6. Uso del modelo de regresión lineal simple.

A continuación se presentan los datos de los pesos de los huevos de 11 nidadas de pingüinos Macaroni, cada nidada tiene dos huevos, uno más pequeño (x) que el otro (y).

x	79	93	100	105	101	96	96	109	70	71	87
y	123	138	154	161	155	149	152	160	117	123	138

I. Ajuste del modelo de regresión.

Ajustaremos una recta de regresión para estimar el peso promedio del huevo mayor (y) dado el peso del huevo menor (x), es decir, la variable dependiente es el peso del huevo más grande y_i y la variable independiente es el peso del huevo menor.

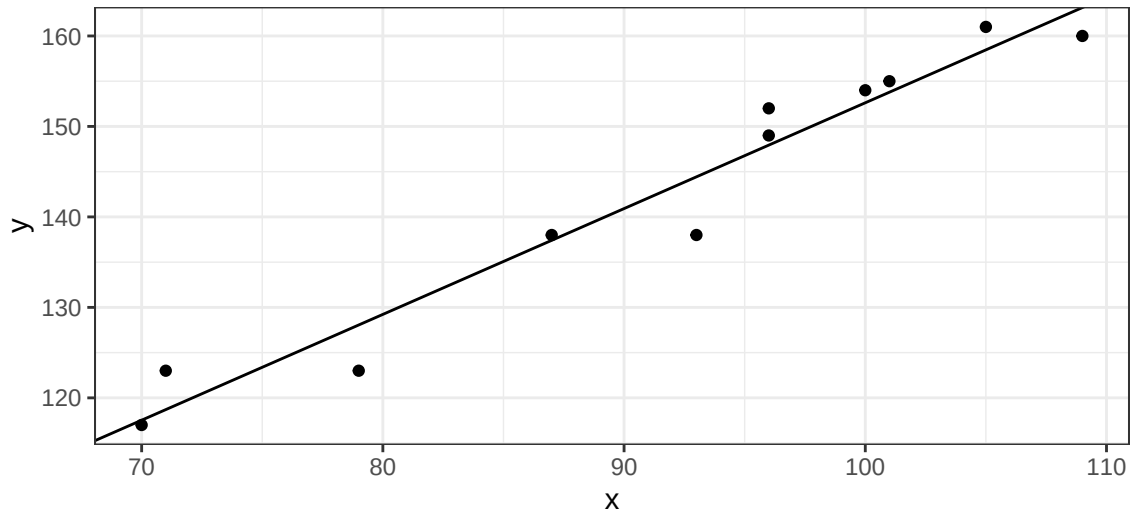
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

En el siguiente Cuadro podemos observar que el p-valor asociado a la prueba F es de menor a 0.05, por lo que se rechaza la hipótesis nula de que los coeficientes asociados a las variables explicativas son cero contra la alternativa de que al menos un coeficiente estimado es distinto de cero. En este caso, como hay una solo variable explicativa, esta prueba coincide con la prueba $t - student$ individual para la β_1= 1.1693983, que también rechaza la hipótesis nula de que $\beta_1 = 0$ contra la alternativa de que $\beta_1 \neq 0$.

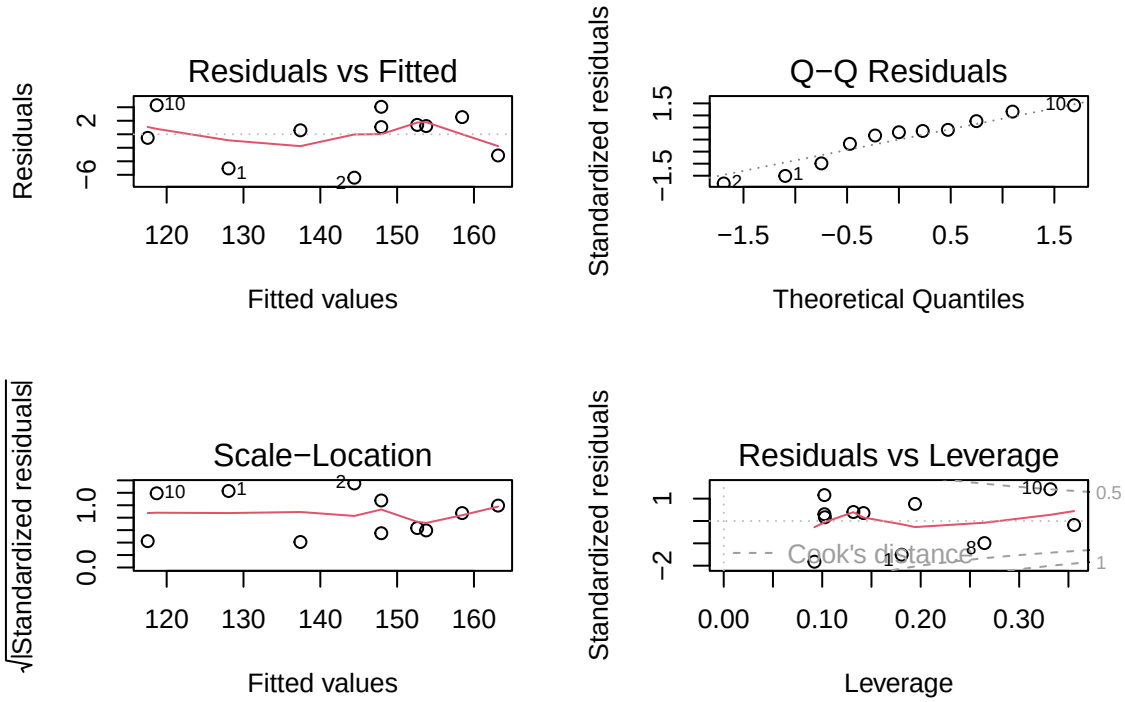
Table 2:

	Dependent variable:
	y
x	1.169***
	s.e.(0.088)
	t-value: 13.225
	Pr(>|t|): 3.35e-07
Constant	35.674***
	s.e.(8.171)
	t-value: 4.366
	Pr(>|t|): 0.00181
Observations	11
R^2	0.951
Adjusted R^2	0.946
Residual Std. Error	3.702 (df = 9)
F Statistic	174.895*** (df = 1; 9); p-value: 3.351e-07
Note:	*p<0.1; **p<0.05; ***p<0.01

La Gráfica siguiente muestra las observaciones y la recta ajustada, parece ser que hay un muy buen ajuste de la recta a los puntos.



En las siguientes Gráficas podemos observar las pruebas gráficas para el cumplimiento de los supuestos del modelo de regresión lineal. La Gráfica **Residuals vs Fitted**, se utiliza para comprobar los supuestos de relación lineal, una línea horizontal, sin patrones distintos, es indicación de una relación lineal, lo que es bueno en nuestro caso. La Gráfica **Normal Q-Q**, se utiliza para examinar si los residuos se distribuyen normalmente, es bueno que los puntos residuales sigan la línea recta discontinua, en nuestro caso parece que todo se ajusta bien. La Gráfica **Scale-Location**, se utiliza para comprobar la homogeneidad de la varianza de los residuos (homoscedasticidad), la línea horizontal con puntos igualmente distribuidos es una buena indicación de homocedasticidad, este es el caso en nuestro ejemplo, donde no tenemos un problema de heterocedasticidad. La Gráfica **Residuals vs Leverage**, se utiliza para identificar casos de valores influyentes, es decir, valores extremos que podrían influir en los resultados de la regresión cuando se incluyen o excluyen del análisis, al parecer ningún valor sale de la distancia de Cook.



En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk, Breusch-Pagan y Durbin-Watson, en todos los casos el p-value asociado es mayor a 0.05, por lo que no hay evidencia para rechzar las hipótesis nulas de normalidad, homoscedasticidad y no autocorrelación, respectivamente.

	1
Normality (Shapiro-Wilk)	0.291
Homoscedasticity (Breusch-Pagan)	0.594
Autocorrelation of residuals (Durbin-Watson)	0.107

II. Prueba de hipótesis. Diferencia entre peso mayor y menor como constante.

Ante la sospecha de que en promedio la diferencia entre el peso mayor y el peso menor es constante (es decir, no depende del peso del huevo menor observado), haremos una prueba de hipótesis. En primer lugar notemos que esto implicaría $H_0 : \beta_1 = 0$, contra la alternativa $\beta_1 \neq 0$, y notamos en el Cuadro anterior de los coeficientes estimados del modelo, que se rechaza $H_0 : \beta_1 = 0$, con la prueba t - student, a favor de la hipótesis alternativa $H_a : \beta_1 \neq 0$. Esto mismo es posible observarlo con el aprueba `Simultaneous Tests for General Linear Hypotheses` que se muestra a continuación, donde se rechaza esta hipótesis nula.

Simultaneous Tests for General Linear Hypotheses				
Fit: lm(formula = Y ∼x, data = Datos6)				
Linear Hypotheses:				
	Estimate	Std. Error	t value	Pr(<t)
1 == 0	1.16940	0.08842	-13.22	3.35e-07*
Signif. codes:	0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			
(Adjusted p values reported – single-step method)				

III. Nueva observación (nidada). ¿Los huevos provienes de pinguinos Macaroni?

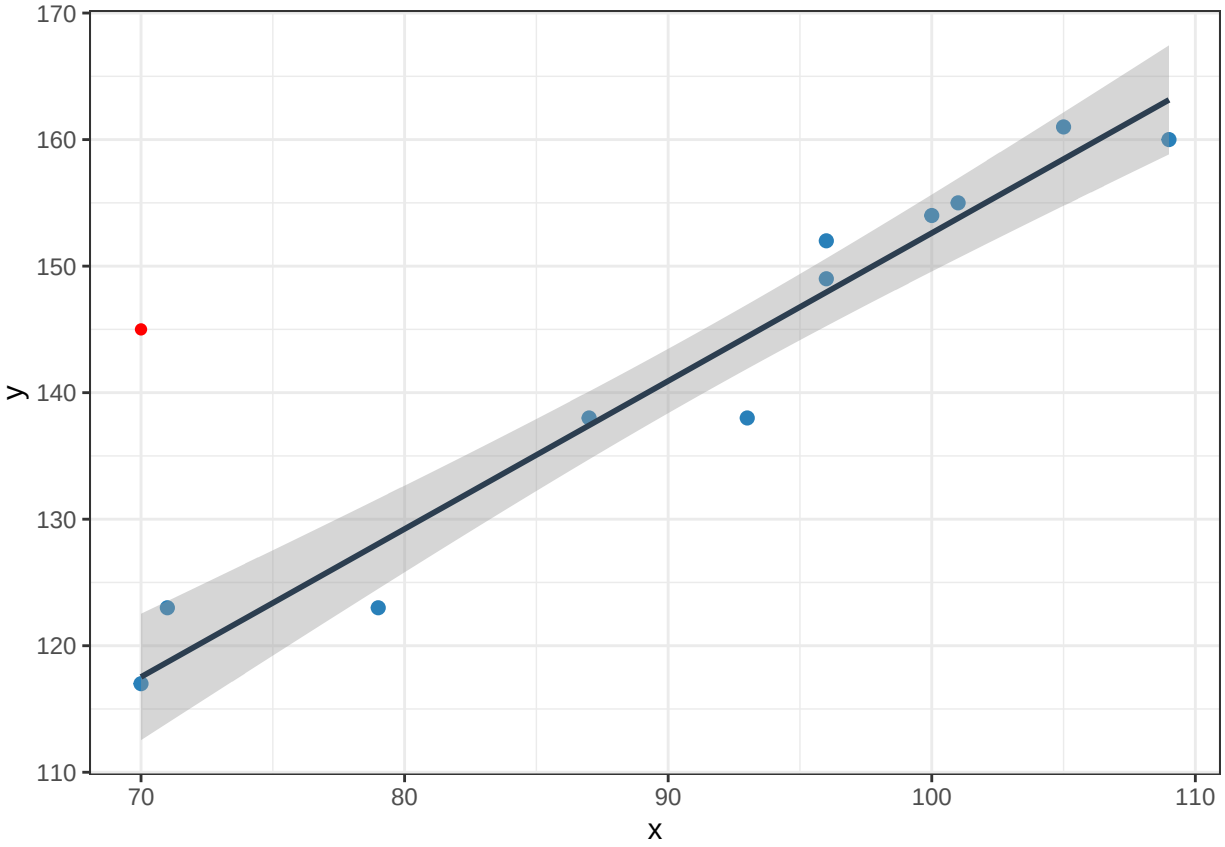
Se observa el peso de los huevos de una nueva nidada, observándose un peso de 70 y 145 gramos. Usando un intervalo de confianza del 95%, veremos si la nidada de huevos sí proviene de pinguinos Macaroni.

Si tomamos en cuenta la recta de regresión anterior, podemos ver que $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i = 35.6741701 + 1.1693983 * (70) = 117.5320539$. Éste valor se ve alejado de los 145 gramos del huevo más grande encontrado. En el siguiente Cuadro se muestran los intervalos de confianza de los parámetros al 95% de confianza.

Table 4:

Constante (β_0)		Pendiente (β_1)	
2.5 %	97.5%	2.5 %	97.5 %
17.2	54.2	0.969	1.369

Al 95% de confianza, podemos predecir que el valor del huevo más grande debería estar entre el valor 112.5371035 y el valor 122.5270044 de acuerdo con un intervalo de confianza al rededor de la recta de regresión. El valor de 145 gramos no cae dentro del intervalo, por lo que podríamos concluir que la nueva observación no corresponde a los huevos de los pingüinos Macaroni, de acuerdo con el modelo planteado.En la siguiente Gráfica podemos observar este intervalo al rededor de la recta de regresión $\hat{y}_h \pm t_{\alpha/2,n-2}(s.e.)_y$ donde $(s.e.)_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-2}}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$.


170
160
150
140
130
120
110
y
70
80
90
100
110
x