TAREA 3

EJERCICIO 1: CRIME (con Alpha=0.05)

1. Tras realizar en Rstudio:

```
modelo=lm( crimes ~ income)
summary(modelo)
```

Resulta el modelo estimado: **crimes=-0,4568+0,3052*income**
El **coeficiente de determinación** resultante es del **69,4%**. Lo que significa que un 69,4% de la variabilidad del crimen viene explicada y relacionada directamente por el income.

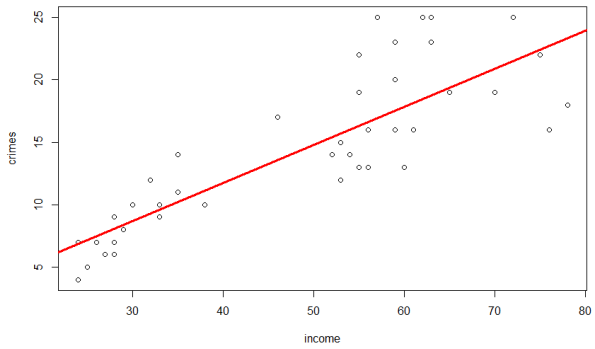
```
Call:
lm(formula = crimes ~ income)

Residuals:
    Min      1Q  Median      3Q     Max
-6.7384 -2.0240 -0.6148  2.4500  8.0604

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4568     1.5805  -0.289    0.774
income        0.3052     0.0309   9.876 1.26e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.345 on 43 degrees of freedom
Multiple R-squared:  0.694,     Adjusted R-squared:  0.6869
F-statistic: 97.54 on 1 and 43 DF,  p-value: 1.259e-12
```

Para el modelo estimado resulta la siguiente gráfica (realizada con la función plot):



2. El valor estimado para $\boldsymbol{\beta_1}$ ha resultado ser **0,3052**. Por cada incremento de una unidad del income ($), la variable crimes aumentará en promedio 0,3052 unidades. Esto significa que el modelo resultante tiene pendiente positiva. A este valor estimado de β_1 le corresponde un **p-valor** de **1,26e-12** que es muy pequeño, por lo que el regresor income es significativo. Se rechaza la hipótesis inicial (Ho: β_1 =0).

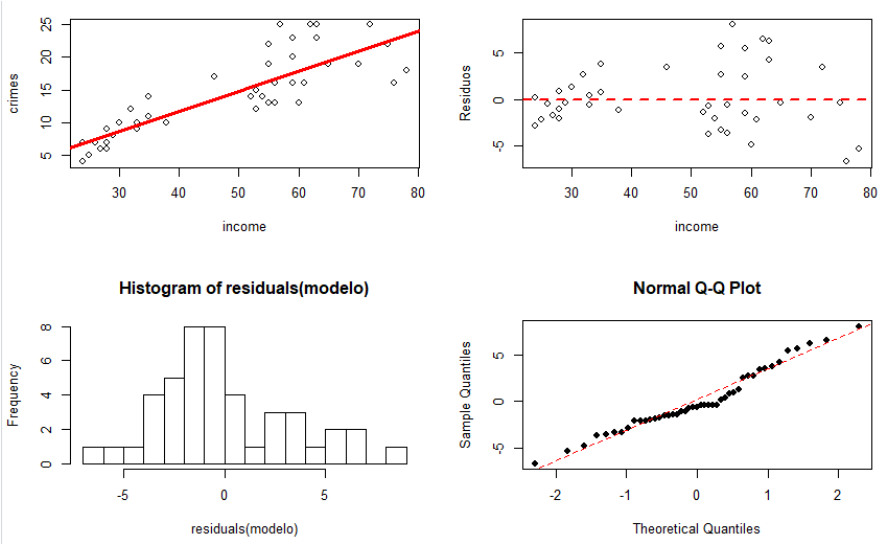
Su valor del estadístico del contraste también es lo suficientemente grande, 9.876. El contraste realizado corresponde con la distribución t-student con 43 grados de libertad.

Además, sabemos que es buena estimación porque un R^2 es de 0,694 que es un valor bastante alto.

3. Para realizar la diagnosis en Rstudio se han realizado:

```
par(mfrow=c(2,2))
plot(income, crimes)
abline(modelo,cex=2, lty=1,col=2,lwd=3)
plot(income, residuals(modelo), ylab="Residuos")
abline(c(0,0), col='red', lty=2, lwd=2)
hist(residuals(modelo), nclas=20)
qqnorm(residuals(modelo), pch=19)
qqline(residuals(modelo), col='red', lwd=1,lty=2)
```

Dando como resultado las siguientes gráficas:



De la primera gráfica del modelo, gráfica que aparece en el primer apartado, podemos observar que no se cumple la hipótesis de linealidad al no haber ninguna tendencia de recta pero tampoco se ve una curva clara. (Con dudas, no podemos afirmar claramente que sea lineal ni que no lo sea, comparar con la transformación)

En las gráficas de la primera fila podemos observar que no se cumple la hipótesis de homocedasticidad ya que hay diferentes anchuras.

De la segunda fila podemos observar que no se llega a cumplir tampoco la hipótesis de normalidad, el histograma no tiene forma de normal y los puntos del Q-Q Plot no están sobre la línea, ni siguen su forma.

La hipótesis de independencia es difícil de estudiar de esta forma por lo que no sabremos si se cumple.

Al no cumplirse ni la hipótesis de homocedasticidad ni la de normalidad **la diagnosis no es la adecuada**.

4. Como la diagnosis no ha sido adecuada, se prueban distintas transformaciones hasta encontrar una posible transformación que mejore la diagnosis.

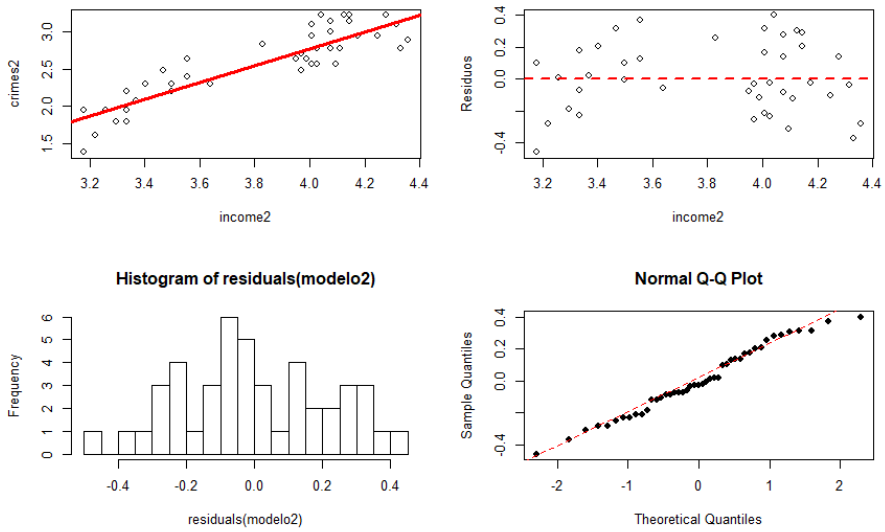
La transformación a la que hemos llegado, que mejoré la diagnosis, ha resultado ser la **transformación** de las variables crimes e income por sus **logaritmos**.
Transformación en Rstudio:

```
crimes2=log(crimes)
income2=log(income)
```

Realizando la diagnosis del modelo2 en Rstudio:

```
par(mfrow=c(2,2))
plot(income2, crimes2)
abline(modelo2,cex=2, lty=1,col=2,lwd=3)
plot(income2, residuals(modelo2), ylab="Residuos")
abline(c(0,0), col='red', lty=2, lwd=2)
hist(residuals(modelo2), nclas=20)
qqnorm(residuals(modelo2), pch=19)
qqline(residuals(modelo2), col='red', lwd=1,lty=2)
```

Dando como resultado las siguientes gráficas:



Como observamos en la primera gráfica se cumple la hipótesis de linealidad, no observamos ninguna tendencia de curva, aunque con dudas.
De la primera fila de gráficas podemos afirmar que la transformación hace que se cumpla la hipótesis de homocedasticidad al tener en general la misma anchura, y al parecer no hay ningún valor atípico.
A partir de la segunda fila de gráficas también podemos concluir que la transformación ha logrado que se cumpla la hipótesis de normalidad, el histograma tiene forma de normal y los puntos de la Q-Q Plot están sobre la línea o siguen su forma.
Podemos afirmar que **la diagnosis es adecuada gracias a las transformaciones** de income y crimes a sus **respectivos logaritmos**.

5. El modelo2 resultante es: **crimes2=-1,7278+1,1244*income2**
   Siendo crimes2 e income2 las transformaciones de sus respectivos logaritmos.
   Tras realizar en Rstudio:

```
crimes2=log(crimes)
income2=log(income)
modelo2=lm( crimes2 ~ income2)
summary(modelo2)
```

```
> summary(modelo2)

Call:
lm(formula = crimes2 ~ income2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.45924 -0.12181 -0.02565  0.16646  0.40073

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.72788    0.33832  -5.107 7.12e-06 ***
income2      1.12440    0.08817  12.753 3.31e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2161 on 43 degrees of freedom
Multiple R-squared:  0.7909,    Adjusted R-squared:  0.786
F-statistic: 162.6 on 1 and 43 DF,  p-value: 3.307e-16
```

De este modelo tomamos que la estimación de $\boldsymbol{\beta_1}$ es **1,124**. Que al ser positivo significa que el modelo tiene pendiente positiva. Esto significa que por un incremento de 1% de income, crimes aumentará un 1,124%. Sabemos que su p-valor es de 3,31e-16 por lo que es muy pequeño y podemos rechazar la hipótesis inicia (Ho: β_1 =0), aceptando la estimación, siendo un valor significativo. Su valor estadístico de contraste también es lo suficientemente grande, 12,753. Contraste correspondiente con la distribución t-student con 43 grados de libertad. De este nuevo modelo también sacamos que esta beta1 nueva tiene un error mayor respecto la estimación del modelo anterior. Y que el nuevo coeficiente de determinación es de un 79,09%, aumentando respecto del 69,4% anterior, por lo que la variabilidad del crime2 viene explicada y relacionada directamente por en un 79,09% por el income2, valor bastante bueno.

6. Realizando en Rstudio:

```
modelom=lm( crimes2 ~ income2 + age)
summary(modelom)
```

Siendo crimes2 e income2 el logaritmo de las variables, el nombre viene de que han sido utilizados para transformación del apartado anterior.

Nos da el siguiente modelo: **log(crime)=-2,398+1,1457*log(income)+0,0248*age**
Siendo **Sr= 0,1781**.

```
> summary(modelom)

Call:
lm(formula = crimes2 ~ income2 + age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40201 -0.12069  0.01446  0.10463  0.39022

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.39847    0.31437  -7.629 1.85e-09 ***
income2      1.14575    0.07281  15.737  < 2e-16 ***
age          0.02480    0.00537   4.617 3.65e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1781 on 42 degrees of freedom
Multiple R-squared:  0.8613,    Adjusted R-squared:  0.8547
F-statistic: 130.4 on 2 and 42 DF,  p-value: < 2.2e-16
```

A continuación, se interpretan los valores y se indican si son significativos:
-$\boldsymbol{\beta_0}$: Para un income2 y una age igual a 0 el término independiente sería -2,398. Su correspondiente p-valor es de 1,85e-9. Será el valor medio de crimes en el origen.
-$\boldsymbol{\beta_1}$: Según este modelo múltiple el valor estimado es 1,1457. Que significa que a igualdad del resto de las variables el aumento de un 1% de income, crimes aumenta 1,1457%. Su correspondiente p-valor es menor que 2e-16.
-$\boldsymbol{\beta_2}$: Se ha obtenido un valor estimado de 0,02480. Esto expresa que a igualdad del resto de las variables el aumento de una unidad del age (años), crimes aumenta en promedio un 2,48%. Su correspondiente p-valor es de 3,65e-5.
Para las tres estimaciones tenemos p-valores tan pequeños que hacen que las variables sean las tres **significativas**. Además, las tres tienen valores del estadístico lo suficientemente grandes, siendo respectivamente: 15,737 y 4,617 para β_1 y β_2 . Contraste correspondiente con la distribución t-student con 42 grados de libertad.
Podemos comprobar con la distribución F-student, al resultar un p-valor muy bajo, que se rechaza la hipótesis múltiple de que todas las betas sean iguales e iguales a cero. Por lo que al menos una es distinta de las demás y de cero.

7. Realizando en Rstudio (para las betas):

```
> confint(modelom, level=0.95)
                2.5 %      97.5 %
(Intercept) -3.03288949 -1.76404597
income2      0.99882122  1.29268374
age          0.01395798  0.03563407
```

Obtenemos para:
-$\boldsymbol{\beta_0}$: [-3.0328,-1.764]
-$\boldsymbol{\beta_1}$: [0.9988,1.2926]
-$\boldsymbol{\beta_2}$: [0.0139, 0.0356]

Realizando en Rstudio (para la **varianza**):

```
> LI=42*(0.1781^2)/qchisq(0.975,42)
> LS=42*(0.1781^2)/qchisq(0.025,42)
> LI
[1] 0.02156513
> LS
[1] 0.05124201
```

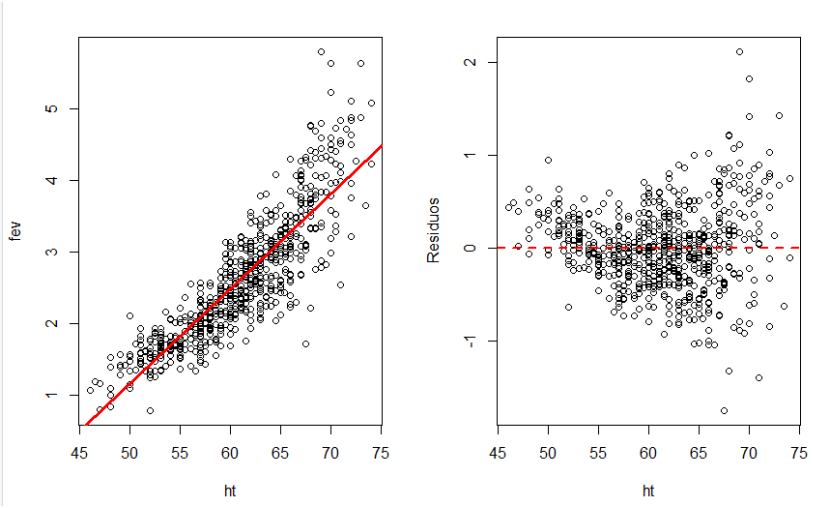
Siendo el intervalo de confianza: [0.02156,0.0051242]

EJERCICIO 2: FEV (con Alpha=0.005)

1. Para realizar el gráfico o gráficos necesarios se han utilizado los siguientes comandos en Rstudio:

```
modelo=lm(fev ~ ht)
par(mfrow=c(1,2))
plot(ht, fev)
abline(modelo,cex=2, lty=1,col=2,lwd=3)
plot(ht, residuals(modelo), ylab="Residuos")
abline(c(0,0), col='red', lty=2, lwd=2)
```

   Resultando los siguientes gráficos, los necesarios para ver linealidad y homocedasticidad del modelo que relaciona fev con ht.

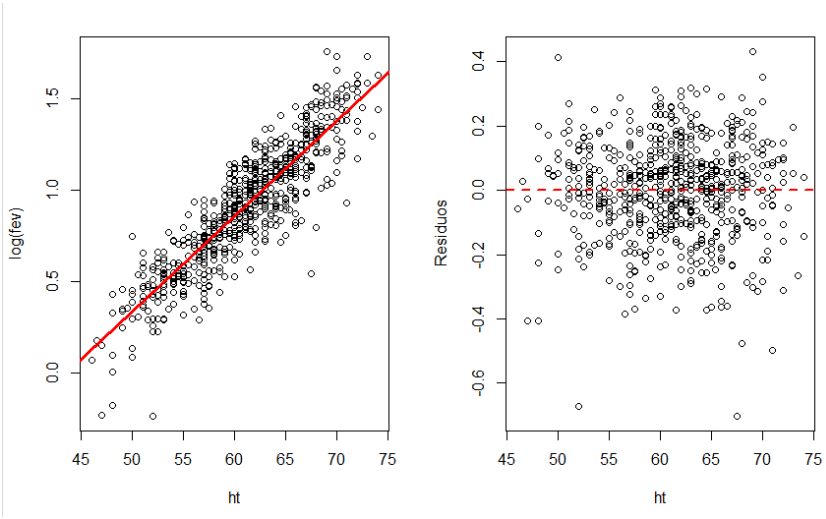


   Se puede observar en ambos gráficos que el conjunto de puntos no toma forma de curva exactamente, pero las zonas con mayor densidad de puntos si, la más oscura. Además la zona de la izquierda de ambas tienden a abrirse por lo que se separan de una posible forma de recta. Aunque sea ligero, la relación **no** es **lineal**, pero si fuese lineal no se puede ver claramente su confirmación (en comparación con la transformación del apartado siguiente).
   En general la anchura de la nube de puntos es igual a lo largo de ambos gráficos, pero en el lado izquierdo de la segunda gráfica vemos como la diferencia de anchura es mayor, por lo que la relación **tampoco** es **homocedástica**.

2. Probando la **transformación** de **log(fev)** en vez de fev podemos observar que ambos gráficos mejoran. La linealidad ahora si que se puede confirmar y se ve claramente que es lineal, sigue la forma de la recta. Y gracias a esta transformación el nuevo modelo tiene relación homocedástica, la anchura de la segunda gráfica es uniforme. (Además, aunque no se pide este nuevo modelo tiene mayor R^2)
   Realización del nuevo modelo en Rstudio:

```
modelo=lm(log(fev) ~ ht)
```

   Gráficas mencionadas:



El nuevo modelo estimado es: **log(fev)=-2.27131+0.052119*ht**

```
> summary(modelo)

Call:
lm(formula = log(fev) ~ ht)

Residuals:
     Min       1Q   Median       3Q      Max
-0.70208 -0.08986  0.01190  0.09337  0.43174

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.271312   0.063531  -35.75   <2e-16 ***
ht           0.052119   0.001035   50.38   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1508 on 652 degrees of freedom
Multiple R-squared:  0.7956,    Adjusted R-squared:  0.7953
F-statistic:  2538 on 1 and 652 DF,  p-value: < 2.2e-16
```

El **coeficiente de determinación** es del **79,56%** . Lo que significa que un 79,56% de la variabilidad del log(fev) viene explicada y relacionada directamente por ht. La varianza residual que resulta del nuevo modelo es el cuadrado de la desviación típica residual, siendo esta 0.1508, por tanto la **varianza residual** es **0.02274**. Esta influye en el cálculo del error de estimación.

3. El valor estimado de $\boldsymbol{\beta_1}$ resultante es **0.052119**, del modelo de regresión simple transformado. Interpretamos que por un incremento en una unidad de ht (pulgadas), el fev aumenta en promedio 5,2119%. Su **p-valor** resulta **ser menor que 2e-16**, valor muy pequeño, lo que significa que se **rechaza** la hipótesis inicial (Ho: β_1 =0). El contraste realizado corresponde con la distribución t-student con 652 grados de libertad. El valor estadístico el contraste es 50,38.

4. Realizando en Rstudio:

```
> LI=652*(0.1508^2)/qchisq(0.975,652)
> LS=652*(0.1508^2)/qchisq(0.025,652)
> LI
[1] 0.02046074
> LS
[1] 0.02542611
```

El intervalo para la **varianza** queda: [0.02046,0.025426]

5. Realizando en Rstudio:

```
> confint(modelo, level=0.95)
                 2.5 %      97.5 %
(Intercept) -2.39606183 -2.14656169
ht           0.05008763  0.05415058
```

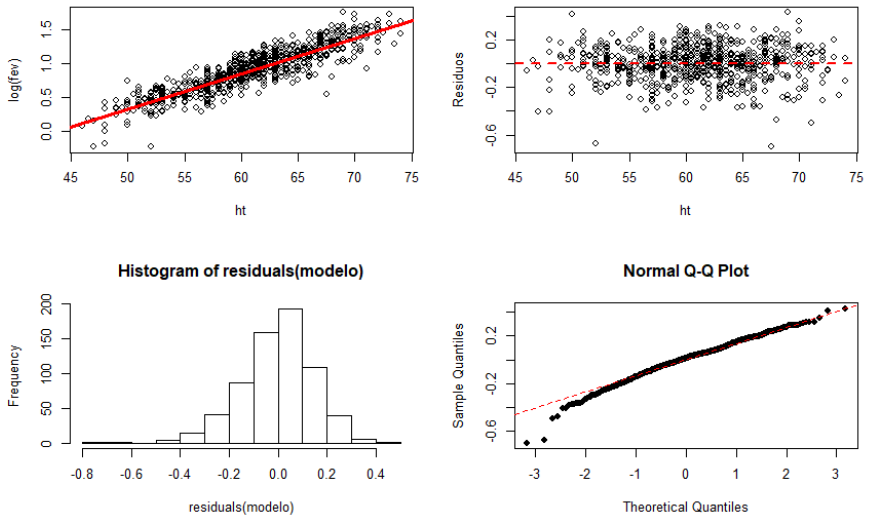
Tenemos que para $\boldsymbol{\beta_1}$ el intervalo de confianza será: [0.050087,0.054150]

6. Como para el nivel de confianza anterior el cero no está incluido la estimación de $\boldsymbol{\beta_1}$ es aceptable y confirmamos lo expresado en el apartado 3.

7. Realizando en Rstudio:

```
par(mfrow=c(2,2))
plot(ht, log(fev))
abline(modelo,cex=2, lty=1,col=2,lwd=3)
plot(ht, residuals(modelo), ylab="Residuos")
abline(c(0,0), col='red', lty=2, lwd=2)
hist(residuals(modelo))
qqnorm(residuals(modelo), pch=19)
qqline(residuals(modelo), col='red', lwd=1,lty=2)
```

Aparecen los siguientes gráficos:



Como podemos observar en las dos primeras gráficas, analizadas en el apartado 2, se cumplen las hipótesis de linealidad y homocedasticidad. (Esto se analizó en el apartado 2 para confirmar que la transformación es favorable). En ambas vemos que la nube de puntos toma forma de recta y no de curva, siendo por tanto lineal. En la segunda podemos ver una anchura uniforme, confirmando la homocedasticidad que se quería lograr.

Además de repetir las dos primeras gráficas para volver a justificar que se cumplen linealidad y homocedasticidad hemos realizado un histograma y un gráfico Q-Q Plot para estudiar la normalidad. En el histograma podemos observar una forma de función normal y en el Q-Q Plot vemos como los puntos están sobre

la recta y siguen su forma por lo que se puede confirmar la normalidad del modelo.
La independencia es difícil de estudiar y la diagnosis no nos da información.
Con todo lo anterior llegamos a la conclusión de que la **diagnosis** es **adecuada**.

8. El modelo estimado con la transformación log(fev) hace que la diagnosis sea adecuada cumpliendo linealidad, homocedasticidad y normalidad. La independencia es difícil de observar a partir de la diagnosis por lo que no lo sabemos.
   El coeficiente de determinación es del 79,56%. Mejorando respecto al modelo sin transformación.
   Con todo lo anterior podemos confirmar que el **modelo** transformado es el **correcto**.

9. Realizando en Rstudio:

```
modelom=lm( log(fev) ~ ht + age)
summary(modelom)
```

Nos da el siguiente modelo: **log(fev)=-1,971147+0,04399*ht+0,019816*age**
Siendo **Sr= 0,1466**.

```
> summary(modelom)

Call:
lm(formula = log(fev) ~ ht + age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64994 -0.08310  0.01055  0.09324  0.42156

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.971147   0.078332  -25.16  < 2e-16 ***
ht           0.043991   0.001647   26.71  < 2e-16 ***
age          0.019816   0.003181    6.23 8.35e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1466 on 651 degrees of freedom
Multiple R-squared:  0.8071,    Adjusted R-squared:  0.8065
F-statistic:  1362 on 2 and 651 DF,  p-value: < 2.2e-16
```

A continuación, se interpretan los valores y se indican si son significativos:
-$\boldsymbol{\beta_0}$: Para un ht y una age igual a 0 el término independiente sería -1,971147. Su p-valor correspondiente es menor de 2e-16. Será el valor medio de fev en el origen.
-$\boldsymbol{\beta_1}$: Según este modelo múltiple el valor estimado es 0,043991. Que significa que para una igualdad del resto de las variables el aumento en una unidad del ht (años), el fev aumenta en promedio 4,3991%. Su correspondiente p-valor es menor que 2e-16.
-$\boldsymbol{\beta_2}$: Se ha obtenido un valor estimado de 0,019816. Esto expresa que para una igualdad del resto de las variables el aumento de una unidad del age (años), el fev aumenta en promedio 1,9816%. Su correspondiente p-valor es de 8,35e-10.

Para las tres estimaciones tenemos p-valores tan pequeños que hacen que las variables sean las tres **significativas**. Además, las tres tienen valores del estadístico lo suficientemente grandes, siendo respectivamente: 26.71 y 6.23 para β_1 y β_2. Contraste correspondiente con la distribución t-student con 651 grados de libertad.
Podemos comprobar con la distribución F-student, al resultar un p-valor muy bajo, que se rechaza la hipótesis múltiple de que todas las betas sean iguales e iguales a cero. Por lo que al menos una es distinta de las demás y de cero.

10. Realizando en Rstudio (para las betas):

```
> confint(modelom, level=0.95)
                  2.5 %      97.5 %
(Intercept) -2.12496090 -1.81733220
ht           0.04075672  0.04722585
age          0.01357088  0.02606161
```

Obtenemos para:
-$\boldsymbol{\beta_0}$: [-2.12496,-1.8173]
-$\boldsymbol{\beta_1}$: [0.040756,0.0472258]
-$\boldsymbol{\beta_2}$: [0.01357, 0.026061]

Realizando en Rstudio (para la **varianza**):

```
> LI=651*(0.1466^2)/qchisq(0.975,651)
> LS=651*(0.1466^2)/qchisq(0.025,651)
> LI
[1] 0.01933536
> LS
[1] 0.02403164
```

Siendo el intervalo: [0.019335,0.024031]