Anova




```
>diverso<-read_csv("diverso.csv")
>diverso_summary<-diverso %>%
 group_by(names)%>%
 summarise(mean_h=mean(div),sd_h=sd(div),n_h=n(),se_h=sd(div)/sqrt(n()))
>ggplot(diverso_summary,aes(names,mean_h))+
 geom_point()+geom_errorbar(aes(x=names, ymin=(mean_h-sd_h), ymax=(mean_h+sd_h), width=0.2))+
 ylab("Diversidad")+xlab("Tipo de hábitat")
>shapiro.test(diverso$div)
>leveneTest(diverso$div, diverso$names)
>diver<-lm(diverso$div~diverso$names)
>anova(diver)
```

Analysis of Variance Table
Response: diverso$div

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diverso$names	3	422.85	140.950	43.819	**1.037e-14 *****
Residuals	56	180.13	3.217		

--- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
>pairwise.t.test(x=diverso$div, g=diverso$names,
                 p.adjust.method=c("bonferroni"))
```

Pairwise comparisons using t tests with pooled SD
data: diverso$div and diverso$names

	bosque	desierto	matorral
desierto	0.0012	-	-
matorral	1.0000	1.8e-05	-
pastizal	5.9e-09	2.7e-15	6.1e-07

P value adjustment method: bonferroni

```
>TukeyHSD(aov(diver))
```

Tukey multiple comparisons of means 95%
family-wise confidence level
Fit: aov(formula = diver)
$`diverso$names`

	diff	lwr	upr	p adj
desierto-bosque	-2.6	-4.3340911	-0.8659089	0.0011565
matorral-bosque	0.8	-0.9340911	2.5340911	0.6159418
pastizal-bosque	4.8	3.0659089	6.5340911	0.0000000
matorral-desierto	3.4	1.6659089	5.1340911	0.0000176
pastizal-desierto	7.4	5.6659089	9.1340911	0.0000000
pastizal-matorral	4.0	2.2659089	5.7340911	0.0000006