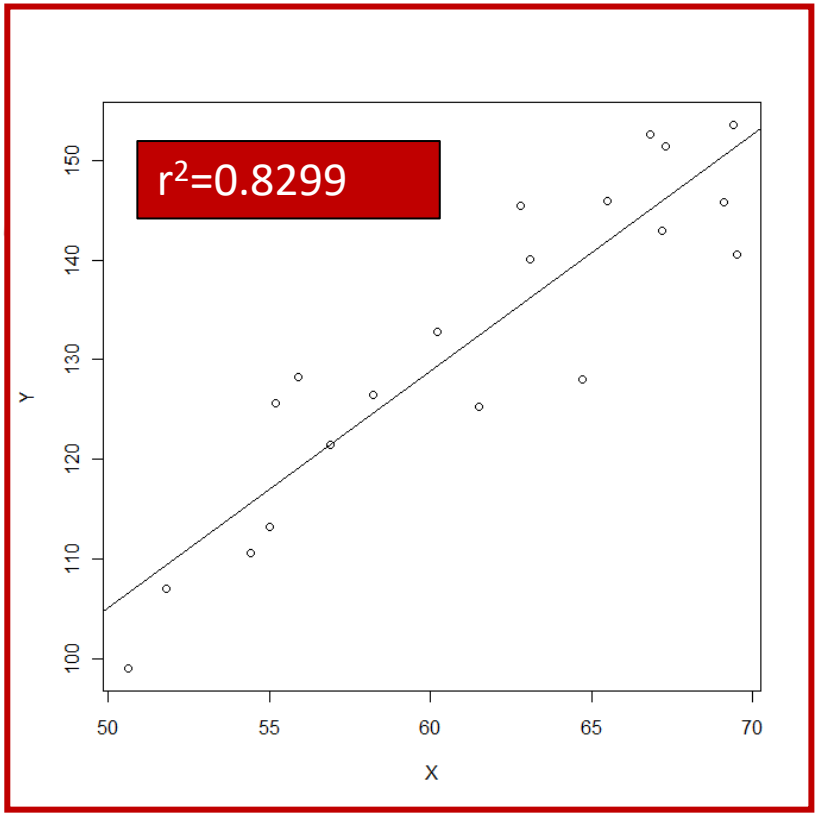
Regresión lineal
Ejemplos

Ejemplo

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.5424    15.5825  -0.869    0.396
X             2.3727     0.2532   9.372  2.4e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 6.784 on 18 degrees of freedom
Multiple R-squared: 0.8299,     Adjusted R-squared: 0.8205
F-statistic: 87.83 on 1 and 18 DF,  p-value: 2.401e-08
```

Adjusted R-squared es una modificación que tiene en cuenta el número de predictores

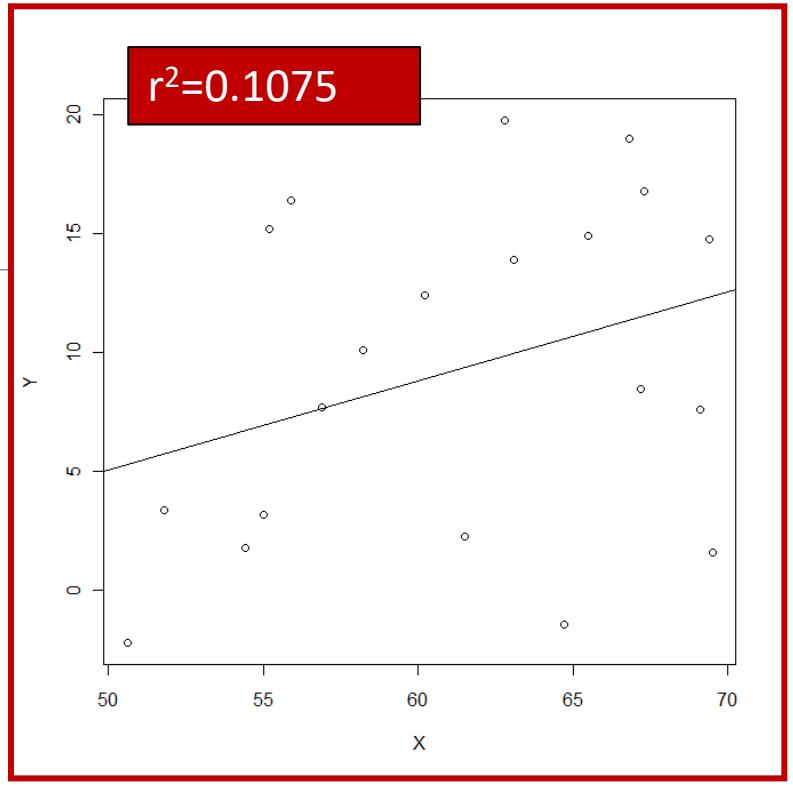
$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$



$$r = \sqrt{R^2} = \sqrt{0.8299} = 0.91$$

Ejemplo

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.5424    15.5825  -0.869    0.396
X             0.3727     0.2532   1.472    0.158

Residual standard error: 6.784 on 18 degrees of freedom
Multiple R-squared: 0.1075,     Adjusted R-squared: 0.05789
F-statistic: 2.168 on 1 and 18 DF,  p-value: 0.1582

> confint(res)
                2.5 %     97.5 %
(Intercept) -46.2800121 19.1951768
X            -0.1591653  0.9046528
> plot(data)
> abline(res)
> detach(data)
```



El IC de la pendiente (coeficiente para X) es (-0.16,0.90). No podemos descartar que la pendiente sea 0 (independencia de los valores de Y respecto a X).

El p-valor evalúa si podemos aceptar que $r^2=0$. En este caso p-valor=0.16, por lo tanto los datos son compatibles con una correlación igual a 0 (independencia entre X e Y)

Beers: Número de cervezas consumidas.
BAL: Blood alcohol level

Ejemplo

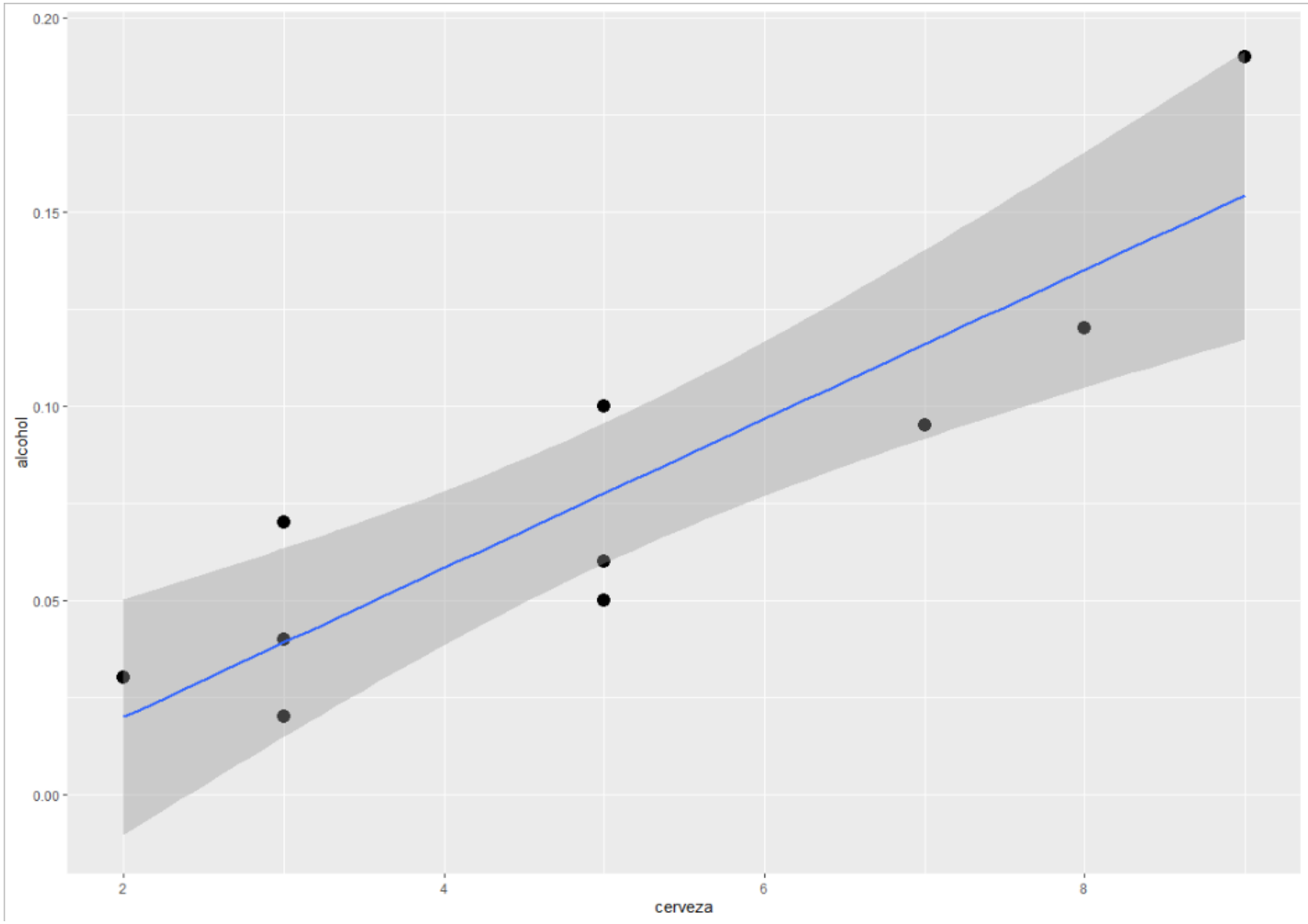
```
beers <- c(5,2,9,8,3,7,3,5,3,5)
BAL <- c(0.1,0.03,0.19,0.12,0.04,0.095,0.07,0.06,0.02,0.05)

data <- data.frame(cerveza = beers, alcohol= BAL)
data
```

	cerveza	alcohol
1	5	0.100
2	2	0.030
3	9	0.190
4	8	0.120
5	3	0.040
6	7	0.095
7	3	0.070
8	5	0.060
9	3	0.020
10	5	0.050

```
res <- lm(alcohol~cerveza,data=data)
summary(res)

Call:
lm(formula = alcohol ~ cerveza, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.0275 -0.0187 -0.0071  0.0194  0.0357

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.018500   0.019230  -0.962 0.364200
cerveza      0.019200   0.003511   5.469 0.000595 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02483 on 8 degrees of freedom
Multiple R-squared:  0.789,     Adjusted R-squared:  0.7626
F-statistic: 29.91 on 1 and 8 DF,  p-value: 0.0005953
```

IC 95% para la media de alcohol en sangre en función del número de cervezas consumidas

```
ggplot(data,aes(x=cerveza,y=alcohol))+
  geom_point(size=4)+
  geom_smooth(method=lm)
```



Intervalos de predicción para los valores esperados de alcohol en sangre en función del número de cervezas consumidas

```
tc <- predict(res,newdata = data.frame(cerveza=1:10),int='conf')
tc
tc<- tc %>% as.data.frame %>% mutate(tomas=c(1:10))
tc %>% round(2)
```

	fit	lwr	upr
	0.0007	-0.03640097	0.03780097
	0.0199	-0.01039281	0.05019281
	0.0391	0.01481172	0.06338828
	0.0583	0.03846870	0.07813130
	0.0775	0.05939658	0.09560342
	0.0967	0.07686870	0.11653130
	0.1159	0.09161172	0.14018828
	0.1351	0.10480719	0.16539281
	0.1543	0.11719903	0.19140097
	0.1735	0.12915586	0.21784414

	fit	lwr	upr	tomas
	0.00	-0.04	0.04	1
	0.02	-0.01	0.05	2
	0.04	0.01	0.06	3
	0.06	0.04	0.08	4
	0.08	0.06	0.10	5
	0.10	0.08	0.12	6
	0.12	0.09	0.14	7
	0.14	0.10	0.17	8
	0.15	0.12	0.19	9
	0.17	0.13	0.22	10

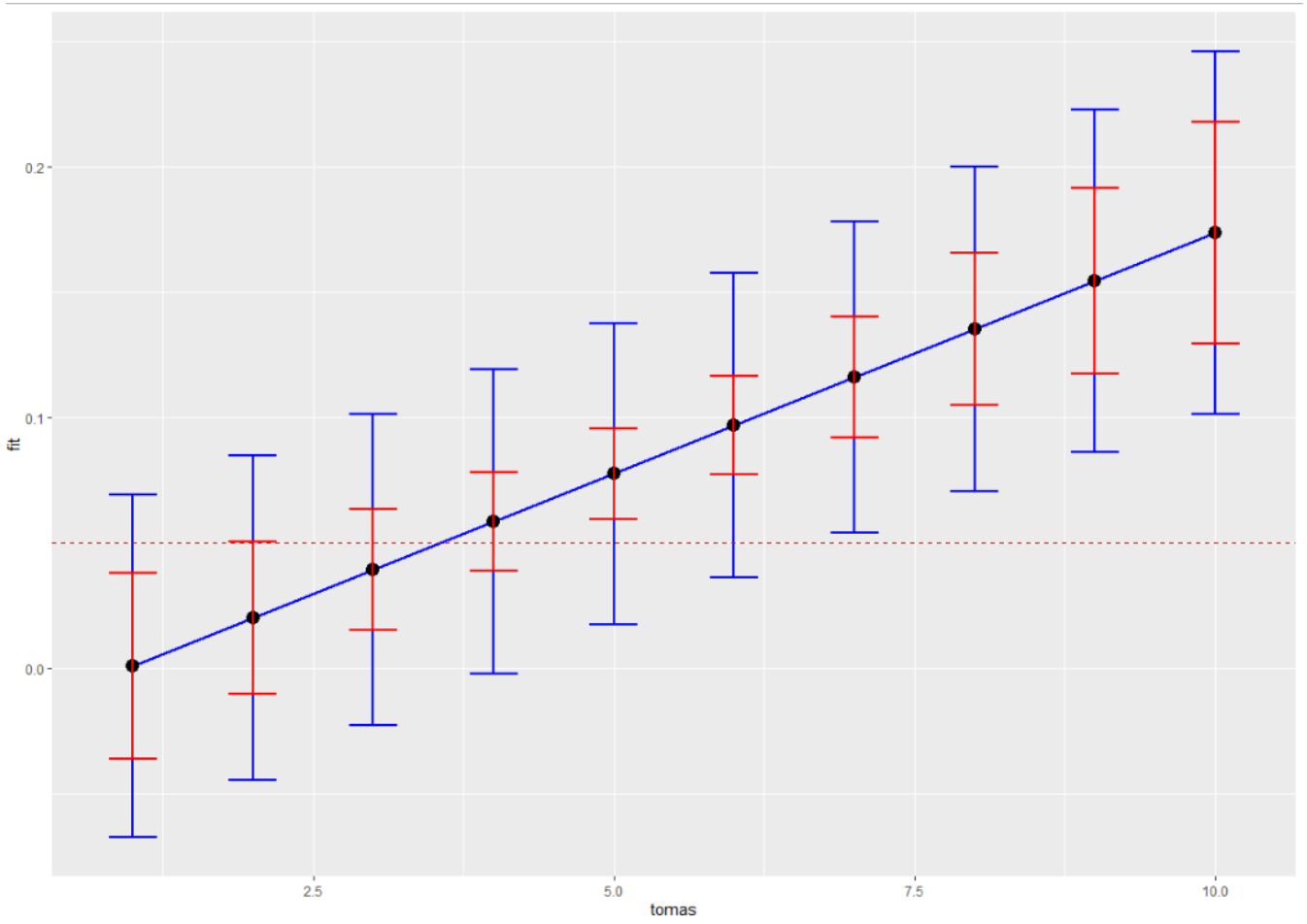
Intervalos de predicción para los valores de alcohol en sangre para una persona en función del número de cervezas consumidas

```
tp <- predict(res,newdata = data.frame(cerveza=1:10),int='prediction')
tp
tp<- tp %>% as.data.frame %>% mutate(tomas=c(1:10))
tp %>% round(2)
```

```
       fit          lwr         upr
0.0007 -0.067518905 0.06891890
0.0199 -0.044868756 0.08466876
0.0391 -0.023087282 0.10128728
0.0583 -0.002285623 0.11888562
0.0775  0.017457756 0.13754224
0.0967  0.036114377 0.15728562
0.1159  0.053712718 0.17808728
0.1351  0.070331244 0.19986876
0.1543  0.086081095 0.22251890
0.1735  0.101086330 0.24591367
```

```
   fit   lwr  upr tomas
0.00 -0.07 0.07     1
0.02 -0.04 0.08     2
0.04 -0.02 0.10     3
0.06  0.00 0.12     4
0.08  0.02 0.14     5
0.10  0.04 0.16     6
0.12  0.05 0.18     7
0.14  0.07 0.20     8
0.15  0.09 0.22     9
0.17  0.10 0.25    10
```

Intervalos de confianza para la media de alcohol en sangre (rojo) y intervalos de predicción para el valor de alcohol en sangre en un idividuo (azul) en función del número de cervezas consumidas.



Ejemplo: Relación altura vs. peso

```
> head(Davis)
  sex weight height repwt repht
1   M     77    182    77   180
2   F     58    161    51   159
3   F     53    161    54   158
4   M     68    177    70   175
5   F     59    157    59   155
6   M     76    170    76   165
```

```
res <- lm(weight~height,data=Davis)
summary(res)
```

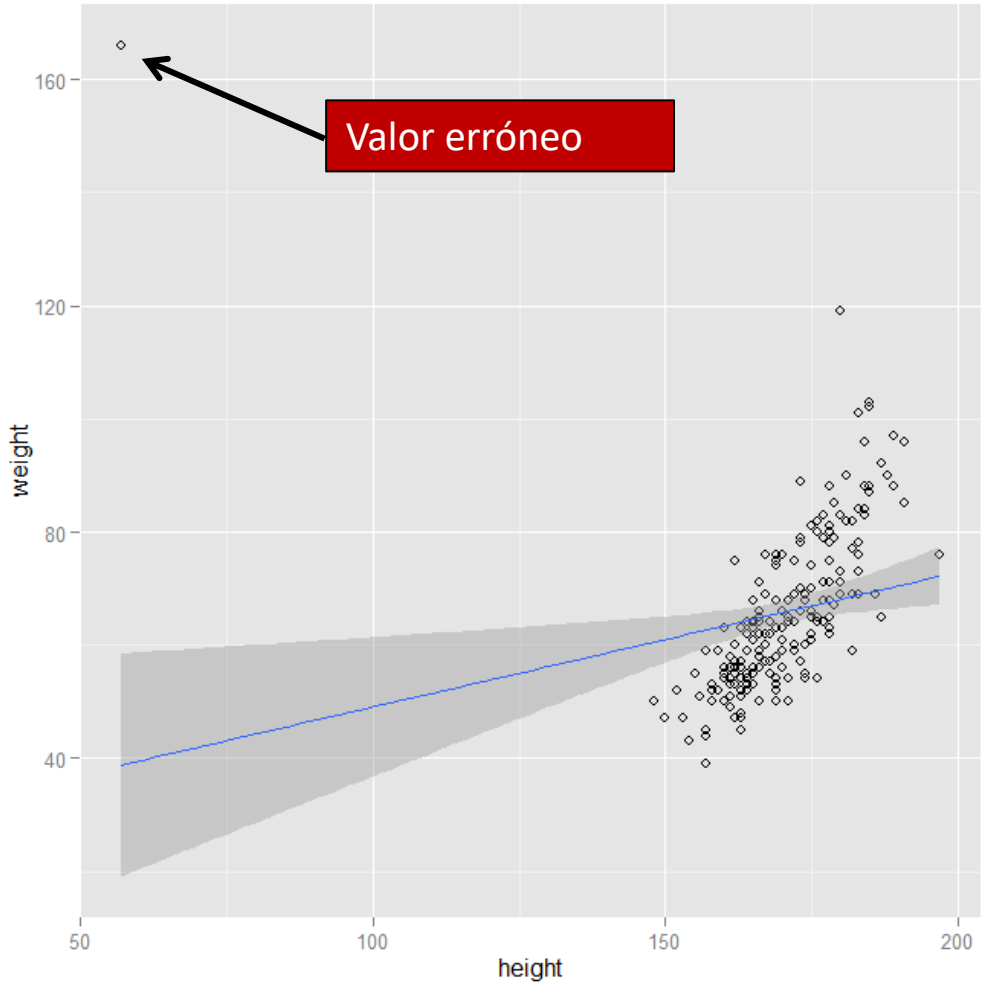
```
Call:
lm(formula = weight ~ height, data = Davis)

Residuals:
    Min      1Q  Median      3Q     Max
-23.696  -9.506  -2.818   6.372 127.145

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.26623   14.95042   1.690  0.09260 .
height       0.23841    0.08772   2.718  0.00715 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.86 on 198 degrees of freedom
Multiple R-squared:  0.03597,   Adjusted R-squared:  0.0311
F-statistic: 7.387 on 1 and 198 DF,  p-value: 0.007152
```

Los resultados indican una R^2=0.03, lo que implicaría que el peso es prácticamente independiente de la altura.

Intuitivamente, este resultado no es muy lógico.

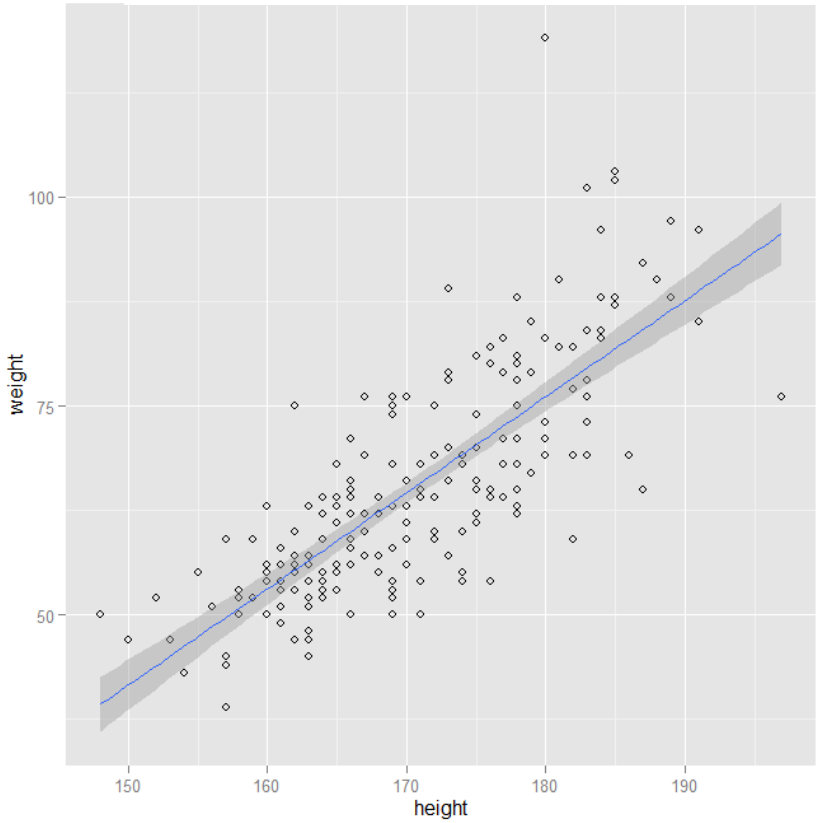
Explorar los datos!!!

El valor extremo determina que la recta estimada se desplace hacia arriba y no ajuste adecuadamente el resto de observaciones.



Efecto de valores extremos

```
data <- Davis %>% filter(height>100)
ggplot(data,aes(x=height,y=weight))+
  geom_point()+
  geom_smooth(method=lm)
```




```
res <- lm(weight~height,data=data)
summary(res)

Call:
lm(formula = weight ~ height, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-19.650  -5.419  -0.576   4.857  42.887

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -130.74698   11.56271  -11.31   <2e-16 ***
height         1.14922    0.06769   16.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.523 on 197 degrees of freedom
Multiple R-squared:  0.594,     Adjusted R-squared:  0.592
F-statistic: 288.3 on 1 and 197 DF,  p-value: < 2.2e-16
```

Intervalos de referencia para un individuo y IC para la media

```
alturas <- seq(160,190,5)
res.pred <- data.frame(
  alturas,
  predict(res,newdata = d,int='pred'))
res.pred %>% round(2)
```

alturas	fit	lwr	upr
160	53.13	36.22	70.04
165	58.87	42.01	75.74
170	64.62	47.77	81.47
175	70.37	53.51	87.23
180	76.11	59.22	93.01
185	81.86	64.90	98.82
190	87.61	70.56	104.65

El valor predicho de peso para una persona de 170 cm de altura es de 64.6 kg, con un intervalo entre 47.8 y 81.5.

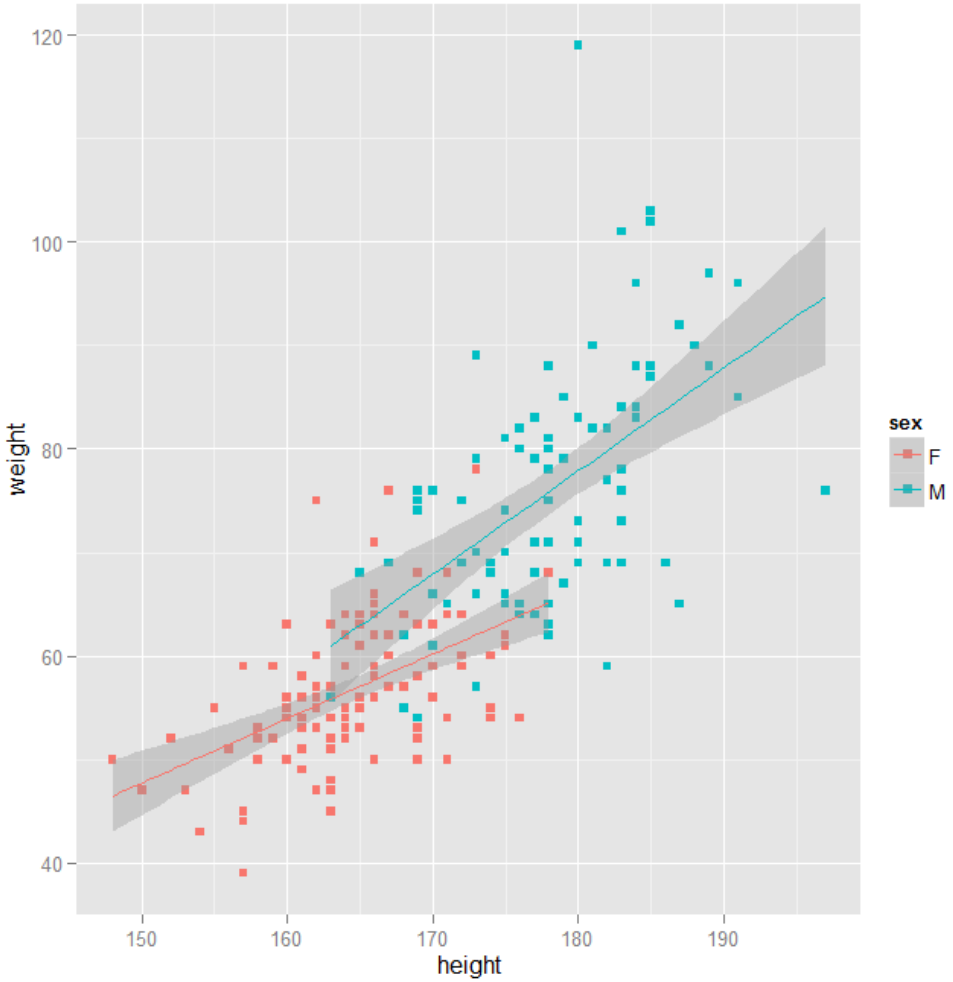
```
res.conf<- data.frame(
  alturas,
  predict(res,newdata = d,int='conf'))
res.conf %>% round(2)
```

alturas	fit	lwr	upr
160	53.13	51.28	54.98
165	58.87	57.47	60.28
170	64.62	63.43	65.81
175	70.37	69.04	71.70
180	76.11	74.38	77.84
185	81.86	79.60	84.12
190	87.61	84.75	90.46

Con una confianza del 95%, el valor medio de peso de las personas de 170 cm de altura se sitúa entre 69.04 y 65.81.

Relación altura-peso según el sexo de la persona

```
ggplot(data,
       aes(x=height,y=weight,group=sex,
           colour=sex))+
  geom_point(aes(colour=sex),shape=15)+
  geom_smooth(aes(colour=sex),method=lm)
```



Las mujeres muestran valores de altura y peso inferiores y la relación entre ambas variables parece tener una pendiente menor.

En este caso, hay que ajustar un modelo que incluya el sexo.

Relación altura-peso según el sexo de la persona

```
res.1 <- lm(weight~height,data=data)
res.2 <- lm(weight~height*sex,data=data)
summary(res.2)
anova(res.1,res.2)
```

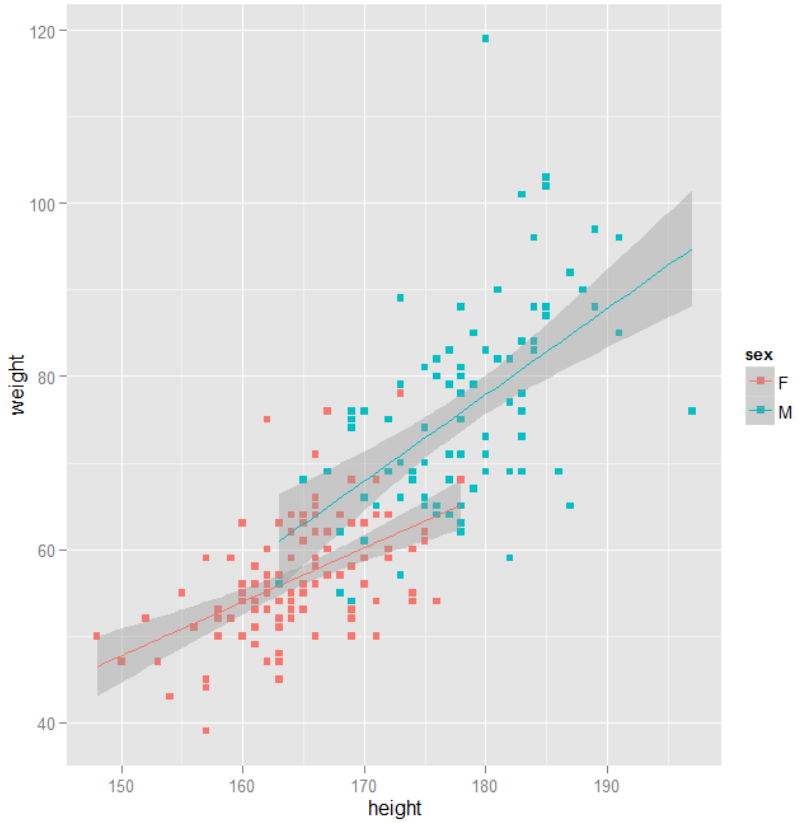
```
Call:
lm(formula = weight ~ height, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-19.650  -5.419  -0.576   4.857  42.887

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -130.74698   11.56271  -11.31   <2e-16 ***
height         1.14922    0.06769   16.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.523 on 197 degrees of freedom
Multiple R-squared:  0.594,     Adjusted R-squared:  0.592
F-statistic: 288.3 on 1 and 197 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = weight ~ height * sex, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.869  -4.848  -0.908   4.546  41.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.7084    22.1942  -2.059   0.0408 *
height        0.6229     0.1347   4.626  6.8e-06 ***
sexM        -55.6216    32.5447  -1.709   0.0890 .
height:sexM   0.3727     0.1897   1.964   0.0509 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.028 on 195 degrees of freedom
Multiple R-squared:  0.6435,    Adjusted R-squared:  0.6381
F-statistic: 117.3 on 3 and 195 DF,  p-value: < 2.2e-16
```

anova(res.1,res.2) compara los resultados de los modelos y evalúa si la introducción de más parámetros asociada a considerar más variables produce una reducción significativa en la suma de cuadrados residual. Si el p-valor es muy pequeño, es una evidencia a favor del modelo con más variables. En caso contrario, el modelo más simple seria suficiente. En este caso, nos quedaríamos con el modelo que considera el sexo y su interacción con la altura.

```
Analysis of Variance Table

Model 1: weight ~ height
Model 2: weight ~ height * sex
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    197 14312
2    195 12567  2    1745.4 13.542 3.112e-06 ***
```

Interpretación de los parámetros del modelo




```
Call:
lm(formula = weight ~ height * sex, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.869  -4.848  -0.908   4.546  41.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -45.7084    22.1942  -2.059   0.0408 *
height        0.6229     0.1347   4.626  6.8e-06 ***
sexM        -55.6216    32.5447  -1.709   0.0890 .
height:sexM   0.3727     0.1897   1.964   0.0509 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.028 on 195 degrees of freedom
Multiple R-squared:  0.6435,    Adjusted R-squared:  0.6381
F-statistic: 117.3 on 3 and 195 DF,  p-value: < 2.2e-16
```

Mujeres (F)

$$\mu_i = -45.71 + 0.62 \times height$$

Hombres (M)

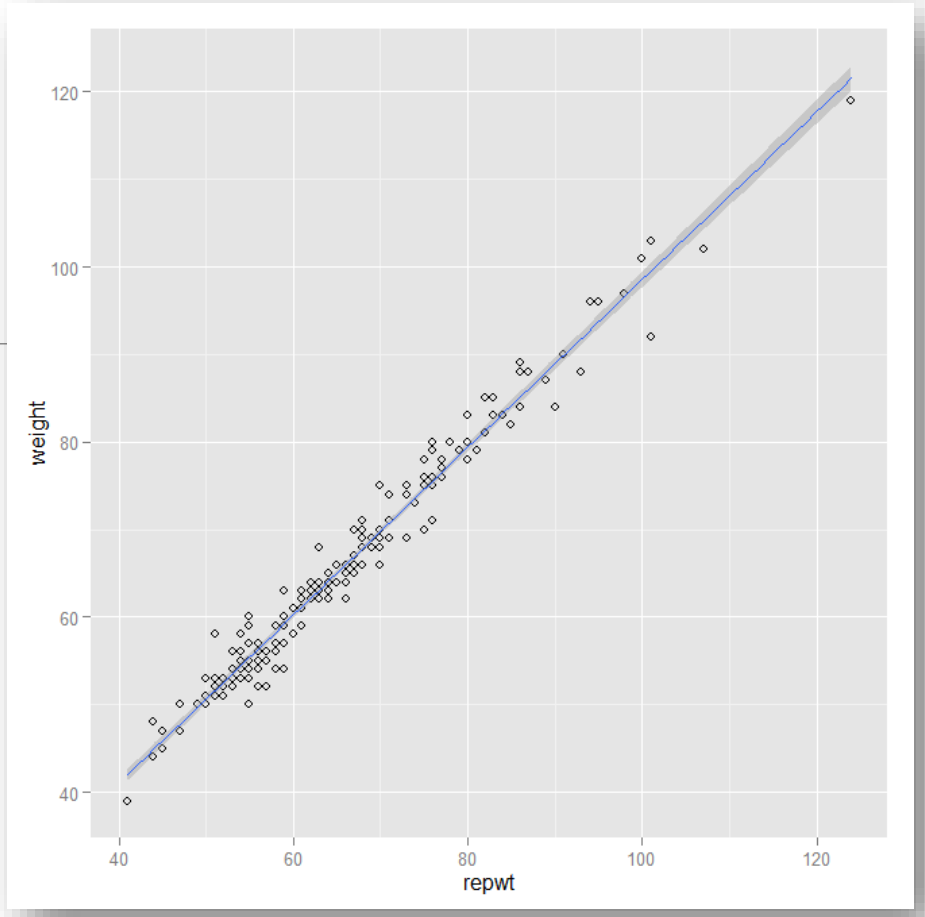
$$\mu_i = (-45.71 - 55.62) + (0.62 + 0.37) \times height$$

$$\mu_i = -101.33 + 0.99 \times height$$

Relación entre el peso real (weight) y el peso indicado por cada persona (perwt).

Utilizamos el fichero donde hemos eliminado el valor extremo.

```
res.peso <- lm(repwt~weight,data=data)
summary(res.peso)

Call:
lm(formula = repwt ~ weight, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-6.8915 -1.1000  0.0661  1.1049  8.6281

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.92797    0.86120  -1.078    0.283
weight       1.01413    0.01285  78.926   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.318 on 180 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.9719,    Adjusted R-squared:  0.9718
F-statistic:  6229 on 1 and 180 DF,  p-value: < 2.2e-16
```



Resumen

El procedimiento de regresión lineal se utiliza para estimar la relación (lineal) entre dos variables cuantitativas.

- La varianza de la variable dependiente debe ser constante para los distintos valores de la variable independiente.
- La variable independiente está controlada por el experimentador.

Análisis típico

- Ajustar la recta de regresión y los IC de los parámetros.
- Obtener el valor de r^2.
- Obtener los IC de predicción para cada valor de la variable independiente.