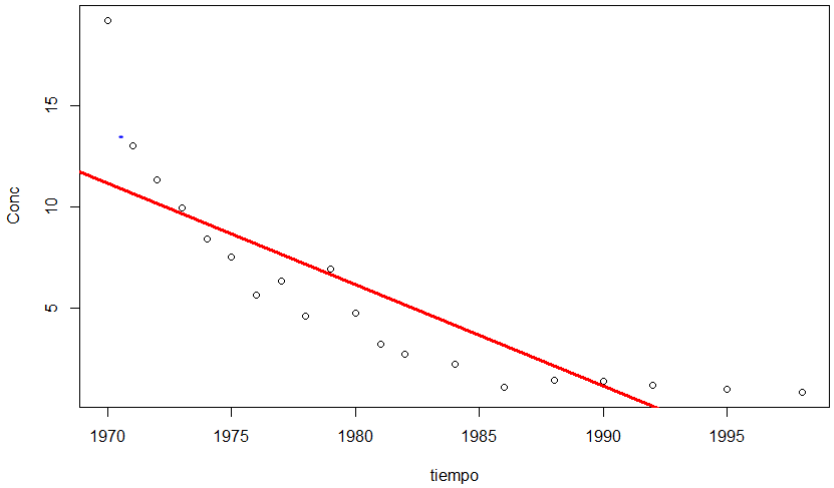
**Tarea 2:** Regresión Simple

EJERCICIO 1: Trabajando a partir del archivo Michigan.txt

1) Para estimar un modelo de regresión hay que aplicar lo aprendido en la tarea anterior y estimar la ecuación de regresión a partir de R con las ordenes modelo=lm( Conc ~ tiempo), y summary(modelo) para obtener la información de dicha ecuación.

```
Call:
lm(formula = Conc ~ tiempo)

Residuals:
   Min     1Q Median     3Q    Max
-2.583 -1.967 -0.730  1.043  8.020

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 997.94410  149.23057   6.687 2.85e-06 ***
tiempo       -0.50090    0.07533  -6.650 3.06e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.692 on 18 degrees of freedom
Multiple R-squared:  0.7107,    Adjusted R-squared:  0.6946
F-statistic: 44.22 on 1 and 18 DF,  p-value: 3.064e-06
```

Dando con esto la siguiente ecuación: y=997,94-0,501x , siendo y la concentración de DDT y x el tiempo. El valor estimado de beta0 es 997,94 que significa que el modelo no pasa por el origen, y el valor estimado de beta1 es -0,501 que da una relación inversa entre la variable concentración y la variable tiempo, al tener el – y por tanto tener la ecuación pendiente negativa. Al haber relación inversa entre las variables un amento del tiempo significará una disminución de la concentración. Todas estas observaciones se ven en la siguiente función gracias a las ordenes plot(tiempo, Conc) y abline(modelo, y las propiedades que queramos en la recta).

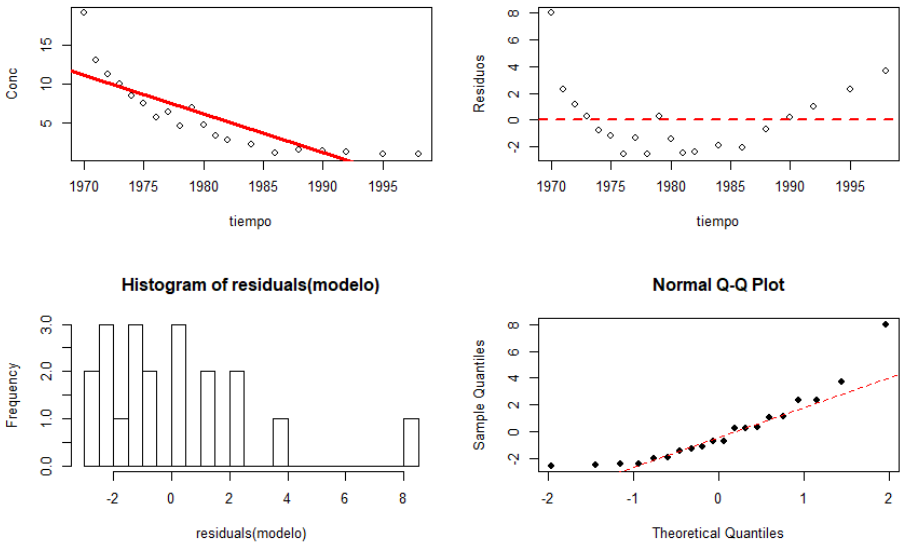



2) La bondad del ajuste ($R^2$) coincide con el coeficiente de determinación, en este caso da una estimación de esta de un 0.7101, lo que significa que un 71,01% de la variabilidad de la concentración está relacionada con el tiempo.
Además en ambos casos, beta0 y beta1, se rechaza la hipótesis inicial (H0=0) ya que en ambas estimaciones el p-valor es muy inferior a cualquier valor de Alpha que se pueda tomar. Esto significa que la ecuación estimada es válida y que ambas variables están relacionadas, como se ve también con la bondad de ajuste.
3) Para realizar la diagnosis en R necesitamos utilizar las ordenes:
-par(mfrow=c(,)) para dividir como queramos las gráficas según su número.
-La función plot utilizada anteriormente.

-plot(tiempo, residuals(modelo)) para la grafica girada o respecto a los residuos, con una función abline(c(0,0), y otras propiedades a elegir) para hacer una recta horizontal de referencia en 0.

-La gráfica de la qqnormal: qqnorm(residuals(modelo),propiedades).

-Y un histograma donde a parte de la qqnormal se puede estudiar la normalidad: hist(residuals(modelo), propiedades).




Vamos a analizar lo que se puede sacar de cada gráfica para comprobar la validez de las hipótesis.

DE la primera gráfica podemos observar que no hay linealidad ya que forman una especie de parábola. Esto se puede observar de forma más clara en la segunda gráfica, donde además podemos ver que si habrá homocedasticidad ya que en ningún momento vemos ninguna forma de trompeta o algo similar. Esto también se pude observar en el gráfico Normal Q-Q Plot al no formarse una forma de campana en los puntos según ascienden. De este gráfico también podemos sacar que no se cumple la condición de normalidad al estar los puntos de manera dispersa sobre la recta y no solo los pocos datos atípicos que hay. Gracias al histograma podemos justificar que no hay normalidad al no tener este forma de función normal, no tener forma de parábola centrada.

4) Mediante prueba y error se prueban distintas transformaciones hasta llegar a una que pueda hacer que se cumplan todas las propiedades anteriores.

   Se llega a la conclusión que la mejor transformación es la de pasar ambas variables a log, hacer conc=log(conc) .

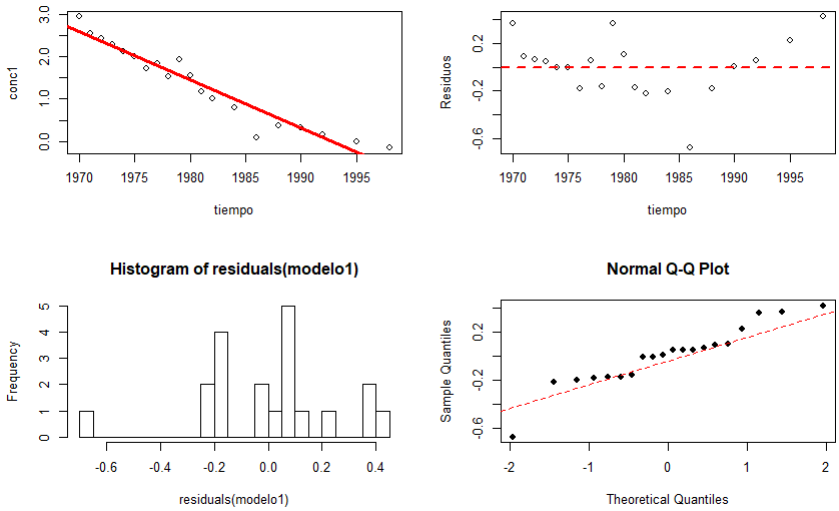
5) Para este nuevo modelo obtenemos la siguiente ecuación: y=225,91-0,113x , a parte de haber cambiado el valor de los coeficientes la R2 y las gráficas sufren otros cambios.

   Empezando por los coeficientes, el nuevo beta0 es 225,91 y el nuevo beta1 es -0,113 , ambos coeficientes han disminuido de valor y en el caso del beta1 sigue haciendo que la relación entre concentración y tiempo sea inversa. Ambas estimaciones son válidas ya que su p-valor es muy inferior a cualquier Alpha por lo que no se cumple ninguna de las hipótesis iniciales. Otro aspecto del modelo que se ve alterado es el R2 aumentando a un 0,9319 lo que quiere decir que un 93,19% de la variabilidad de la concentración viene explicada por el tiempo, siendo este un cambio favorable.

   Respecto a la gráfica podemos observar en las dos primeras que ahora si que hay linealidad y homocedasticidad, aunque también se observan cerca de tres datos

atípicos. Observando las dos últimas no podemos concluir que cumpla la propiedad de normalidad, pero tiene una forma más cercana a la que teníamos en el modelo sin transformación.

```
> summary(modelo1)

Call:
lm(formula = conc1 ~ tiempo)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67895 -0.17283  0.02955  0.09308  0.42358

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 225.913846  14.308314   15.79 5.45e-12 ***
tiempo       -0.113363   0.007223  -15.70 6.02e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2581 on 18 degrees of freedom
Multiple R-squared:  0.9319,    Adjusted R-squared:  0.9281
F-statistic: 246.4 on 1 and 18 DF,  p-value: 6.022e-12
```




6) Realizando los siguientes cálculos obtenemos:

```
> tasa=((datos[20,2]-datos[1,2])/datos[1,2])*100
> tasa
[1] -95.57061
> tiempo=(225.913846-log(datos[1,2]/2))/0.113363-datos[1,1]
> tiempo
[1] 2.888896
```

Siendo los datos empleados el valor final e inicial.
La tasa da negativa por estar relacionados inversamente y decrecer.
El tiempo se halla despejando la x de la ecuación del modelo de los apartados 4 y 5.

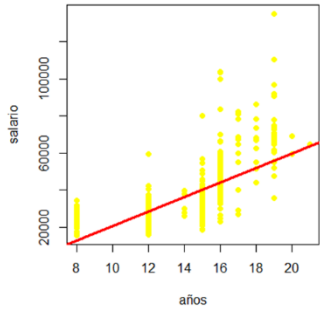
EJERCICIO 2: Habiendo realizado la Tarea 1 (respecto de su correspondiente ejercicio 3)

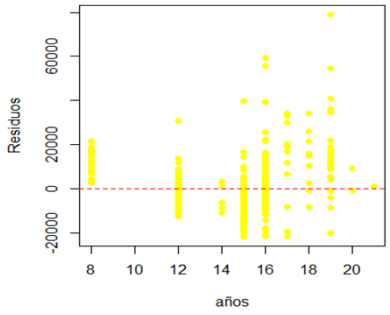
La ecuación de regresión estimada era: y=3909,9x-18331,2 correspondiendo a la y un valor de salario y a la x uno de años.

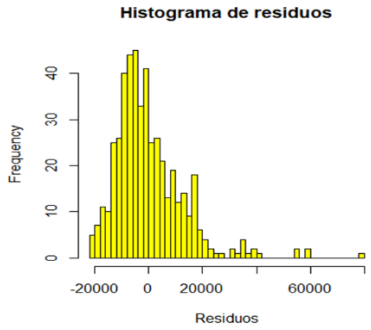
El 3909,9 corresponde al valor estimado de beta1 y sus unidades serán de dolares.

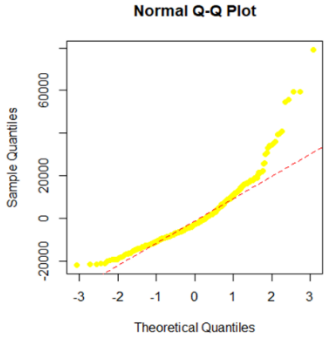
El -18331,2 corresponde al valor estimado de beta0 y sus unidades serán de dólares anuales.

```
Call:
lm(formula = salario ~ aÃ.os)

Residuals:
   Min     1Q Median     3Q    Max
-21567  -8210  -2503   5877  79043

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -18331.2     2821.9  -6.496  2.1e-10 ***
aÃ.os         3909.9      204.5  19.115  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12830 on 472 degrees of freedom
Multiple R-squared:  0.4363,    Adjusted R-squared:  0.4351
F-statistic: 365.4 on 1 and 472 DF,  p-value: < 2.2e-16
```



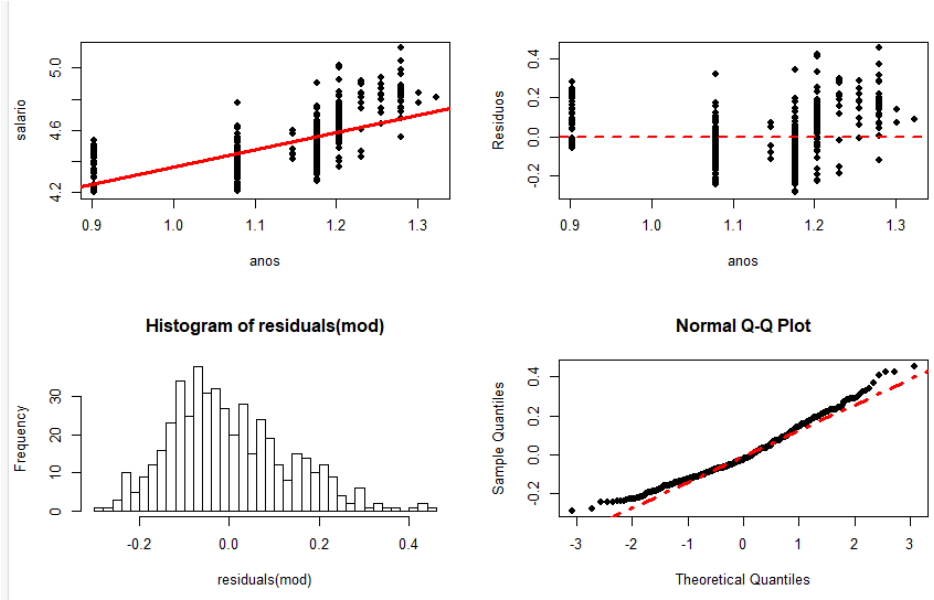




A continuación, vamos a realizar la transformación log en base 10 de los datos años y salario, con esta podemos lograr mejor p-valor para beta0 y hacer que el modelo cumpla la linealidad homocedasticidad y normalidad como se puede observar en las siguientes gráficas.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.25615    0.06895   47.22   <2e-16 ***
anos         1.10942    0.06136   18.08   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1328 on 472 degrees of freedom
Multiple R-squared:  0.4092,    Adjusted R-squared:  0.4079
F-statistic: 326.9 on 1 and 472 DF,  p-value: < 2.2e-16
```

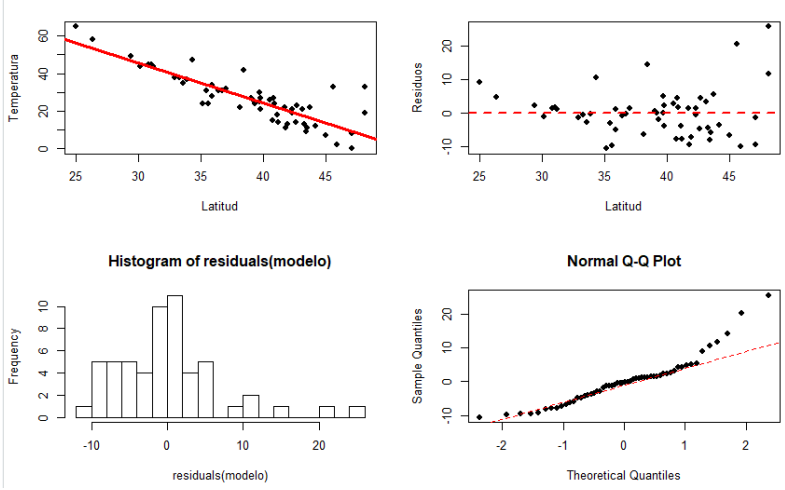



EJERCICIO 3: A partir del fichero Temperatura.txt.

**A.** 1) Realizando el modelo obtenemos la siguiente ecuación: y=108,73-2,11x , siendo la y un valor de temperatura y la x uno de latitud.
Respecto a este resultado el valor de beta0 estimado es de 108,73 que significa que el modelo no pasa por el origen de coordenadas. La beta1 toma un valor de -2,11 que significa que la pendiente de la recta es negativa y por tanto la relación entre temperatura y latitud es inversa.
La varianza residual que resulta del modelo, con 54 grados de libertad, es el cuadrado de la desviación típica residual, siendo esta un 7,156, por tanto la varianza es 51,208. Esta influye en el cálculo del error de estimación.
La R2 es 0,7192 lo que significa que la variabilidad de la temperatura viene explicada un 71,92% por la latitud.

2) EL regresor Latitud si es significativo. Una explicación es la relación expresada en el R2 y otra sería que el valor estimado de beta1 es acepado y distinto de cero por lo que la ecuación relaciona la temperatura con la latitud. Es aceptable ya que el p-valor es muy inferior para cualquier valor de Alpha dado, por lo que se rechaza la hipótesis inicial, beta1=0.

```
Call:
lm(formula = Temperatura ~ Latitud)

Residuals:
     Min       1Q   Median       3Q      Max
-10.6812  -4.5018  -0.2593   2.2489  25.7434

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 108.7277     7.0561   15.41   <2e-16 ***
Latitud      -2.1096     0.1794  -11.76   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.156 on 54 degrees of freedom
Multiple R-squared:  0.7192,    Adjusted R-squared:  0.714
F-statistic: 138.3 on 1 and 54 DF,  p-value: < 2.2e-16
```




3) Para la ordenada en el origen (beta0) y la pendiente de la recta (beta1) se obtienen los siguientes intervalos de confianza al 95%.

```
> confint(modelo)
                2.5 %    97.5 %
(Intercept) 94.581106 122.87438
Latitud     -2.469256  -1.74992
```

Siendo para beta0: de 94,58 a 122,87.
Y para beta1: -2,47 a -1.75.
Para la varianza del modelo obtenemos:

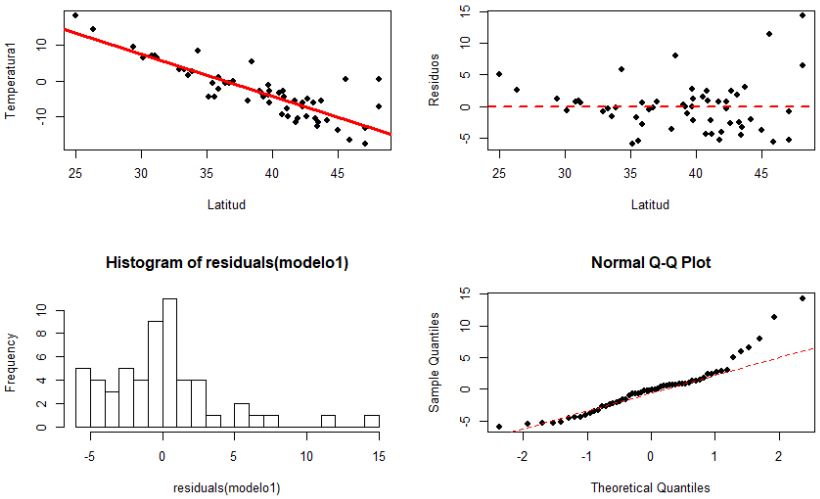
```
> 54*(7.1556^2)/qchisq(0.975,54)
[1] 36.2891
> 54*(7.1556^2)/qchisq(0.025,54)
[1] 77.69669
```

Por tanto el intervalo de la varianza queda: de 36,29 a 77,69.

4) El coeficiente de correlación coincide con la raíz del coeficiente de determinación. Si el coeficiente de determinación es 0.7192, el de correlación es 0.8481.

5) El resultado del modelo en Celsius es el siguiente:

```
Call:
lm(formula = Temperatura1 ~ Latitud)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9340 -2.5010 -0.1441  1.2494 14.3019

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.62652    3.92006   10.87 3.24e-15 ***
Latitud     -1.17199    0.09966  -11.76  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.975 on 54 degrees of freedom
Multiple R-squared:  0.7192,    Adjusted R-squared:  0.714
F-statistic: 138.3 on 1 and 54 DF,  p-value: < 2.2e-16
```

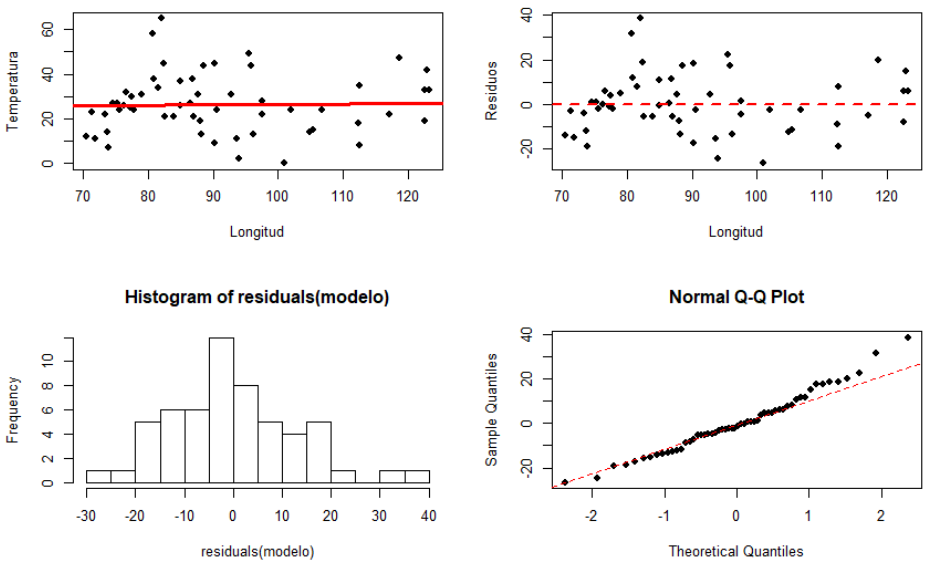



Por lo que el coeficiente de correlación será en este caso el mismo que en el caso anterior ya que resultan del mismo coeficiente de determinación.

6) Como se ha mencionado en el apartado anterior en ambos casos coinciden los coeficientes de determinación y por tanto también el de correlación.
Esto es debido a que la transformación es un cambio de unidades no cambia la relación entre las variables ni la forma de las gráficas, solo valores de temperatura, no se ven afectada ni linealidad ni el resto de sus propiedades.

**B.** 7) Realizando el modelo obtenemos la siguiente ecuación: y=24,87+0,018x , siendo la y un valor de temperatura y la x uno de longitud.
Respecto a este resultado el valor de beta0 estimado es de 24,87 que significa que el modelo no pasa por el origen de coordenadas. La beta1 toma un valor de 0,018 que significa que la pendiente de la recta es positiva y por tanto la relación entre temperatura y longitud es directa.
La varianza residual que resulta del modelo, con 54 grados de libertad, es el cuadrado de la desviación típica residual, siendo esta un 13,5, por tanto la varianza es 182,25. Esta influye en el cálculo del error de estimación.
La R2 es 0,0004165 lo que significa que la variabilidad de la temperatura viene explicada un 0,042% por la longitud, por lo que es un dato peor que la latitud para trabajar sobre la temperatura y crear un modelo. Esto también se puede ver con el p-valor ya que tiene p-valor es alto por lo que la estimación de beta1 no sería apropiada.

```
Call:
lm(formula = Temperatura ~ Longitud)

Residuals:
    Min      1Q  Median      3Q     Max
-26.697  -8.289  -1.754   6.350  38.645

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.87483   11.10076   2.241   0.0292 *
Longitud     0.01805    0.12030   0.150   0.8813
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.5 on 54 degrees of freedom
Multiple R-squared:  0.0004165, Adjusted R-squared:  -0.01809
F-statistic: 0.0225 on 1 and 54 DF,  p-value: 0.8813
```




Al no ser aceptadas las estimaciones, y aceptar la hipótesis inicial (beta1=0) , no hay relación entre la temperatura y la longitud. Por eso la primera gráfica tiene esa forma al ser la ecuación y=24,87.

8) El p-valor del contraste de la F tiene como resultado 0.8813 y tiene 54 grados de libertad. Se observa que el p-valor es grande para cualquier Alpha, por ello se aceptaría la hipótesis inicial (=0) del contraste de la F. Por lo que la longitud no es significativa.

9) Para la ordenada en el origen (beta0) y la pendiente de la recta (beta1) se obtienen los siguientes intervalos de confianza al 95%.
Siendo para beta0: de 2,62 a 47,13.
Y para beta1: -0,22 a 0,26.
El intervalo de la varianza queda: de 129,167 a 276,55.

```
> confint(modelo)
                2.5 %     97.5 %
(Intercept)  2.6191452 47.1305170
Longitud    -0.2231476  0.2592396
> 54*(13.5^2)/qchisq(0.975,54)
[1] 129.167
> 54*(13.5^2)/qchisq(0.025,54)
[1] 276.5527
```

**C.** 10) El mejor de los modelos es el primero, tanto en Celsius como en Fahrenheit, ya que tiene un coeficiente de determinación (R2) mucho mayor y también debido a que las estimaciones de beta0 y beta1 son válidas debido a sus p-valores tan pequeños.
Otra justificación es que la estimación del modelo longitud ha sido rechazado aceptando la hipótesis inicial, beta1=0 , por lo que la longitud no se relacionaría con la temperatura en la ecuación resultante, la temperatura es independiente de la longitud. (Ver la primera gráfica de este modelo).